USC-SIPI REPORT #420

Enhancing Speech to Speech Translation Through Exploitation of Bilingual Resources and Paralinguistic

by

Andreas Tsiartas

May 2014

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

Enhancing Speech to Speech translation through exploitation of bilingual resources and paralinguistic information.

by

Andreas Tsiartas

May 2014

Acknowledgements

Special thanks to Shri Narayanan for advising and supporting me during my PhD. I want to thank Panayiotis Georgiou for his valuable advising in my work. Also, I want to thank Prasanta Ghosh for his valuable ideas and close collaboration in the works presented in chapters 2-4. In addition, Theodora Chaspari for her valuable contribution in the work presented in chapter 2. Additionally, I want to thank Ming Lee, Maarten Van Segbroeck, Nassos Katsamanis and all SAIL lab members for their support and close collaboration. Finally, I want to thank my family and Margarita Siamptani for their support.

Table of Contents

List of Figures

List of Tables

Abstract

This thesis focuses on developing a speech-to-speech (S2S) translation system that utilizes paralinguistic acoustic cues for achieving successful cross-lingual interaction. To enable that goal, research is needed at both the foundational speech and language processing, as well as in applying and validating the extracted rich information in translation.

Techniques have been developed that enable more robust signal acquisition through robust voice activity detection (VAD) and cross-talk detection. This can enable hands free S2S communication. The benefits are shown on multiple datasets.

To support rapid technology translation in new language pairs, I have developed novel techniques for extracting parallel audio and text from commonly available bilingual resources such as movies. Also, I have developed a method for aligning subtitles and show performance benefits for translation of spoken utterances by exploiting the timing information of the subtitles to extract high-quality bilingual pairs.

Paralinguistic cues are a big part of spoken communication. To investigate the importance of such cues, I have developed a method to extract bilingual audio pairs from dubbed movies by exploiting the parallel nature of the audio signals and the show performance on English dubbed movies in French. Using these and acted data, I show through perceptual experiments that transfer of paralinguistic acoustic cues from a source language to a target language is correlated with the quality of the spoken

translation for a case study of English-Spanish pair. In addition, a method to represent bilingual paralinguistic acoustic codes is presented.

Chapter 1:

Introduction

The goal of Speech-to-speech (S2S) translation is to allow human interactions across language barriers and enable or aid communication between people with limited or no knowledge of a certain spoken language. S2S translation is expected to be an emerging technology in the years ahead. An industry that can provide a great boost to S2S translation is the tourism is industry. According to the world trade organization, currently, there are 940 million arrivals per annum worldwide that are projected to rise to 1.6 billion by 2020 [66]. S2S translation can prove a great tool to break the language barrier between locals and tourists by aiding them in communicating.

Moreover, a rapidly growing segment of the internet involves the streaming of audiovisual content online (i.e, youtube.com by Google). Google reported that 60% of the youtube.com visits are from people whose primary language is not English [1]. Taking into account the vast amount of audiovisual content added online every day, an interesting application of S2S would be to automatically dub videos/radio talk shows etc. breaking the language barrier and letting the information reach a broader audience.

[1] http://gigaom.com/video/youtube-global-language-stats/

Furthermore, communication and cultural barriers exist within a country. For example, in the US the primary official language is English and there are 50 million people that do not speak English as their first language and 20% do not speak English fluently. In the health domain, studies have showed that translation quality improves health-care access and delivery [64, 90]. In addition, Federal and State health-care providers are mandated to provide language access to non-native English speakers. As an example, the California Bill 853 requires an interpreter present in any language requested. Usually, in large medical facilities and hospitals in urban centers, dedicated interpreters for certain languages are available. However, it is not always possible to facilitate any language in urban areas, let alone rural areas and smaller health centers. In such cases, S2S can be used as a cost effective, efficient and widely deployable solution for facilitating the communication between people that do not share a common language.

Commercial S2S systems include commercial applications that have been developed on desktop systems, for example, Microsoft Windows systems. Recently, there has been development of S2S systems for i-Phone or Android smart-phones. While these systems implement basic S2S functionality, lack the quality and the features for wide spread usage for applications in tourism and the medical domain. Closer to State-of-the-art systems for S2S applications, include systems developed under the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) DARPA program to enable robust, spontaneous two-way tactical speech communications between U.S. war-fighters and native speakers.

Current limitations of the state-of-the-art S2S systems include a pipelined architecture of speech recognition (ASR), machine translation (MT) and speech synthesis (TTS). By the nature of this pipeline approach, the rich information present in speech and spoken discourse is ignored by converting the audio signal into lexical only information, thus, ignoring the paralinguistic acoustic cues in the components following speech

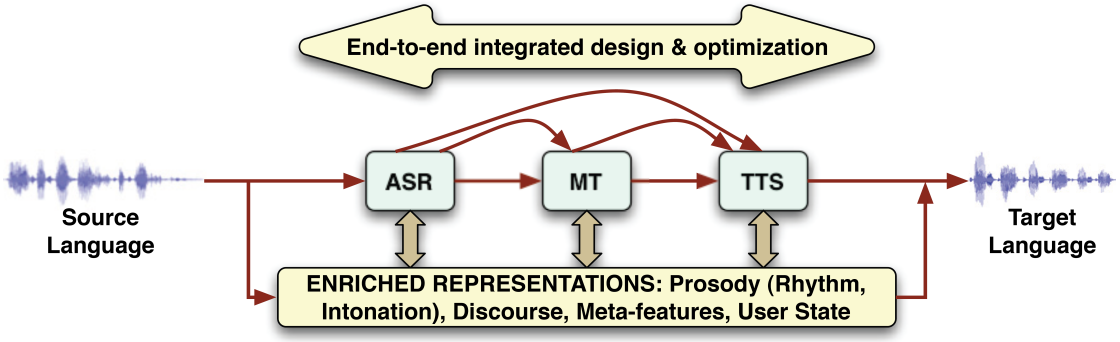

Figure 1.1: Overview of the proposed design.

recognition. Our main premise is that the S2S system should utilize the paralinguistic acoustic cues for achieving a successful cross-lingual interaction.

In addition, due to the pipelined architecture of the S2S systems, there is a lack of a tighter component integration that would enable us to use enriched and contextual information. My vision is an S2S system that translates jointly the contextual, lexical and paralinguistic information with S2S components that can transfer the additional information contained in speech. Fig. 5.1 shows the pipelined approach implemented in current state-of-the-art S2S systems along with our goal to enrich the pipelined approach with acoustic and non-verbal cues, in general. Such cues can be intonational and prosodic patterns including local and global loudness, duration, speech rate, pitch trajectory etc. To make this vision a reality, we need a tighter component integration, for example, the translation component might need to receive and translate acoustic cues from the signal directly or from ASR.

Limitations of state-of-the-art S2S systems include the lack of robust voice activity detection (VAD) especially in noisy environments to seamlessly allow live interactions without the need of a push-to-talk button. Also, the VAD has to be robust in detecting the person speaking in two-way conversations. In this thesis, I tackle the problem of not only robust VAD but also VAD in close-talk conversations.

Additional limitations include lack of vast and scalable sources of bilingual transcriptions that match the domain and speaking style of daily or domain specific conversations. Such data can enable large-scale multiple translation pairs that can scale across different languages. In this work, algorithms are presented to extract such data from movie subtitles. In particular, two approaches are presented exploiting language and timing information of subtitles.

In addition to spoken text transcriptions, there is a need of bilingual spoken data. Bilingual spoken utterances can be extracted from dubbed and can be used to model and represented the speech aspect of speech to speech translation, for example, palanguistic cues translation. I present a language independent method that exploits audio information to extract bilingual audio utterances that can scale in across different language pairs and test the method for the English-French pair.

In chapter 2, I present an approach to detect voice in noisy environments. Also, an application of the VAD in a 3-way two-channel conversation is presented (Patient-doctor-interpreter interaction). In chapter 3, I discuss a method we developed to align parallel bilingual subtitles of movies and their usage in training a machine translation for S2S interactions. In chapter 4, I expand the subtitles alignment to the detection of parallel speech audio segments in original and dubbed movies. An additional section presents an approach to detect the clean speech segments that can be used for S2S translation analysis and modeling. In chapter 5, perceptual experiments are presented asking the question whether the paralinguistic acoustic cues transfer can affect the quality of the S2S translation. In chapter 6, a method to find a representation of bilingual acoustic cues is presented based on objective information theoretic measures. Finally, I conclude the work of this thesis.

Chapter 2:

Voice activity detection

Works presented in this chapter have been carried out in collaboration with Prasanta Gosh and parts with Theodora Chaspari.

2.1 Robust voice activity detection using long-term signal variability [68]

In this chapter, We propose a novel long-term signal variability (LTSV) measure, which describes the degree of non-stationarity of the signal. We analyze the LTSV measure both analytically and empirically for speech and various stationary and non-stationary noises. Based on the analysis, we find that the LTSV measure can be used to discriminate noise from noisy speech signal and, hence, can be used as a potential feature for voice activity detection (VAD). We describe an LTSV-based VAD scheme and evaluate its performance under eleven types of noises and five types of signal-to-noise ratio (SNR) conditions. Comparison with standard VAD schemes demonstrates that the accuracy of the LTSV-based VAD scheme averaged over all noises and all SNRs is ∼6% (absolute) better than that obtained by the best among the considered VAD schemes, namely

AMR-VAD2. We also find that, at -10dB SNR, the accuracies of VAD obtained by the proposed LTSV-based scheme and the best considered VAD scheme are 88.49% and 79.30% respectively. This improvement in the VAD accuracy indicates the robustness of the LTSV feature for VAD at low SNR condition for most of the noises considered.

2.1.1 Introduction

Voice activity detection (VAD) refers to the problem of distinguishing speech from non-speech regions (segments) in an audio stream. The non-speech regions could include silence, noise, or a variety of other acoustic signals. VAD is challenging in low signal-to-noise ratio (SNR), especially in non-stationary noise, because both low SNR and a non-stationary noisy environment tend to cause significant detection errors. There is a wide range of applications for VAD, including mobile communication services [51], real-time speech transmission on the Internet [10], noise reduction for digital hearing aid devices [52], automatic speech recognition [29], and variable rate speech coding [28].

Being a critical component in many applications, VAD has had a lot of attention in the research community over the last few decades. Researchers have proposed a variety of features exploiting the spectro-temporal properties of speech and noise to detect the speech segments present in a noisy observed signal. Many existing algorithms for VAD use features that depend on energy [41, 86, 35]. Some algorithms use a combination of zero-crossing rate (ZCR) and energy [16]; others have used correlation coefficients [5], the wavelet transform coefficients [22], Walsh basis function representation [57], and a distance measure of the cepstral features [46]. More complex algorithms use more than one feature to detect speech [86, 83]. Among the various other proposed features, negentropy has been shown to be robust for VAD [75, 69] at low SNR. Negentropy is the entropy computed using the probability density function (pdf) obtained from normalized short-time spectrum. All of the above-mentioned features are typically computed from

the signal along short-term analysis frames (usually 20 msec long), based on which VAD decisions are taken at each frame. In contrast to the use of frame level features, Ramirez et al [47] proposed the use of long-term spectral divergence between speech and noise for VAD, although they assign the VAD decision directly to the frame in the middle of the chosen long analysis window. Also, the long-term feature proposed in [47] requires average noise spectrum magnitude information, which might not be accurately available in practice. In general, no particular feature or specific set of features has been shown to perform uniformly well under different noise conditions. For example, energy-based features do not work well in low SNR [74] and similarly, under colored noise, negentropy fails to distinguish speech regions from noise with good accuracy due to the colored spectrum of speech. Also, SNR estimation is a critical component in many of the existing VAD schemes, which is particularly difficult in non-stationary noise [19]. Thus, the VAD problem still remains challenging and requires the design of further robust features and algorithms.

Recent works on VAD have been mostly statistical model based [48, 6, 23, 26]. In this approach, VAD is posed as a hypothesis testing problem with statistical models of speech and noise, although assumptions made about the statistics of noise [48, 6, 40, 26] do not always hold in practice. In the short-term frame-level analysis framework, this hypothesis testing problem can be stated as follows: given a frame of observed signal $\{x(n)\}_{n=0}^{N_w-1}$ (N_w is the frame duration in number of samples), the goal is to determine whether the given frame belongs to only noise ($H_0 : \ x(n) = N(n)$) or noisy speech ($H_1 : \ x(n) = s(n) + N(n)$). $s(n)$ and $N(n)$ denote the samples of speech and noise respectively. To check the robustness of both feature-based and model-based approaches, the corresponding VAD performances should be evaluated on a wide range of noises (stationary, non-stationary, impulsive) and under different SNR conditions (particularly at low SNR such as -10dB or -5dB).

Signal characteristics of speech and non-speech sounds have different variability profiles. This can be advantageously used to discriminate them. For example, a person with average speaking rate produces approximately 10-15 phonemes per second [56]. These phonemes have different spectral characteristics. Due to this variability in signal characteristics over time, the speech signal is non-stationary. On the other hand, ideally there is no change over time in the statistics of the stationary noises (both white and colored). The signal characteristics of non-stationary noises change with time; however, we need a metric to compare the variability of non-stationary noise with that of speech. For computing such a metric, we need to analyze the signal over longer duration in contrast to the usual short-term analysis.

In this work, we have proposed a novel long-term signal variability (LTSV) measure, by which the degree of non-stationarity in various signals can be compared. We have demonstrated the usefulness of the LTSV measure as a feature for VAD and have experimentally evaluated its performance under a variety of noise types and SNR conditions (eleven noise types including white, pink, tank, military vehicle, jet cockpit, HF channel, F16 cockpit, car interior, machine gun, babble, and factory noise in five different SNR conditions: -10dB, -5dB, 0dB, 5dB and 10dB). For the proposed signal variability measure, the analysis window goes beyond the usual short-term frame size. The short-term analysis assumes that the speech signal is slowly varying and stationary over 20 msec. The rationale behind our choice of a long analysis window is to obtain a realistic measure of non-stationarity or variability in signal characteristics, which cannot be captured with a short window of 20 msec. We hypothesize that the proposed long-term variability measure for speech will be distinctly greater compared to that obtained for commonly encountered noises. We theoretically show that in additive stationary noise, even at low SNR, it is possible to distinguish speech regions from stationary noisy regions using our proposed method, which is not possible using short-time energy-based features

[74]. Energy-based features depend on the signal amplitude and hence change when the signal is scaled, but our feature is based on the degree of non-stationarity of the signal, which does not get affected by scaling the signal. For additive non-stationary noises, we show experimentally that it is possible to distinguish speech from non-stationary noise with an accuracy as good as that for stationary noise, unless the non-stationary noise and speech have a similar degree of variability, measured by the proposed LTSV metric.

From the LTSV measure, we obtain an indication whether there is a speech signal in the respective long analysis window. However, this decision is not assigned to a 10-20 msec frame in the middle of the long window, unlike [47], for example. We repeat the analysis process across the signal stream with a small shift (this shift determines how coarse we want to make VAD). Thus, we obtain decisions over long windows for each shift. We assimilate all these decisions to arrive at the final frame-level VAD decision. We find that, by utilizing signal information over a longer window, we can make VAD more robust even at low SNR.

2.1.2 Long-term signal variability measure

The long-term signal variability (LTSV) measure at any time is computed using the last R frames of the observed signal $x(n)$ with respect to the current frame of interest. The LTSV, $\mathcal{L}_x(m)$, at the m^{th} ($m \in \mathcal{Z}$) frame is computed as follows:

$$\mathcal{L}_x(m) \triangleq \frac{1}{K}\sum_{k=1}^{K}\left(\xi_k^x(m) - \overline{\xi^x(m)}\right)^2 \tag{2.1}$$

$$\text{where, } \overline{\xi^x(m)} = \frac{1}{K}\sum_{k=1}^{K}\xi_k^x(m)$$

$$\text{and } \xi_k^x(m) \triangleq -\sum_{n=m-R+1}^{m}\frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m}S_x(l,\omega_k)}\log\left(\frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m}S_x(l,\omega_k)}\right) \tag{2.2}$$

$S_x(n, \omega_k)$ is the short time spectrum at ω_k. It is computed as

$$S_x(n, \omega_k) = |X(n, \omega_k)|^2 \text{, where } X(n, \omega_k) = \sum_{l=(n-1)N_{sh}+1}^{N_w+(n-1)N_{sh}} w(l-(n-1)N_{sh}-1)x(l)e^{-j\omega_k l} \quad (2.3)$$

$w(i),\ 0 \leq i < N_w$ is the short-time window, N_w is the frame length, and N_{sh} is the frame shift duration in number of samples. $X(n, \omega_k)$ is the short-time Fourier transform (STFT) coefficient at frequency ω_k, computed for the n^{th} frame.

$\xi_k^x(m)$ is essentially an entropy measure on the normalized short-time spectrum computed at frequency ω_k over R consecutive frames, ending at the m^{th} frame. The signal variability measure $\mathcal{L}_x(m)$ is the sample variance of $\{\xi_k^x(m)\}_{k=1}^K$, i.e., the sample variance of entropies computed at K frequency values. $\mathcal{L}_x(m)$ is, therefore, dependent on the choice of K frequency values $\{\omega_k\}_{k=1}^K$, R, and K itself.

Note that $\mathcal{L}_x(m)$ is invariant to amplitude scaling of the observed signal $x(n)$. $\mathcal{L}_x(m)$ is significantly greater than zero only when the entropies $\{\xi_k^x(m)\}_{k=1}^K$ computed at K frequencies are significantly different from each other. $\mathcal{L}_x(m)$=0 if $\{\xi_k^x(m)\}_{k=1}^K$ are identical for all $k = 1, ..., K$.

2.1.3 Stationary noise case

Let $x(n)$ be a stationary noise (need not be white) $N(n)$. Since $N(n)$ is stationary, the ideal noise spectrum does not change with time, i.e., $S_N(n, \omega_k)$ is ideally constant for all n. Let us assume the actual spectrum of noise is known and $S_N(n, \omega_k) = \sigma_k,\ \forall n$. Thus, using eqn. (2.2), $\xi_k^N(m) = \log R,\ \forall k$[1] and, hence, using eqn. (2.1) $\mathcal{L}_N(m)$=0.

Now consider $x(n)$ to be speech in additive stationary noise, i.e., $x(n) = S(n)+N(n)$. This means that, ideally, $S_x(n, \omega_k) = SS+\sigma_k$ (assuming noise is uncorrelated to signal), where SS is the actual speech spectrum. Thus,

[1]Note that $\xi_k^x(m)$ being an entropy measure can take a maximum value of $\log R$, for a fixed choice of R.

$$\xi_k^{S+N}(m) = -\sum_{n=m-R+1}^{m} \frac{SS+\sigma_k}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k} \log\left(\frac{SS+\sigma_k}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k}\right) \quad (2.4)$$

Note that $\log R = \xi_k^N(m) \geq \xi_k^{S+N}(m) \geq \xi_k^S(m) \geq 0$, where $\xi_k^S(m)$ is the entropy measure when $x(n)$ is only speech $S(n)$ (without any additive noise) and $\xi_k^S(m)$ can be obtained from eqn. (2.4) by setting $\sigma_k = 0$ (see appendix .1 for proof). This means that if there is only speech at frequency ω_k over R consecutive frames, the entropy measure will have a smaller value compared to that of speech plus noise; with more additive stationary noise, the entropy measure increases, and it takes the maximum value when there is no speech component (only noise) at frequency ω_k over R frames. This means that the proportion of the energies of speech and noise (SNR) plays an important role in determining how large or small the entropy measure $\xi_k^{S+N}(m)$ is going to be. Let us denote the SNR at frequency ω_k at n^{th} frame by $SNR_k(n)$ $\left(\triangleq \frac{SS}{\sigma_k}\right)$. It is easy to show that eqn. (2.4) can be rewritten as follows:

$$\begin{aligned}\xi_k^{S+N}(m) = {} & \frac{\sum_{l=m-R+1}^{m} SNR_k(l)}{\sum_{l=m-R+1}^{m} SNR_k(l)+R}\xi_k^S(m) + \frac{R}{\sum_{l=m-R+1}^{m} SNR_k(l)+R}\xi_k^N(m) \\ & - \sum_{n=m-R+1}^{m} \frac{SNR_k(n)}{\sum_{l=m-R+1}^{m} SNR_k(l)+R} \log\left(\frac{1+\frac{1}{SNR_k(n)}}{1+\frac{R}{\sum_{l=m-R+1}^{m} SNR_k(l)}}\right) \\ & - \sum_{n=m-R+1}^{m} \frac{1}{\sum_{l=m-R+1}^{m} SNR_k(l)+R} \log\left(\frac{SNR_k(n)+1}{\frac{\sum_{l=m-R+1}^{m} SNR_k(l)}{R}+1}\right) \quad (2.5)\end{aligned}$$

The first two terms jointly are equal to the convex combination of $\xi_k^S(m)$ and $\xi_k^N(m)$. The third and fourth terms are additional terms. If $SNR_k(n) \gg 1$, $\frac{1}{SNR_k(n)} \approx 0$, $\frac{1}{SNR_k(n)+R} \approx 0$, then the second, third and fourth terms turn very small and, hence, negligible; thus for high SNR at ω_k, $\xi_k^{S+N}(m) \approx \xi_k^S(m)$. Similarly, for low SNR ($SNR_k(n) \ll 1$), $\xi_k^{S+N}(m) \approx \xi_k^N(m)$.

As $\mathcal{L}_{S+N}(m)$ is an estimate of variance of $\left\{\xi_k^{S+N}(m)\right\}_{k=1}^{K}$ and $\xi_k^{S+N}(m)$ depends on SNR_k, the value of $\mathcal{L}_{S+N}(m)$ also depends on the SNRs at $\{\omega_k\}_{k=1}^{K}$. If $SNR_k(m) \ll$

1, $\forall k$, then $\xi_k^{S+N}(m) \approx \xi_k^N(m) = \log R$, $\forall k$ and hence $\mathcal{L}_{S+N}(m) \approx 0$. On the other hand, let us consider the case where the signal contains speech with high SNR. However, it is well known that speech is a low pass signal; the intensity of the speech component at different frequencies varies widely, even up to a range of 50 dB[82]. Thus, in additive stationary noise, SNR_k also varies widely across frequency depending on the overall SNR. Thus, we expect $\mathcal{L}_{S+N}(m)$ to be significantly greater than zero.

Although for the sake of analysis above, we assumed the actual spectrum of noise and speech are known, in practice, we don't know σ_k for any given stationary noise. Thus, we empirically investigate the LTSV measure when the spectrum is estimated from real signal. In this work, we estimate both SS and σ_k by the periodogram method [36] of spectral estimation (eqn. (2.3)). Fig. 2.1(a) shows the histogram of $\log_{10}(\mathcal{L}_N(m))$ for stationary white noise and histogram of $\log_{10}(\mathcal{L}_{S+N}(m))$ for speech in additive stationary white noise at 0dB SNR. For demonstrating the histogram properties, we consider the logarithm of the LTSV feature for better visualization in the small value range of LTSV. In this example, the number of realizations of $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$ are 375872 and 401201, respectively. These samples were computed at every frame from noisy speech obtained by adding white noise to the sentences of the TIMIT training corpus [30] at 0dB SNR. Note that the LTSV computed at the m^{th} frame (i.e. $\mathcal{L}_x(m)$) is considered to be $\mathcal{L}_{S+N}(m)$ if there are speech frames between $(m-R+1)^{\text{th}}$ and m^{th} frame. The sampling frequency for speech signal is F_s=16kHz. The Hanning window is used as the short-time window, $w(i)$ (as in eqn. (2.3)), and we chose the following parameter values N_w=320, N_{sh}=$\frac{N_w}{2}$, R=30, K=448, and $\{\omega_k\}_{k=1}^K$ uniformly distributed between 500 and 4000 Hz. As the spectrum of the noise and the signal plus noise are both estimated using the periodogram, $\mathcal{L}_N(m)$ is not exactly zero ($\mathcal{L}_N(m) \to 0$ is equivalent to $\log_{10}(\mathcal{L}_N(m)) \to -\infty$) although $\mathcal{L}_{S+N}(m) > 0$. The periodogram estimate of spectrum is biased and has high variance [36]. In spite of this, on average, the

values of $\mathcal{L}_N(m)$ are closer to zero compared to that of $\mathcal{L}_{S+N}(m)$; this demonstrates that the entropy measure $\xi_k^x(m)$ varies more over different ω_k when there is speech compared to when there is only noise in the observed signal. As the proportion of speech and noise in the observed signal determines $\mathcal{L}_{S+N}(m)$, we can interpret the above statement in the following way: LTSV captures how SNR_k varies across K frequencies over R frames without explicitly calculating SNRs at $\{\omega_k\}_{k=1}^K$. Fig. 2.1(a) also shows the histogram of $\log_{10}(\mathcal{L}_S(m))$. The sample mean and sample standard deviation (SD) of the realizations of $\mathcal{L}_N$, $\mathcal{L}_{S+N}$, and $\mathcal{L}_S$ are tabulated in the figure. It is clear that in additive noise the mean LTSV decreases (from 14.16×10^{-2} to 6.65×10^{-2}). The SD of $\mathcal{L}_N$ (1.77×10^{-3}) is less than that of $\mathcal{L}_{S+N}$ (5.46×10^{-2}). Thus, in the presence of speech in R frames, LTSV can take a wider range of values compared to that in the absence of speech. We see that there is overlap between the histograms of $\log_{10}(\mathcal{L}_N)$ and $\log_{10}(\mathcal{L}_{S+N})$. We calculated the total misclassification error among these realizations of $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$, which is the sum of the speech detection error (this happens when there is speech over R frames, but gets misclassified as noise) and non-speech detection error. This was done using a threshold obtained by the equal error rate (EER) of the region operating characteristics (ROC) curve [31]. The total misclassification error turned out to be 6.58%.

We found that a better estimate (unbiased with low variance) of the signal spectrum and the noise spectrum leads to a better estimate of $\mathcal{L}_N(m)$ and $\mathcal{L}_{S+N}(m)$ (see appendix .2 for details). Therefore, we use the Bartlett-Welch method of spectral estimation [36]; we estimate the signal spectrum by averaging spectral estimates of M consecutive frames. Thus, eqn. (2.3) is modified to

$$S_x(n,\omega_k) = \frac{1}{M}\sum_{p=n-M+1}^{n}\left|\sum_{l=(p-1)N_{sh}+1}^{N_w+(p-1)N_{sh}} w(l-(p-1)N_{sh}-1)x(l)e^{-j\omega_k l}\right|^2 \tag{2.6}$$

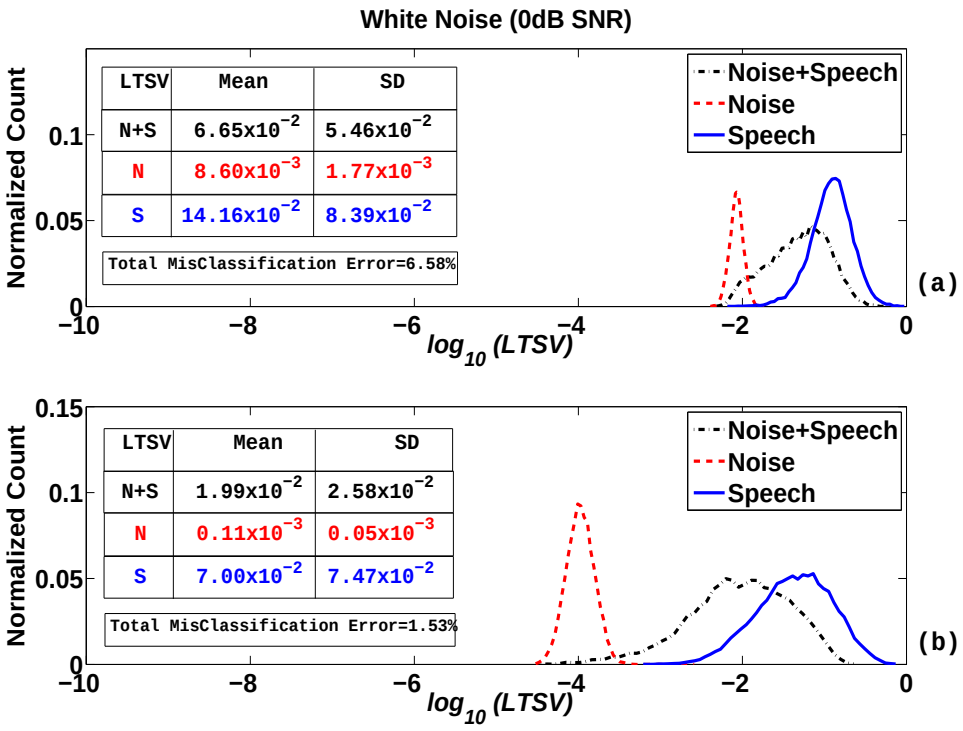

Figure 2.1: *Histogram of the logarithmic LTSV ($\log_{10}(LTSV)$) measure of white noise and speech in additive white noise (0dB SNR) using (a) the periodogram estimate of spectrum (b) the Bartlett-Welch estimate of spectrum. The tables in the figures show the sample mean and sample SD of the realizations of $\mathcal{L}_N$, $\mathcal{L}_{S+N}$, and $\mathcal{L}_S$.*

Fig. 2.1(b) shows the histograms of $\log_{10}(\mathcal{L}_N(m))$, $\log_{10}(\mathcal{L}_{S+N}(m))$, and $\log_{10}(\mathcal{L}_S(m))$, where the spectral estimates are obtained by the Bartlett-Welch method (M=20). We observe that the mean of $\mathcal{L}_N$ has moved closer to 0 compared to that obtained using the periodogram method in Fig. 2.1(a). The SD of $\mathcal{L}_N$ has also decreased; these suggest that the estimate of $\mathcal{L}_N$ is better using the Bartlett-Welch method compared to the periodogram. The mean and SD of both $\mathcal{L}_{S+N}$ and $\mathcal{L}_S$ have also decreased. However, since we don't know the true values of LTSV for speech (and speech+noise), we can't really comment on how good the estimates of $\mathcal{L}_S$ and $\mathcal{L}_{S+N}$ are in Fig. 2.1(b) compared to those of Fig. 2.1(a). The total misclassification error turned out to be 1.53%. The

reduction in misclassification error from 6.58% (Fig. 2.1(a)) to 1.53% (Fig. 2.1(b); 76.75% relative reduction) suggests that the estimate of $\mathcal{L}_x$ using the Bartlett-Welch method improves the speech hit rate and thus is useful for VAD.

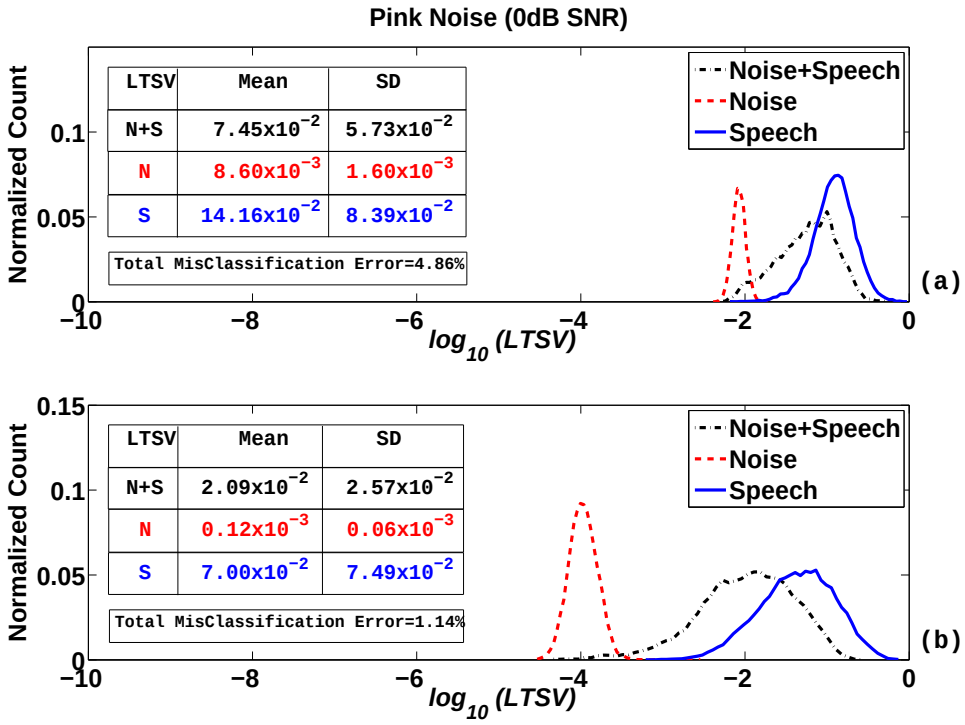

Figure 2.2: *Histogram of the logarithmic LTSV ($\log_{10}(LTSV)$) measure of pink noise and speech in additive pink noise (0dB SNR) using (a) the periodogram estimate of spectrum (b) the Bartlett-Welch estimate of spectrum. The tables in the figures show the sample mean and sample SD of the realizations of $\mathcal{L}_N$, $\mathcal{L}_{S+N}$, and $\mathcal{L}_S$.*

Fig. 2.2 repeats Fig. 2.1 for stationary pink noise. We observe similar trends from Fig. 2.2(a) to Fig. 2.2(b) as seen from Fig. 2.1(a) to Fig. 2.1(b). Since pink noise is colored, $\mathcal{L}_{S+N}(m)$ for pink noise is not the same as that of white noise. However, for both white and pink noises, the mean of $\mathcal{L}_N$ using the periodogram method is of the order of 10^{-2} and that using the Barlett-Welch method is of the order of 10^{-4}. Thus, on average, $\mathcal{L}_N$ obtained by the Bartlett-Welch method is closer to its theoretical value (0)

compared to that obtained by the periodogram method. The misclassification error for additive pink noise reduces from 4.86% for the periodogram method to 1.14% (76.54% relative reduction) for the Bartlett-Welch method. This suggests that the temporal averaging of the short-time spectrum is appropriate to make the estimate of LTSV more robust for VAD in the case of stationary noise even when it is colored.

2.1.4 Nonstationary noise case

The spectrum of nonstationary noise varies with time. Thus, when $x(n)$ is a nonstationary noise, $\mathcal{L}_x(m)$ is no longer 0 even when the actual spectrum of the signal is known and used to compute $\mathcal{L}_x(m)$. $\mathcal{L}_x(m)$ depends on the type of noise and its degree of non-stationarity and hence becomes analytically intractable in general. Speech is a non-stationary signal. Thus, speech in additive nonstationary noise makes the analysis even more challenging. However, the following observations can be made about $\mathcal{L}_x(m)$ when noise is nonstationary:

- $\xi_k^x(m)$ depends on how rapidly $S_x(n, \omega_k)$ changes with n, and $\mathcal{L}_x(m)$ depends on how different $\xi_k^x(m),\ k = 1, ..., K$ are.
- If $S_x(n, \omega_k)$ is slowly varying with n for all $\{\omega_k\}_{k=1}^K$, $\xi_k^x(m)$ is expected to be higher for all $\{\omega_k\}_{k=1}^K$ and hence, $\mathcal{L}_x(m)$ will be close to 0.
- Similarly, if $S_x(n, \omega_k)$ varies rapidly over n for all $\{\omega_k\}_{k=1}^K$, $\xi_k^x(m)$ becomes small $\forall k$ and hence $\mathcal{L}_x(m)$ will also be close to 0.
- However, if $S_x(n, \omega_k)$ varies with n slowly at some ω_k and largely at some other ω_k, $\mathcal{L}_x(m)$ would tend to take a high value. When $x(n) = S(n) + N(n)$, how $S_x(n, \omega_k)$ varies with ω_k depends on the SNR_k.

For nonstationary noises, we demonstrate the efficacy of the LTSV measure by simulations. We obtained samples of non-stationary noises, namely tank, military vehicle, jet

cockpit, HFchannel, F16 cockpit, car interior, machine gun, babble, and factory noise from the NOISEX-92 database [15]. We added these noise samples to the sentences of the TIMIT training corpus [30] at 0dB SNR to compute realizations of $\mathcal{L}_N$, $\mathcal{L}_{S+N}$, and $\mathcal{L}_S$. For illustrations, Fig. 2.3 (a) and (b) show the histograms of $\log_{10}(\mathcal{L}_N)$, $\log_{10}(\mathcal{L}_{S+N})$, and $\log_{10}(\mathcal{L}_S)$ using the Bartlett-Welch spectral estimates for additive (0dB) car interior noise and jet cockpit noise, respectively. The parameter values for Fig. 2.3 are chosen to be the same as those used for Fig. 2.1.

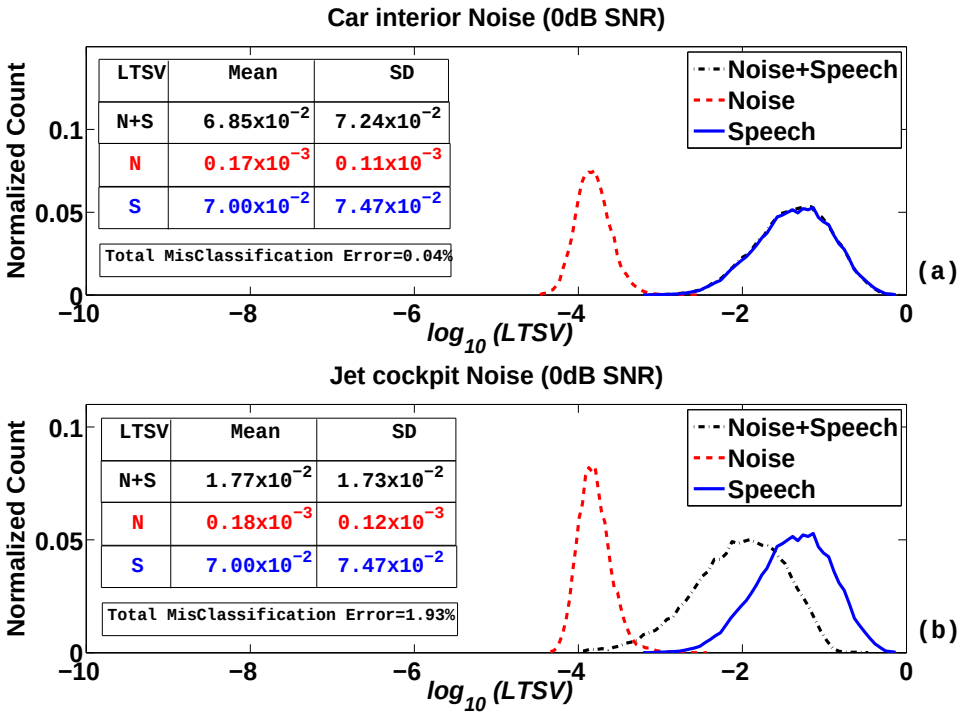

Figure 2.3: *Histogram of the logarithmic LTSV ($\log_{10}(LTSV)$) measure using the Bartlett-Welch spectral estimates for (a) car interior noise and speech in additive car interior noise (0dB), and (b) jet cockpit noise and speech in additive jet cockpit noise (0dB). The tables in the figures show the sample mean and sample SD of the realizations of $\mathcal{L}_N$, $\mathcal{L}_{S+N}$, and $\mathcal{L}_S$.*

In Fig. 2.3(a) (car interior noise), it is seen that the histogram of $\log_{10}(\mathcal{L}_{S+N})$ is not much different from that of $\log_{10}(\mathcal{L}_S)$. This means that in additive car interior noise the

LTSV measure of noisy speech does not change significantly compared to that of speech only. Computation of SNR_k at $\{\omega_k\}_{k=1}^{K}$ for additive car noise (at 0dB SNR) reveals that the average SNR_k is 15dB in the range of frequencies between 500 and 4000 Hz. The spectrum of car interior noise has a relatively large low-pass component below 500 Hz. $\{SNR_k\}_{k=1}^{K}$ being high, $\mathcal{L}_S$ and $\mathcal{L}_{S+N}$ are similar. From Fig. 2.3 (a) it can also be seen that the mean of $\mathcal{L}_N$ is approximately 500 times smaller than the mean of $\mathcal{L}_{S+N}$ and, also, the overlap between the histograms of $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$ is negligible, resulting in a very small misclassification error of 0.04%. Similar to the case of stationary white and pink noises, the use of the Bartlett-Welch method for spectral estimation provides a 63.63% relative reduction in misclassification error compared to that of the periodogram method (0.11%) in additive car noise. Similarly, the misclassification error reduces from 18.11% (the periodogram method) to 1.93% (89.34% relative reduction) (the Bartlett-Welch method) for additive jet cockpit noise.

For a comprehensive analysis and understanding, the total misclassification errors for both stationary and non-stationary noises are presented in Table 2.1 using both the periodogram and the Bartlett-Welch methods. Noises are added to speech at 0dB SNR, and M is chosen to be 20 for the Bartlett-Welch method for this experiment.

Except for the case of factory and babble noise in Table 2.1, we observe a consistent reduction in misclassification error when LTSV is computed using the Bartlett-Welch method (M=20) compared to that using the periodogram method. The percentages of relative reduction indicate that the temporal smoothing with a fixed M in the Bartlett-Welch estimate does not consistently reduce the total misclassification error for all noises compared to that obtained by periodogram estimate.

In particular, the misclassification error is high for machine gun and speech babble noise using both the periodogram and the Bartlett-Welch method. The LTSV measure cannot distinguish these two noises from the corresponding noisy speech. For machine

Noise Type	Total Misclassification Error		
	Periodogram	Bartlett-Welch	Relative reduction
White	6.58	1.53	76.44%
Pink	4.86	1.14	76.54%
Tank	0.88	0.68	22.72%
Military Vehicle	0.27	0.22	18.51%
Jet Cockpit	18.11	1.93	89.34%
HFchannel	3.67	1.9	48.22%
F16	2.93	0.9	69.28%
Factory	1.88	2.23	-18.61%
Car	0.11	0.04	63.63%
Machine Gun	40.64	34.6	14.86%
Babble	14.56	18.59	-27.67%

Table 2.1: Total misclassification errors (in percent) for using the periodogram and the Bartlett-Welch method (M=20) in estimating LTSV measure. R is chosen to be 30.

gun noise, the histogram of $\log_{10}(\mathcal{L}_N)$ shows a bimodal nature (Fig. 2.4(a)) when the Bartlett-Welch method is used. This is due to the fact that machine-gun noise is composed of mainly two different signals - the sound of the gun firing and the silence in between firing. When R consecutive frames belong to silence they yield a very small value of $\mathcal{L}_N$ but, when R frames include portions of the impulsive sound of firing (nonstationary event), the value of $\mathcal{L}_N$ becomes high. This creates a hump in the histogram exactly where the main hump of the histogram of $\mathcal{L}_{S+N}$ is (Fig. 2.4(a)). This causes a considerable amount of misclassification error.

A similar observation can be made for the case of speech babble noise. As the noise is speech-like, it is nonstationary and causes similar values of $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$, resulting in significant overlap between the histograms of $\log_{10}(\mathcal{L}_N)$ and $\log_{10}(\mathcal{L}_{S+N})$ (Fig. 2.4(b)). Due to this, a large misclassification error is obtained.

From the simulations of non-stationary noise cases, we found that the mean LTSV of the non-stationary noise is higher than that of stationary noise. Except for machine gun noise, the mean LTSV of all noises is lower than that of speech. Thus, the LTSV

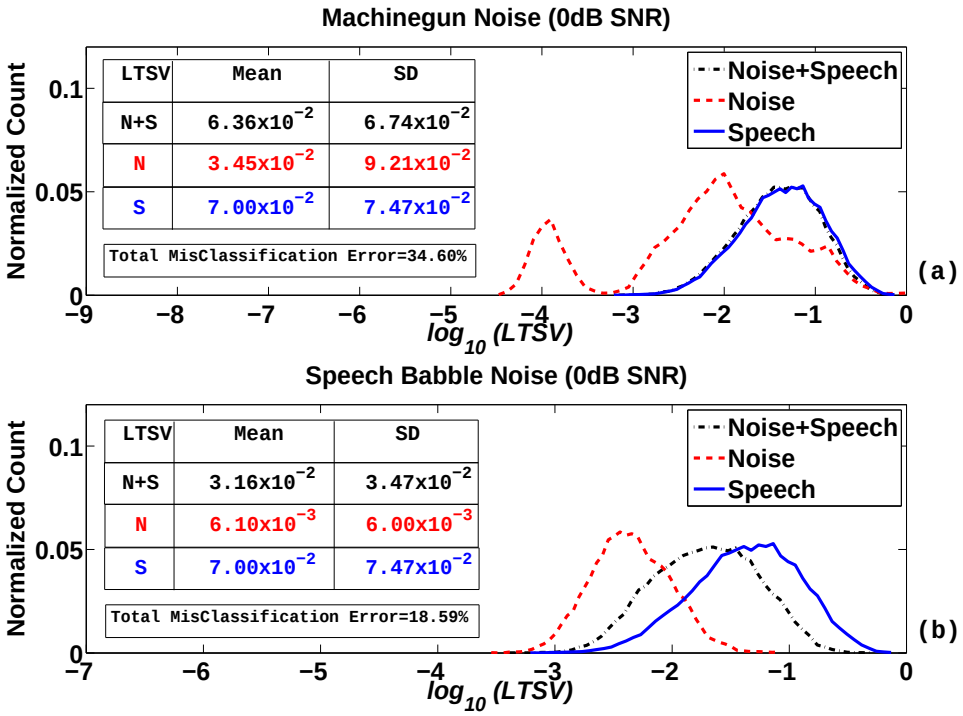

Figure 2.4: *Histogram of the logarithmic LTSV* ($\log_{10}(LTSV)$) *measure using the Bartlett-Welch spectral estimates for (a) machine gun noise and speech in additive machine gun noise (0dB), and (b) babble noise and speech in additive babble noise (0dB). The tables in the figures show the sample mean and sample SD of the realizations of* $\mathcal{L}_N$, $\mathcal{L}_{S+N}$, *and* $\mathcal{L}_S$.

measure reflects the degree of non-stationarity or variability in the signal, and the signal variability in speech is more than that in all noises considered here, except machine gun noise. When the degree of non-stationarity of noise is similar to that of noisy speech as measured by LTSV, noise and noisy speech can not be distinguished effectively. This happens for machine gun and babble noise, resulting in high misclassification errors (Table 2.1).

2.1.5 Selection of $\{\omega_k\}_{k=1}^{K}$, R and M

2.1.5.1 Selection of $\{\omega_k\}_{k=1}^{K}$

From the analysis of LTSV for different noises, we realize that the higher the SNR, the better the separation between the histograms of $\mathcal{L}_{S+N}$ and $\mathcal{L}_N$. Thus, for a better discrimination between $\mathcal{L}_{S+N}$ and $\mathcal{L}_N$, we need to select the frequency values $\{\omega_k\}_{k=1}^{K}$, for which SNR_k is high enough. Computation of SNR_k for various noises reveals that SNR_k is high for frequency values below 4kHz. This is particularly because speech in general is a low pass signal. It is also known that the 500Hz to 4kHz frequency range is crucial for speech intelligibility [3]. Hence we decided to choose ω_k in this range. The exact values of $\{\omega_k\}_{k=1}^{K}$ are determined by the sampling frequency F_s and the order N_{DFT} of discrete Fourier transform (DFT), used to compute the spectral estimate of the observed signal. Thus $K = N_{DFT}\left(\frac{4000-500}{F_s}\right)$. For example, $N_{DFT} = 2048$ and $F_S = 16000$ yield $K = 448$; $\{\omega_k\}_{k=1}^{448}$ are uniformly distributed between 500Hz and 4kHz.

2.1.5.2 Selection of R and M

R and M are parameters used for computing $\mathcal{L}_x(m)$ (see eqn. (2.6) and (2.2)). Our goal is to choose R and M such that the histograms of $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$ are maximally discriminative since the better the discrimination between $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$, the better the final VAD decision. We computed the total misclassification error (sum of two types of detection errors) as a measure of discrimination between the histograms of $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$ for given values of R and M denoted by:

$$\mathcal{M}(R, M) = \text{Speech Detection Error} + \text{Noise Detection Error}$$

We used receiver operating characteristics (ROC) [31] to compute $\mathcal{M}(R, M)$. ROC curve is a plot of speech detection error and non-speech detection error for varying

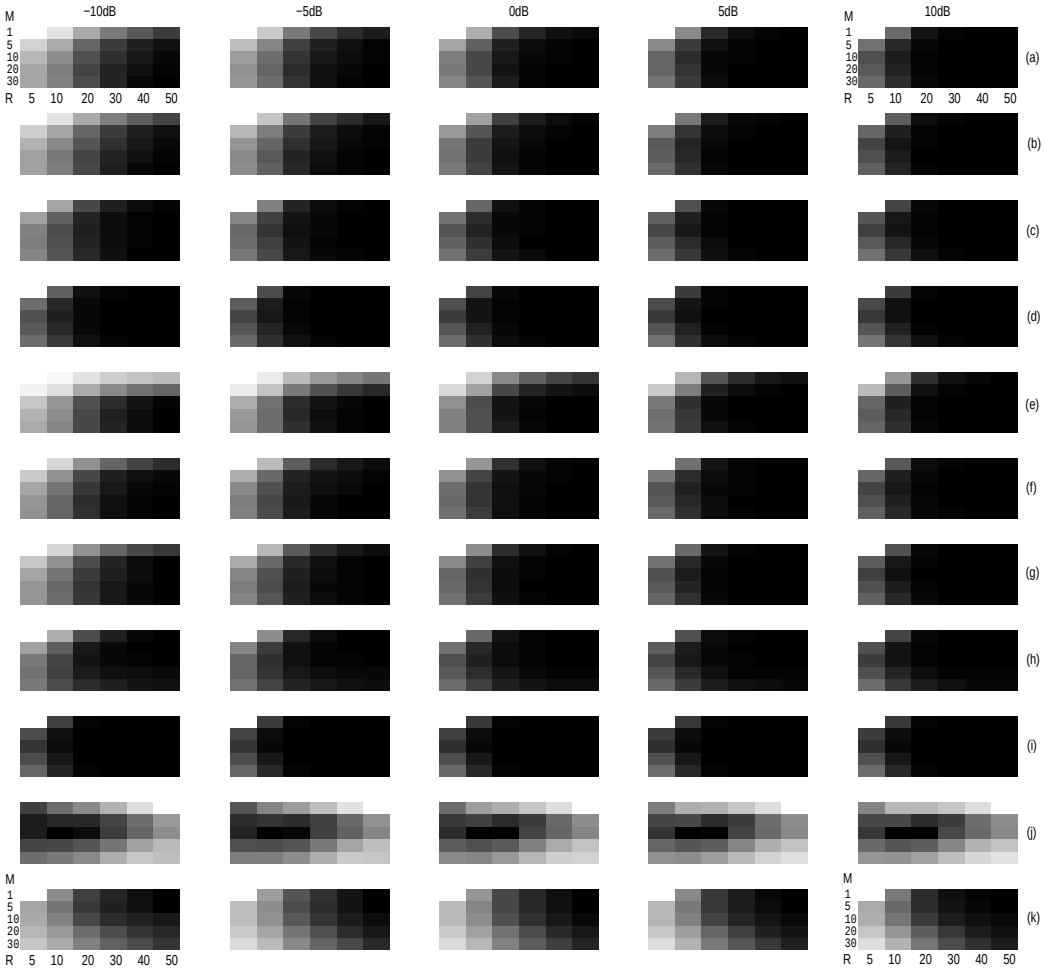

Figure 2.5: *Gray valued representation of $\mathcal{M}(R, M)$ for R=5, 10, 20, 30, 40, 50 and M=1, 5, 10, 20, 30. Darker box indicates lower value: (a) white, (b) pink, (c) tank, (d) military vehicle, (e) jet cockpit, (f) HFchannel, (g) F16 cockpit, (h) factory, (i) car, (j) machine gun, (l) babble noise.*

threshold γ, above which the LTSV measure indicates speech. We chose $\mathcal{M}(R, M)$ and the corresponding threshold for which two types of detection errors are equal, which is known as equal error rate (EER). Eleven different types of noises (as mentioned in Table 2.1) were added to TIMIT training sentences to generate realizations of $\mathcal{L}_{S+N}$.

$\mathcal{L}_N$ were also computed for all these different noises. The Hanning window is used as the short-time window, $w(i)$ (as in eqn. (2.6)), and we chose the following parameter values N_w=320 (corresponds to 20 msec), N_{sh}=$\frac{N_w}{2}$, K=448 and $\{\omega_k\}_{k=1}^{K}$ uniformly distributed between 500 and 4000 Hz (as determined in section 3.1). $\mathcal{M}(R, M)$ are computed for R=5, 10, 20, 30, 40, 50 (corresponding to 50 msec to 500 msec) and M=1, 5, 10, 20, 30 (corresponding to 10 msec to 300 msec). This experiment was systematically performed for 11 types of noises and 5 different SNR conditions, i.e., -10dB, -5dB, 0dB, 5dB, 10dB. The misclassification errors for all combinations of R and M are shown in Fig. 2.5 using gray valued representation. The darker the box, the lower the value of $\mathcal{M}(R, M)$. Rows in Fig. 2.5 correspond to different noise types and columns correspond to different SNRs (as mentioned at the top of the columns). Except for machine gun noise, it was found that for any choice of M, $\mathcal{M}(R, M)$ monotonically decreases with increasing R. However, above R=30 the reduction in $\mathcal{M}(R, M)$ is not significant for most of the noises (which can be seen from insignificant changes in gray values in Fig. 2.5). Also a larger R leads to a larger delay in VAD decision. Hence, we restrict possible values of R up to 30 frames i.e., R=5, 10, 20, 30. We report the combination of R and M corresponding to the minimum value of $\mathcal{M}(R, M)$ in Table 2.2.

We observe that, except for machine gun noise, the best choice of R is 30 (which means that the LTSV is computed over 0.3 sec). For machine gun noise, the best choices of R and M both are found to be 10 frames (0.1 sec) for all SNR conditions. Machinegun noise consists of two types of sounds, namely, gun-shot and silence between gun-shots. For a high value of R like 30 frames, the long analysis window would include both types of sounds increasing the non-stationarity resulting in more overlap between $\mathcal{L}_N$ and $\mathcal{L}_{S+N}$ compared to the case of R=10. From Table 2.2, it is also clear that the spectral averaging over long duration such M=20 (following Bartlett-Welch method) is only advantageous if the noise is not highly non-stationary like factory, machinegun and

Noise Type	SNR specific (R, M)				
	-10dB	-5dB	0dB	5dB	10dB
White	30, 30	30, 30	30, 20	30, 20	30, 30
Pink	30, 30	30, 30	30, 30	30, 30	30, 30
Tank	30, 20	30, 20	30, 20	30, 10	30, 10
Military Vehicle	30, 20	30, 20	30, 10	30, 10	30, 10
Jet Cockpit	30, 20	30, 20	30, 20	30, 20	30, 20
HFchannel	30, 20	30, 20	30, 20	30, 20	30, 20
F16	30, 20	30, 20	30, 20	30, 20	30, 20
Factory	30, 5	30, 5	30, 5	30, 5	30, 5
Car	30, 20	30, 20	30, 10	30, 10	30, 10
Machine gun	10, 10	10, 10	10, 10	10, 10	10, 10
Babble	30, 5	30, 5	30, 1	30, 1	30, 1

Table 2.2: Best choices of R, M for different noises and different SNRs.

babble noise. For most of the noises and SNR conditions, it can be seen that $M > 1$ ($M = 1$ means no averaging, which is equivalent to periodogram method) is found to result in minimum $\mathcal{M}(R, M)$; this means that the low variance spectral estimation (using the Bartlett-Welch method) improves the discrimination between $\mathcal{L}_{S+N}$ and $\mathcal{L}_N$ for most of the noises.

2.1.6 The LTSV-based voice activity detector

The block diagram of the implemented system for VAD using LTSV measure is shown in Fig. 2.6. The input speech signal is first windowed (20 msec length and 10 msec shift) using the Hanning window and the spectrum of the windowed signal is estimated using the Bartlett-Welch method. At the l^{th} window, the LTSV measure $\mathcal{L}_x(l)$ is computed using the previous R frames. $\mathcal{L}_x(l)$ is thresholded to decide whether there was speech in the last R frames. This is denoted by D_l. If $D_l = 0$, it means there is no speech in the last R frames ending at l^{th} frame; if $D_l = 1$, it means there is speech over the last R frames ending at l^{th} frame. However, the final VAD decision is made on every 10 msec

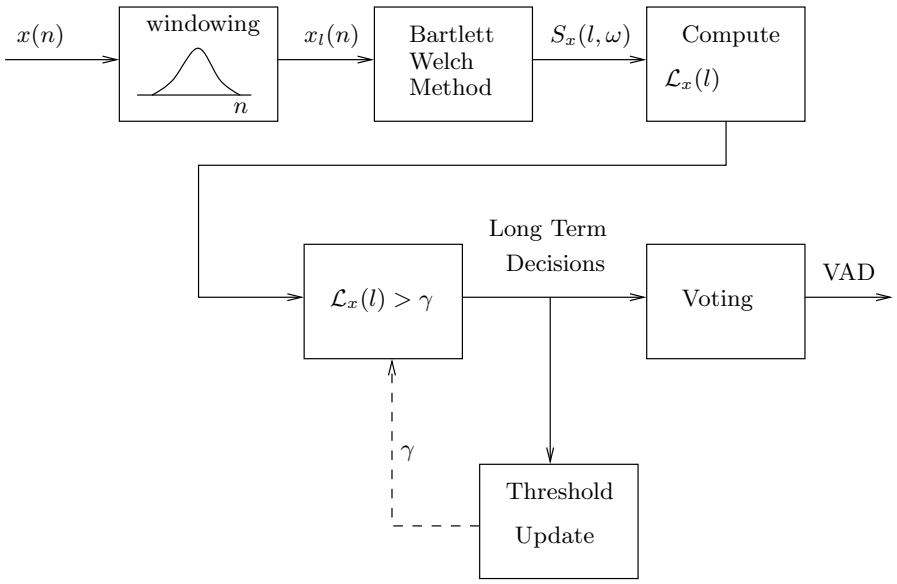

Figure 2.6: *Block diagram of the LTSV-based VAD system*

interval by using a voting scheme[2] as shown in Fig. 2.7. To take a VAD decision on a target 10 msec interval indexed by l, (R+1) decisions D_l, $D_{l+1}, ..., D_{l+R+1}$ are first collected on the long windows, which overlap with the target 10 msec interval. If c% of these decisions are speech, the target 10 msec interval is marked as speech; otherwise it is marked as noise. Experiments on the TIMIT training set shows that a high value of c leads to higher VAD accuracy at 10dB SNR, while a low value of c leads to higher VAD accuracy at -10dB SNR. In our experiment, we chose c=80, which provided the maximum VAD accuracy at 0dB SNR.

Noise and SNR specific best choices of R, M, and threshold γ were obtained in section 3.2. However, to deploy the VAD scheme in practice, we need to update these parameters on the fly according to the type of noise. For our current implementation,

[2]As an alternative to the voting scheme, we also modeled the observed noisy speech as a sequence of segments (each of duration .3 sec with 50% overlap) which are either speech or silence or speech-silence boundary or silence-speech boundary. The probability of transition from one type of segment to another is learned from the training data and used to decode best segment sequence given the LTSV measure for each segment. However, the performance was not significantly improved compared to the voting scheme for all noises.

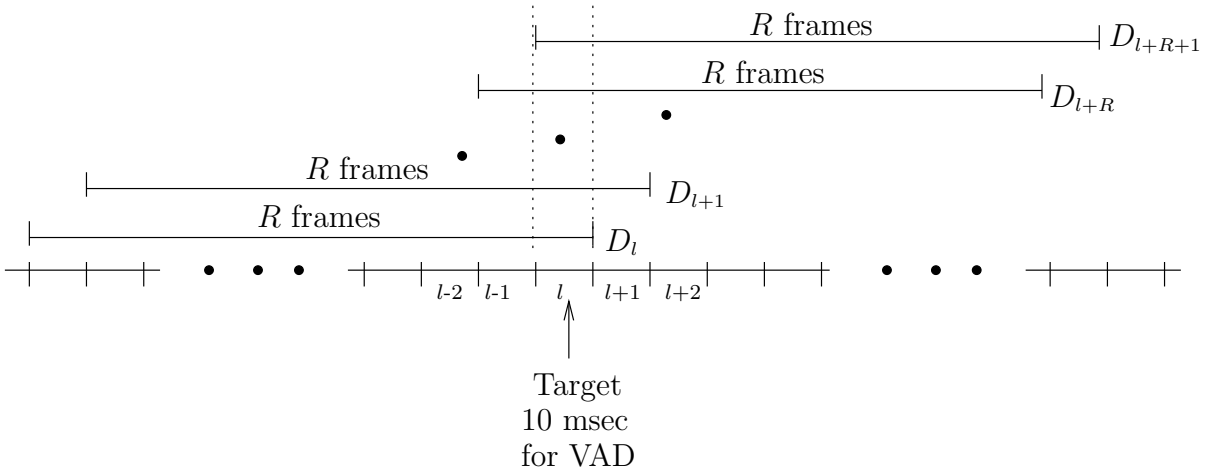

Figure 2.7: *Long windows for voting on a 10 msec interval.*

we have fixed R=30 and M=20, since most of the noises turn out to have minimum misclassification errors for this choice of R and M (Table 2.2). However, a fixed value of γ does not work well for all noises. Hence, we designed an adaptive threshold selection scheme. γ is the threshold for LTSV measure between two classes - noise and noisy speech. To update $\gamma(m)$ at m^{th} frame, we used two buffers $\mathcal{B}_N(m)$ and $\mathcal{B}_{S+N}(m)$. $\mathcal{B}_N(m)$ stored the LTSV measures of the last 100 long-windows, which were decided as containing noise only; similarly, $\mathcal{B}_{S+N}(m)$ stored the LTSV measures of the last 100 long-windows, which were decided as having speech. One hundred long-windows (with 10 msec shift) in each buffer is equivalent to 1 sec. Since we are interested in measuring signal variability over long duration, we assume that the degree of non-stationarity of the signal does not change drastically over 1 sec. $\gamma(m)$ is computed as the convex combination between the minimum of the elements of $\mathcal{B}_{S+N}(m)$ and maximum of the elements of $\mathcal{B}_N(m)$ as follows:

$$\gamma(m) = \alpha \min\left(\mathcal{B}_{S+N}(m)\right) + (1-\alpha)\max\left(\mathcal{B}_N(m)\right) \tag{2.7}$$

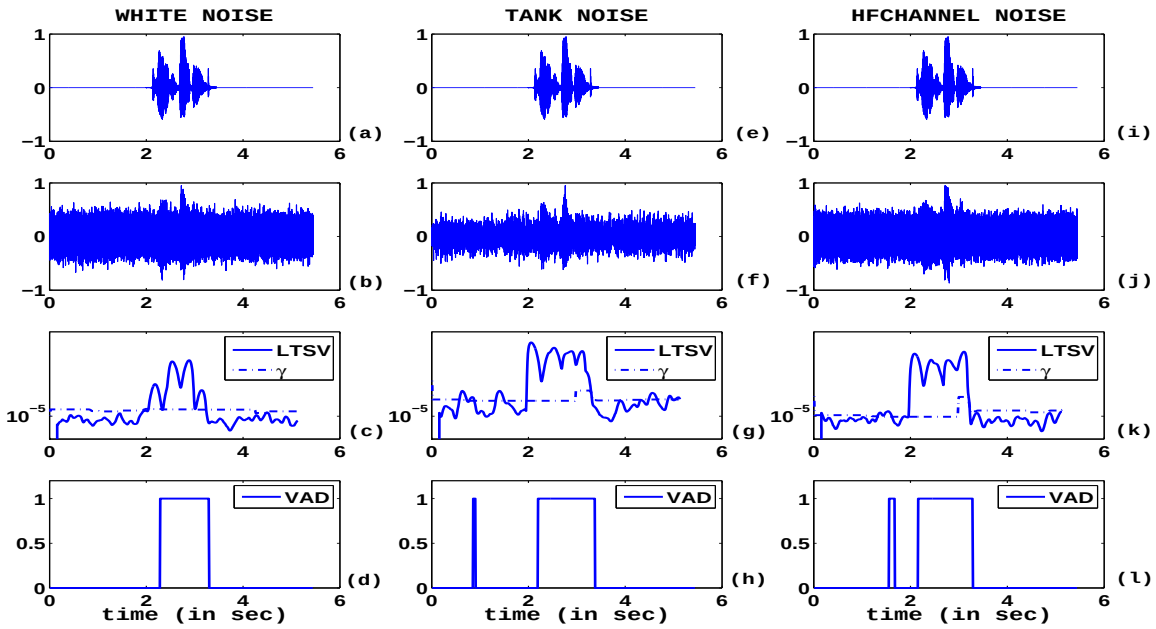

Figure 2.8: *Illustrative example of VAD using LTSV with adaptive threshold on a randomly chosen sentence from TIMIT test set: (a): Clean speech; (b): White Noise added at -10dB SNR; (c): $\mathcal{L}_x(m)$, $\gamma(m)$ computed on (b); (d): VAD decisions on (b); (e)-(h): (a)-(d) repeated for Tank Noise; (i)-(l): (a)-(d) repeated for HFchannel Noise.*

where α is the parameter of the convex combination[3]. We experimentally found that $\alpha = 0.3$ results in maximum accuracy in VAD decisions over the TIMIT training set. To initialize γ, when the LTSV-based VAD scheme starts operating, we proceed in the following way:

We assume that the initial 1 second of the observed signal $x(n)$ is noise only. From this 1 second of $x(n)$, we obtain 100 realizations of $\mathcal{L}_N$. Let μ_N and σ_N^2 be the sample mean and sample variance of these 100 realizations of $\mathcal{L}_N$. We initialize $\gamma = \mu_N + p\sigma_N$, where p is selected from a set of $\{1.5, 2, 2.5, 3, 3.5, 4, 4.5\}$ to obtain the maximum accuracy of VAD decisions on the TIMIT training set. The best choice of p was 3. Since

[3]As an alternative, we also performed experiments by convex combination of the average of $\mathcal{B}_{S+N}(m)$ and $\mathcal{B}_N(m)$, but the performance of VAD decisions was worse compared to that obtained by using eqn. (2.7).

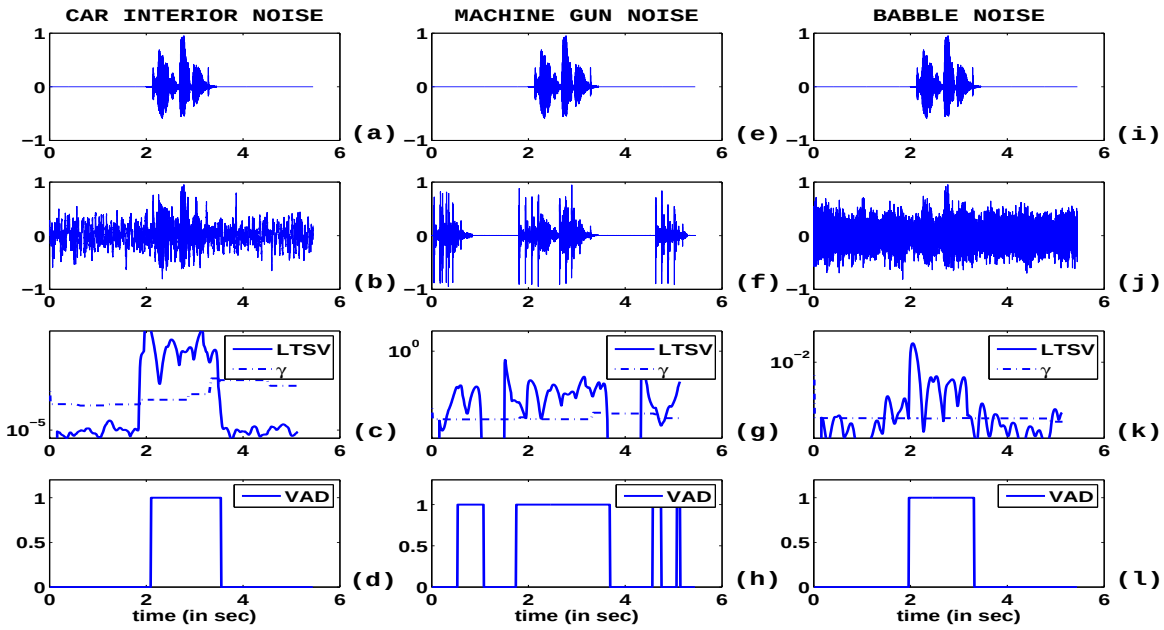

Figure 2.9: *Illustrative example of VAD using LTSV with adaptive threshold on a randomly chosen sentence from TIMIT test set: (a): Clean speech; (b): Car Interior Noise added at -10dB SNR; (c): $\mathcal{L}_x(m)$, $\gamma(m)$ computed on (b); (d): VAD decisions on (b); (e)-(h): (a)-(d) repeated for Machine gun Noise; (i)-(l): (a)-(d) repeated for Babble Noise.*

on average the LTSV of noisy speech is more than that of noise only (as seen in Fig. 2.1-2.4), γ should be more than the mean LTSV of noise (μ_N). The choice of p was done to select the threshold between mean values of the LTSV of noise and noisy speech.

Fig. 2.8 and 2.9 illustrate $\mathcal{L}_x(m)$, $\gamma(m)$ and the VAD decisions for a randomly chosen sentence from TIMIT test set in additive (-10dB SNR) white, tank, HFchannel, car interior, machine gun, and babble noise. It should be noted that before adding noise samples, silence of two seconds has been added in the beginning and at the end of the utterance. This silence padded-speech is shown in Fig. 2.8 and 2.9 (a), (e), (i) to visually compare with the VAD decisions obtained using LTSV and the adaptive threshold scheme for six types of noises. Each column in both Fig. 2.8 and 2.9 corresponds to one type of noise, which is mentioned on top of each column. The second row in both figures shows the speech signal in additive noise at -10dB SNR. The third row in both figures shows $\mathcal{L}_x(m)$ and $\gamma(m)$. Y-axes of these plots are shown in log scale to clearly show the variation in very small values. It can be seen that the threshold γ varies with time as computed by eqn. (2.7). The respective VAD decisions for six noises are shown in the last row of both Fig. 2.8 and 2.9. Value 1 in these plots corresponds to speech and 0 corresponds to noise. Even at -10dB SNR, VAD decisions for additive white, car, and babble noise appear approximately correct from Fig. 2.8 (d), 2.9 (d), and (l) respectively. For machine gun noise, many noise frames are detected as speech. For tank and HFchannel noise also, a few noise frames are detected as speech. However, systematic performance evaluation is required for understanding the accuracy of VAD in various noises.

2.1.7 Evaluation and results

Evaluation of a VAD algorithm can be performed both subjectively and objectively. Subjective evaluation is done through a listening test, in which human subjects detect

VAD errors [80]; on the other hand, objective evaluation is done using an objective criterion, which can be computed numerically. However, subjective evaluation is often not sufficient to examine the VAD performance because listening tests like ABC [80] fail to consider the effect of the false alarm [6]. Hence, we use objective evaluation strategy to report the performance of the proposed VAD algorithm.

We closely follow the testing strategy proposed by Freeman et al [51] and by Beritelli et al [79], in which the labels obtained by the proposed VAD algorithm are compared against known reference labels. This comparison is performed through five different parameters reflecting the VAD performance:

(i) *CORRECT*: Correct decisions made by the VAD.

(ii) *FEC* (*front end clipping*): Clipping due to speech being misclassified as noise in passing from noise to speech activity.

(iii) *MSC* (*mid speech clipping*): Clipping due to speech misclassified as noise during an utterance.

(iv) *OVER* (*carry over*): Noise interpreted as speech in passing from speech activity to noise due to speech information carried over by the LTSV measure.

(v) *NDS* (*noise detected as speech*): Noise interpreted as speech within a silence period.

FEC and *MSC* are indicators of true rejection, while *NDS* and *OVER* are indicators of false acceptance. *CORRECT* parameter indicates the amount of correct decisions made. Thus all four parameters *FEC*, *MSC*, *NDS*, *OVER* should be minimized and the *CORRECT* parameter should be maximized to obtain the best overall system performance.

For VAD evaluation in this work, we used the TIMIT test corpus [30] consisting of 1680 individual speakers of eight different dialects, each speaking 10 phonetically

balanced sentences. Silence of an average duration of 2 sec was added before and after each utterance, and then noise of each category was added at 5 different SNR levels (-10dB, -5dB, 0dB, 5dB, 10dB) to all 1680 sentences. The test set for each noise and SNR thus consisted of 198.44 minutes of noisy speech of which 62.13% was only noise. The noise samples of eleven categories were taken from the NOISEX-92 database. The beginning and end locations of the speech portions of the silence-padded TIMIT sentences were computed using the start time and the end time of the sentence obtained from the TIMIT transcription. The final VAD decisions were computed for every 10 msec interval. Thus, for reference, each 10 msec interval was tagged as speech or noise using the beginning and end of speech. If a 10 msec interval overlapped with speech, it was tagged as speech, and otherwise as noise.

The proposed adaptive-threshold LTSV (we denote this by LTSV-Adapt scheme) based VAD scheme was run followed by the voting scheme to obtain VAD decisions at 10 msec frame level. The noisy TIMIT test sentences were concatenated and presented in a contiguous manner to the LTSV-Adapt VAD scheme so that the threshold could be adapted continuously. In order to do a comparative analysis, the performance of the proposed LTSV-Adapt scheme was compared against three modern standardized VAD schemes. These schemes are the ETSI AMR VADs option 1 & 2 [2] and ITU G.729 AnnexB VAD [44]. The implementations were taken from [1] and [45] respectively. The VAD decisions at every 10 msec obtained by the standard VAD schemes and our proposed VAD scheme were compared to the references, and five different parameters (*CORRECT*, *FEC*, *MSC*, *NDS*, *OVER*) were computed for eleven noises and five SNRs. In addition to the performance of the LTSV-Adapt scheme, we report performance for the case using noise and SNR specific R, M and γ (we denote this by LTSV-opt scheme), assuming that we know the correct noise category and SNR. This was done to analyze how much the VAD performance degrades when the noise information is not available or

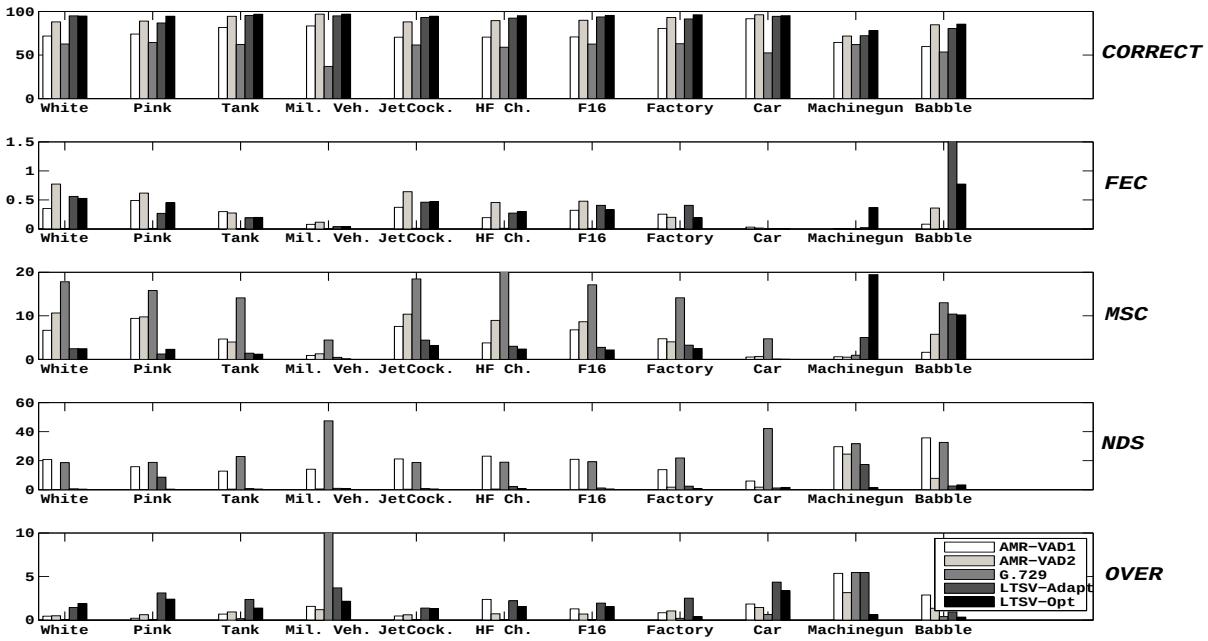

Figure 2.10: *CORRECT, FEC, MSC, NDS and OVER averaged over all SNRs for eleven noises as obtained by five VAD schemes - AMR-VAD1, AMR-VAD2, G.729, LTSV-Adapt scheme and LTSV-opt scheme.*

not estimated. However, for comparing against standard VAD schemes, we used LTSV-Adapt scheme-based VAD decisions. Fig. 2.10 shows five different scores (*CORRECT*, *FEC*, *MSC*, *NDS*, *OVER*), averaged over 5 SNRs for each noise, computed for AMR-VAD1, AMR-VAD2, G.729, LTSV-Adapt, and LTSV-opt schemes. Fig. 2.11 shows the same result for -10dB SNR.

We observe a consistent reduction in average *CORRECT* score from LTSV-opt scheme to LTSV-Adapt scheme for all noises (Fig. 2.10). The significant reduction happens for speech babble (from 85.3% for LTSV-opt scheme to 80.3% for LTSV-Adapt scheme) and for machine gun noise (from 78.0% for LTSV-opt scheme to 72.0% for LTSV-Adapt scheme). While machine gun noise is impulsive in nature, speech babble

is speech-like and, hence, the best choices of R and M for these noises are different (see Table 2.2) compared to the R=30 and M=20 combination, which is used in LTSV-Adapt scheme. This mismatch in R and M causes a significant difference in the *CORRECT* score between the LTSV-opt scheme and the LTSV-Adapt scheme. A suitable noise categorization scheme prior to LTSV-opt scheme can improve the VAD performance compared to LTSV-Adapt scheme. From Fig. 2.10, it is clear that in terms of the *CORRECT* score, AMR-VAD2 is the best among all three standard VAD schemes considered here. Hence, the LTSV-Adapt scheme is compared with the AMR-VAD2 among three standard VAD schemes. We see that on an average, the LTSV-Adapt scheme is better than the AMR-VAD2 in terms of *CORRECT* score for white (6.89%), tank (0.86%), jet cockpit (5.02%), HFchannel (2.88%), F16 cockpit (3.86%), and machine gun (0.35%), and worse for pink (2.14%), military vehicle (2.01%), factory (1.61%), car interior (1.73%), and babble (4.38%) noises. The percentage in the bracket indicates the absolute *CORRECT* score by which one scheme is better than the other. LTSV-Adapt scheme has a smaller *MSC* score compared to that of AMR-VAD2 for white (8.19%), pink (8.52%), tank (2.57%), military vehicle (0.82%), jet cockpit (5.95%), HFchannel (5.92%), F16 cockpit (5.88%), Factory Noise (0.75%), and car interior (0.61%) and a larger *MSC* for machine gun (4.52%) and babble noise (4.63%). The percentage in the bracket indicates the absolute *MSC* score by which one scheme is better (has lower *MSC*) than the other. Thus, AMR-VAD2 has a larger *MSC* score compared to the LTSV-Adapt scheme for all noises except machine gun and babble noise. This means, on an average, AMR-VAD2 loses more speech frames compared to the LTSV-Adapt scheme. For babble noise, the *CORRECT* score of AMR-VAD2 is greater than that of LTSV-Adapt scheme due to the fact that we use M=20, which is not the best choice for speech babble as shown in Table 2.2. Speech babble being non-stationary noise, long temporal smoothing does not help. The *OVER* score of the LTSV-Adapt scheme for

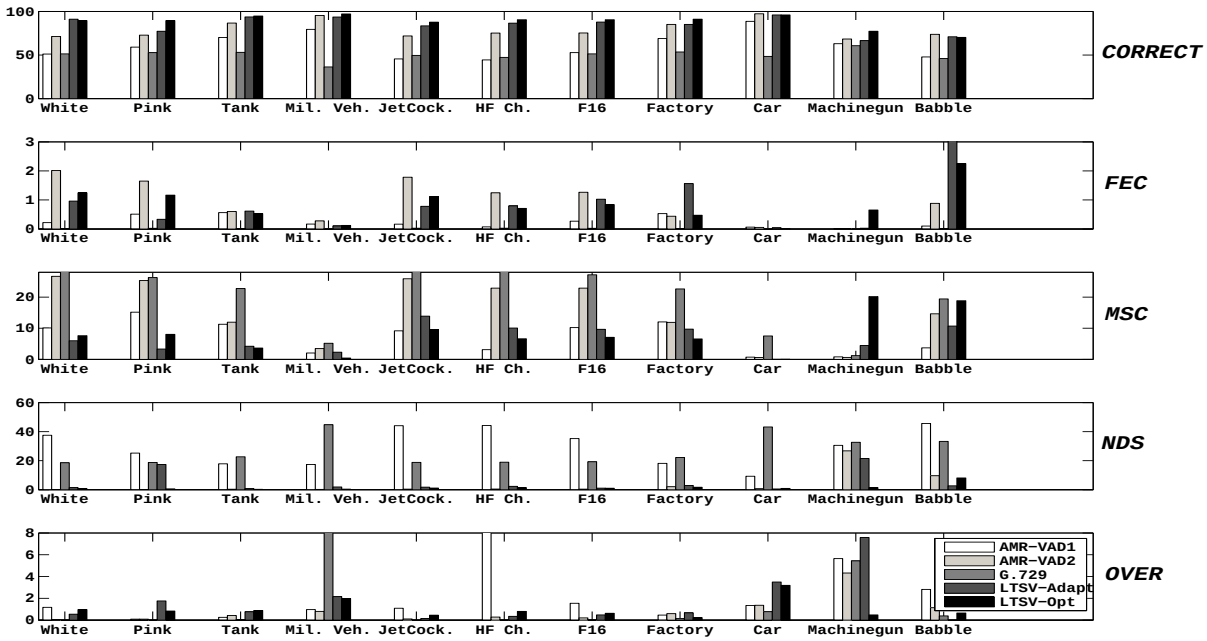

Figure 2.11: *CORRECT, FEC, MSC, NDS and OVER at -10dB SNR for eleven noises as obtained by five VAD schemes - AMR-VAD1, AMR-VAD2, G.729, LTSV-Adapt scheme and LTSV-opt scheme.*

additive car interior noise is more than that of AMR-VAD2. This happens for pink, military vehicle, factory, and babble noise, too. Higher values of *OVER* for these noises result in a lower value of the *CORRECT* score of the LTSV-Adapt scheme compared to that of AMR-VAD2. High value of *OVER* implies that noise frames at the speech-to-noise boundary are detected as speech. Depending on the application, such errors can be tolerable compared to high *MSC* and high *NDS*. High *MSC* is harmful for any application since high *MSC* implies that speech frames are decided as noise frames.

It should be noted that in the LTSV-Adapt scheme, we are neither estimating the SNR of the observed signal nor estimating the type of noise. This is an SNR independent scheme; however, the LTSV-Adapt scheme performs consistently well in all SNRs. In

particular, from Fig. 2.11, we observe that at -10dB SNR, the LTSV-Adapt scheme has a higher *CORRECT* score than that of AMR-VAD2 for white (19.88%), pink (4.43%), tank (6.95%), jet cockpit (11.6%), HFchannel (11.36%), F16 cockpit noise (12.48%), and factory noise (0.07%) and lower for military vehicle (1.72%), car interior (1.37%), machine gun (1.88%), and babble noise (2.76%). These are noises where we have mismatch between the fixed R and M used for the LTSV-Adapt scheme with the best R and M as indicated by Table 2.2. Also at -10dB SNR, *MSC* of the LTSV-Adapt scheme is lower than that of AMR-VAD2 for white (20.7%), pink (22.01%), tank (7.73%), military vehicle (1.16%), jet cockpit (12.02%), HFchannel (12.83%), F16 cockpit (13.23%), Factory Noise (2.18%), car interior (0.58%), and babble noise (3.95%) and greater for machine gun (3.82%) noise. Compared to AMR-VAD2, the LTSV-Adapt scheme has lower *NDS*, too. All these imply that the LTSV-Adapt scheme has smaller speech frame loss compared to AMR-VAD2 at -10dB SNR, and hence is robust in low SNR.

2.1.8 Conclusions

We presented a novel long-term signal variability (LTSV) based voice activity detection method. Properties of this new long-term variability measure were discussed theoretically and experimentally. Through extensive experiments, we show that the accuracy of the LTSV based VAD averaged over all noises and all SNRs is 92.95% as compared to 87.18% obtained by the best available commercial AMR-VAD2. Similarly, at -10dB SNR, the accuracies are 88.49% and 79.30% respectively, demonstrating the robustness of LTSV feature for VAD at low SNR. While the energy-based features for VAD are affected by signal scaling, the proposed long-term variability is not. It has also been found that for non-stationary noises, which have similar LTSV measure as that of speech, the proposed VAD scheme fails to distinguish speech from noise with good accuracy. However, additional modules such as noise category recognition might help improve the

result by allowing for noise-specific solutions and improve the VAD performance. If we have knowledge of the background noise in any application or if we can estimate the category of noise and accordingly choose the R, M and γ for minimum misclassification error on the training set, we expect to achieve the performance of LTSV-opt scheme, which is better than that of LTSV-Adapt scheme. We also observed that the optimum choice of c varies with SNR. Thus, adaptively changing c by estimating the SNR of the observed signal can improve the VAD performances. Also, a choice of low value of c improves *FEC* score while increases *OVER* score at high SNR. On the other hand, a high value of c reduces the *OVER* score while increases *FEC* score at low SNR. Thus, the choice of c should be tuned considering the trade-off between *FEC* and *OVER* scores. These are part of our future works.

To improve upon the LTSV measure, we have explored using mean entropy ($\overline{\xi^x(m)}$) for VAD. Theoretically, it is easy to prove that mean entropy for noise $\geq$ mean entropy of S+N. But, in practice, their histograms overlap more than those of their variance. We observed that the correlation between mean and variance of the LTSV feature is high (in the range of -0.6 to -0.9); hence, using mean LTSV as an additional feature, we did not obtain any significant improvement in VAD performance. We also performed experiments with additional features like subband energy, subband LTSV, derivatives of LTSV, and with choices of different frequency bands. In some cases, additional features provided improvements for some noises. Thus, in noise-specific applications, these additional features could be useful. Also, it can be seen that we have not used the usual hangover scheme as done in frame based VAD schemes [6]. This is because our approach inherently takes a long-term context through variability measure. So, there is no additional need of the hangover scheme.

One advantage of using LTSV for VAD is that there is no need for explicit SNR estimation. At the same time, it should be noted that, depending on the choice of the

longer window length, any VAD related application is expected to suffer a delay equal to the duration of the window. Thus, a trade-off between the delay and the robustness of VAD, particularly in low SNR, should be examined carefully before using LTSV-based VAD scheme in a specific application.

2.2 Multi-band long-term signal variability features for robust voice activity detection

[100]

In this paper, we propose robust features for the problem of voice activity detection (VAD). In particular, we extend the long term signal variability (LTSV) feature to accommodate multiple spectral bands. The motivation of the multi-band approach stems from the non-uniform frequency scale of speech phonemes and noise characteristics. Our analysis shows that the multi-band approach offers advantages over the single band LTSV for voice activity detection. In terms of classification accuracy, we show 0.3%-61.2% relative improvement over the best accuracy of the baselines considered for 7 out 8 different noisy channels. Experimental results, and error analysis, are reported on the DARPA RATS corpora of noisy speech.

2.2.1 Introduction

Voice activity detection (VAD) is the task of classifying an acoustic signal stream into speech and non-speech segments. We define a speech segment as a part of the input signal that contains the speech of interest, regardless of the language that is used, possibly along with some environment or transmission channel noise. Non-speech segments are the signal segments containing noise but where the target speech is not present. Manual or automatic speech segment boundaries are necessary for many speech processing systems. In large-scale or real-time systems, it is neither economical nor feasible to employ human labor (including crowd-sourcing techniques) to obtain the speech boundaries as a key first step. Thus, the fundamental nature of the problem has positioned VAD

as a crucial preprocessing tool to a wide range of speech applications, including automatic speech recognition, language identification, spoken dialog systems and emotion recognition.

Due to the critical role of VAD in numerous applications, researchers have focused on the problem since the early days of speech processing. While some VAD approaches have shown robust results using advanced back-end techniques and multiple system fusion [65], the nature of VAD and diversity of environmental sounds suggests the need of robust VAD front-ends. Various signal features have been proposed for separating speech and non-speech segments in the literature. Taking into account short-term information ranging from 10ms to 40ms, various researchers [41, 86, 35] have proposed energy-based features. In addition to energy features, researchers have used zero-crossing rate [16], wavelet-based features [22], correlation coefficients [5] and negentropy [69, 75] which has been shown to perform well in low SNR environments. Other works have used long-term features in the range of 50-100ms [47] and above 150ms [68]. Long-term features have been shown to perform well on noisy speech conditions under a variety of environmental noises. Notably, they offer theoretical advantages for stationary noise [68] and capture information that short-term features lack.

The long-term features proposed in the past focus on extracting information from a two-dimensional (2-D) time-frequency window. Limiting the extracted feature information from 2-D spectro-temporal windows fails to capture some useful auditory spectrum properties of speech. It is well known that the human auditory system utilizes a multi-resolution frequency analysis with non-linear frequency tiling reflected in the Mel-scale [91] representation of audio signals. Mel-scale provides an empirical frequency resolution that approximates the frequency resolution of the human auditory system. Inspired by this property of the human auditory system and the fact that the discrimination of

various noise types can be enhanced at certain different frequency levels, we expand the LTSV feature proposed in [68] to use multiple spectral resolution.

We compare the proposed approach with two baselines: the MFCC [107] features and the single-band (1-band) long-term signal variability (LTSV) [68] and show significant performance gains. Unlike [94] where standard MFCC features have been used for this task and experimented with various back-end systems, we use a fixed back-end and focus only on comparing features for the VAD task using a K-Nearest Neighbor (K-NN) [32] classifier. We perform our experiments on the DARPA RATS data [104] for which an off-line batch processing is required.

2.2.2 Proposed VAD Features

In this subsection, we describe the proposed multi-band extension of the LTSV feature introduced in [68]. LTSV has been shown to have good discriminative properties for the VAD task especially in high SNR noise conditions. We try to exploit this property by capturing dynamic information of various spectral bands. For example, impulsive noise which degrades the performance of LTSV features is often limited to certain band regions in the spectrum. The aim of this work is to investigate the use of a multi-band approach to capture speech variability across different bands. Also, speech variability might be exemplified in different regions for different phonemes. Thus, a multi-band approach could have advantages over the 1-band LTSV.

2.2.2.1 Frequency smoothing

The low pass filtering process is important for the LTSV family of features because it removes the high frequency noise on the spectrogram. Also, it was shown that it improves robustness in stationary noise [68], such as white noise.

Let $S(\hat{f}, j)$ represent the spectrogram, where $\hat{f}$ is the frequency bin of interest and j is j^{th} frame. As in [68], we smooth S using a simple moving average of window of size M (assumed to contain even number of samples for our notation) as follows:

$$S_M\left(\hat{f}, j\right) = \frac{1}{M} \sum_{k=j-\frac{M}{2}}^{j+\frac{M}{2}-1} S\left(\hat{f}, k\right) \tag{2.8}$$

2.2.2.2 Multi-Band LTSV

In order to define multiple bands, we need a parameterization to set the warping of the spectral bands. For this purpose, we use the warping function from the warped discrete Fourier transform [8] which is defined as:

$$F_W(f, \alpha) = \frac{1}{\pi} \arctan\left(\frac{1+\alpha}{1-\alpha} \tan(2\pi f)\right) \tag{2.9}$$

where f represents the frequency to be warped starting from uniform bands and α is the warping factor and takes values in the range $[-1, 1]$. A warping factor of -1 implies a high resolution for high frequencies and, of 1 implies a high resolution for low frequencies. A warping factor of 0 results in uniform bands. To define the multi-resolution LTSV, we first define the normalized spectrogram across time over an analysis window of R frames as:

$$S_R\left(\hat{f}, j\right) = \frac{S_M\left(\hat{f}, j\right)}{\sum_{k=j-\frac{R}{2}}^{j+\frac{R}{2}-1} S_M\left(\hat{f}, k\right)} \tag{2.10}$$

Hence, we define the multi-band LTSV feature of window size R and warping factor α at the i^{th} frequency band and j^{th} frame as:

$$L_i(\alpha, R, j) = V_{\hat{f} \in F_i} \left(\sum_{k=j-\frac{R}{2}}^{j+\frac{R}{2}-1} S_R\left(\hat{f}, k\right) \log\left(S_R\left(\hat{f}, k\right)\right) \right) \tag{2.11}$$

V is the variance function defined as:

$$V_{f \in F}\left(a(f)\right) = \frac{1}{|F|} \sum_{f \in F} \left(a(f) - \frac{1}{|F|} \sum_{f \in F} a(f) \right)^2$$

where $|F|$ is the cardinality of set F. The set F_i includes the frequencies $F_W(f, \alpha)$ for $f \in \left[\frac{N_s * (i-1)}{2N} \cdots \frac{N_s * i}{2N} \right]$, N is the number of bands to be included and N_s denotes the sampling frequency.

2.2.3 Experimental setup

To compare across the various features, we used a K-NN classifier for all the experiments. We used 70 hours of data from the RATS[4] corpus (dev1_v2 set) for training and 11 hours for testing for each channel; the RATS data comprises of speech data transmitted through eight different channels (A through H), resulting in varying signal qualities and SNRs. To optimize the parameters, we used a small set of 1 hour for training and a 1 hour development set for each channel. As a post-processing step, we applied a median filter to the output of the classifier to impose continuity on the local detection based output. For each experiment, we searched for the optimal K-NN neighborhood size $K \in [1 \cdots 100]$ and the optimal median filter length for various windows sizes ($[100, 300, 500, 700, 900]$ms). This optimization procedure was performed for each

[4] www.darpa.mil/Our_Work/I2O/Programs/Robust_Automatic_Transcription_of_Speech_(RATS).aspx

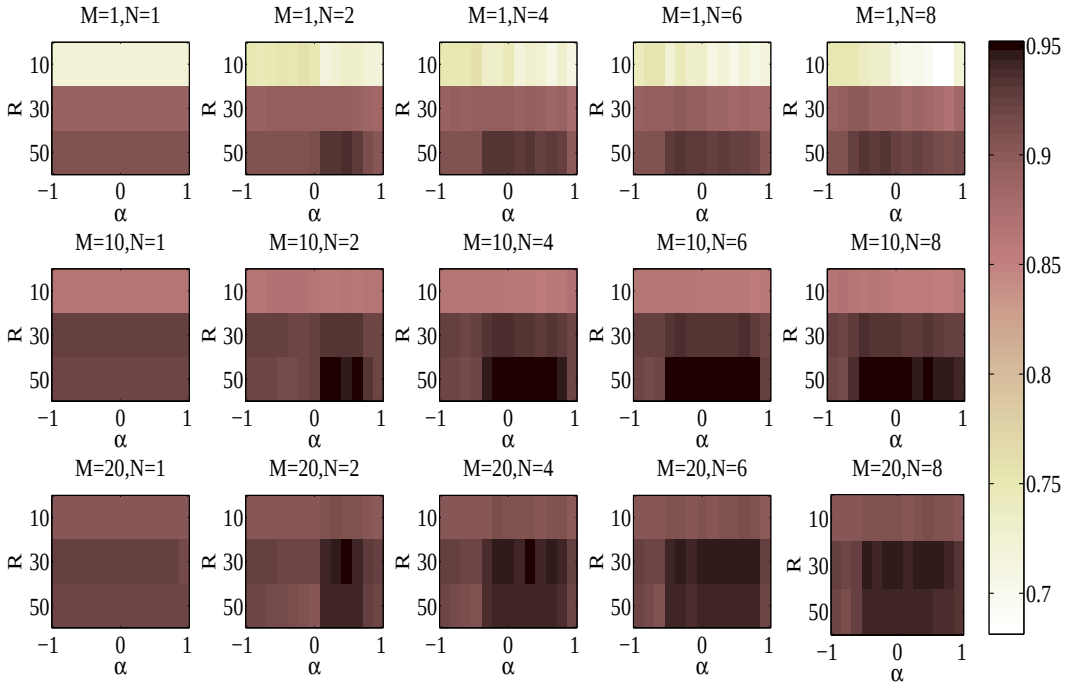

Figure 2.12: This figure shows the VAD frame accuracy for the development set of channel A for various parameters of the multi-band LTSV. R represents the analysis window length, M the frequency smoothing, α the warping factor and N the number of filters. The bar on the right represents the frame accuracy. This figure indicates that for channel A increasing the number of bands (N) improves the accuracy. Also, indicates that smoothing ($M \geq 100$) and analysis window (R) are crucial parameters for the multi-band LTSV as observed in the original LTSV [68].

channel separately. We set as baselines the MFCC and 1-band LTSV features and compare against the proposed multi-band LTSV. We experimented with all A-H channels included in the RATS data set.

The test set results have been generated using the DARPA speech activity detection evaluation scheme [39] which computes the error at the frame level and considers the following:

- Does not score 200ms from the start/end speech annotation towards the speech frames.
- Does not score 500ms from the start/end speech annotation towards the non-speech frames.
- Converts to non-speech, speech segments less than 300ms.
- Converts to speech, non-speech segments less than 700ms.

2.2.4 Emprical selection of algorithm parameters

In this subsection, we describe the pilot experiments we performed to choose the optimal parameters for the LTSV-based features. Fig. 2.12 shows the accuracy for channel A for all the parameters used to fine-tune the optimal LTSV features. To select the set of parameters, we run a grid search over a range of parameters for each channel separately. In particular, we experimented with 15 different warping factors uniformly in the range $[-0.95 \cdots 0.95]$. We also computed the spectrogram smoothing parameter M as defined in Sec. 2.2.2.1. $M = 1$ corresponds to no smoothing whereas $M = [100, 200]$ correspond to smoothing of 100 and 200ms, respectively. In addition, we searched different analysis window sizes $R = [100, 300, 500]$ms. The final parameter we experimented with was the number of bands $N = [1, 2, 4, 6, 8]$. Fig. 2.12 shows that for channel A the optimal number of filters is 6. The optimal values consist of warping factor $\alpha = 0.3$ with smoothing $M = 200$ms and analysis window $R = 300$ms. Channel A contains bandpass speech in the range 400-4000Hz. This might be one of the reasons a warping factor of 0.3 has been chosen for this channel. Smoothing M and analysis window R depend on how fast the noise varies with time. Very slow varying noise types, i.e. stationary noises can afford to have high values for M and R. However, if impulsive noises are of interest, smaller windows are preferable. The warping factor α depends on which

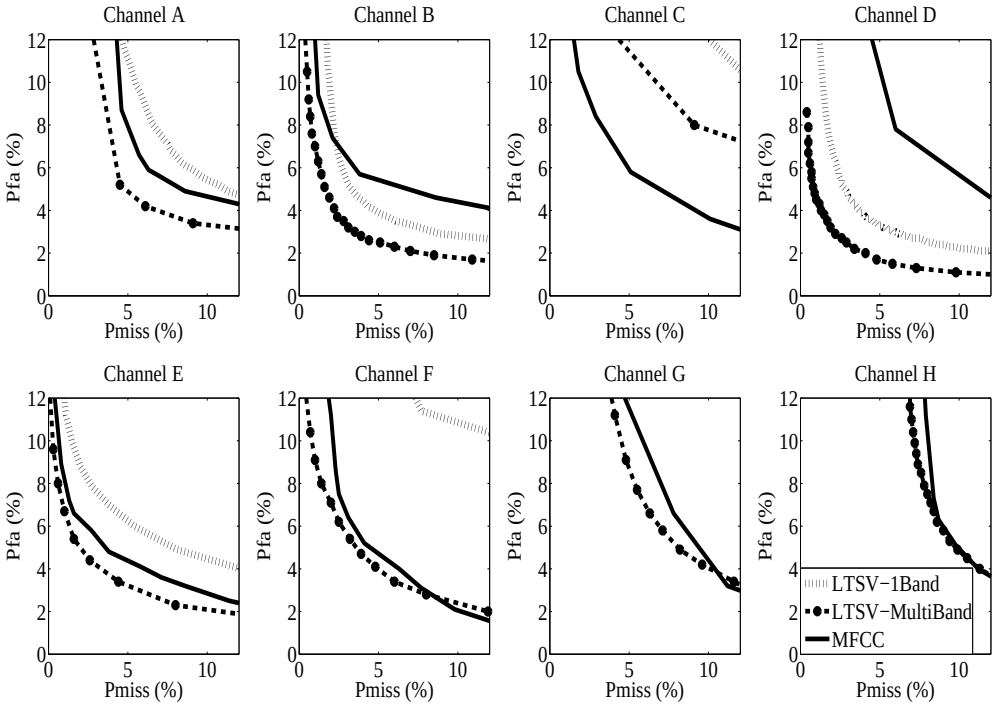

Figure 2.13: This figure shows the ROC curve of Pfa vs Pmiss for channels A-H of the multi-Band LTSV (LTSV-MultiBand) and the two baselines (1-band LTSV and MFCC). For channels G and H the 1-band LTSV ROCs are out of the boundaries of the plots, hence they do not appear in the figure. The same legend applies to all subfigures.

frequency bands have prominent formants. For instance, if strong formants appear in low frequency ranges, values around 0.6 are preferable (i.e. close to Mel-scale).

For all pilot experiments, we have optimized K of K-NN using the Mahalanobis distance [58] and the median filter length. We have observed that a median filter of 700-900ms is best for most of the experiments. This suggests that extracting features with longer window lengths can further improve the accuracy.

2.2.5 Results and discussion

Fig. 2.13 shows the Receiver Operating Characteristics (ROC) curve between false alarm probability (Pfa) and miss probability (Pmiss) for the eight different channels of noisy speech and noise data considered. Channels A-D contain stationary channel noise but non-stationary environmental noise which imposes challenges for the 1-band LTSV. Channels G-H consist of varying channel and environmental noise, causing poor performance for the 1-band LTSV features with equal error rate (EER) exceeding 12%.

Poor classification results due to the non-stationarity of the noise can be improved using multi-band LTSV features. Multi-band LTSV features achieve the best performance compared to both baselines, except for channel C where MFCC has the lowest EER. In addition, we did an error analysis of individual channels to investigate the cases for which the algorithm fails to classify correctly the two classes. On the miss side at the equal error rate (EER), a common error for all channels was due to the presence of filler words, laughter etc. Also, for channels D and E almost half of the errors contributing to the miss rate were due to background/degraded speech. Filler words have slower varying spectral characteristics than verbal speech. If noise has higher spectral variability than filler words, the LTSV features fail to discriminate them.

On the false alarm side, the error analysis at EER reveals that there were a variety of errors including background/robotic speech, filler words and kids background speech/cry. Such errors are expected since background speech shares the spectral variability characteristics of foreground speech; in fact, the classification of background speech by annotators is often based on semantics rather than low-level signal characteristics.

Apart from the speech-like sounds where the multi-band LTSV shows degraded performance, there are non-speech sounds that the multi-band LTSV failed to classify. In particular, false alarms (FA) in channels A,B,D,E and H have been associated with

constant tones appearing at different frequencies over time and impulsive noises at varying frequencies. FA in channel C are composed of noise with spectral variability appearing at different frequencies with one strong frequency component up to 2500Hz and bandwidth greater than the speech formants bandwidth. The limited frequency discriminability (although improved in the multi-band version) is an inherent weakness of the LTSV features. Thus, for channel C, LTSVs performed very poorly, even worse than MFCC. FAs of multi-band LTSV in channel G stem from the variability of the channel and not the environmental noise.

Overall, the multi-band LTSV, performs better than the two baselines considered: the 1-band LTSV and MFCC. From the error analysis, we found that the multi-band LTSV not only retains the discrimination of the 1-band LTSV for stationary noises but also improves discrimination in noise environments with variability, even in impulsive noise cases where the 1-band LTSV fails. However, the multi-band LTSV fails to discriminate impulsive noises appearing at different frequencies over time. For speech miss errors, filler words/laughter are challenging for LTSV due to their lower spectral variability over long time relative to the actual speech. Finally, besides channel C where MFCC gives the best performance, the multi-band LTSV gives the best accuracy showing the benefits of capturing additional information using a multi-resolution LTSV approach.

2.2.6 Conclusion and future work

In this paper, we extended the LTSV [68] feature to multiple spectral bands for the voice activity detection (VAD) task. We found that the multi-band approach improves the performance in different noise conditions including impulsive noise cases in which the 1-band LTSV suffers. We compare the multi-band approach against two baselines: the 1-band LTSV and MFCC features and we found that we gain significantly in performance for 7 out of the 8 channels tested.

In future work, we plan to include delta features along with additional long-term and short-term features that capture the information the multi-band LTSV fails to capture. One aspect that needs further investigation is how to improve the accuracy at the fine-grained boundaries of the decision due to the long-term nature of the feature set. Also, it would be interesting to explore the potential of these features with various machine learning algorithms including deep belief networks.

2.3 Robust voice activity detection in stereo recording with crosstalk [67]

Crosstalk in a stereo recording occurs when the speech from one participant is leaked into the close-talking microphones of the other participants. This crosstalk causes degradation of the voice activity detection (VAD) performance on individual channels, in spite of the strength of the crosstalk signal being lower than that of the participant's speech. To address this problem, we first detect speech using a standard VAD scheme on the merged signal obtained by adding the signals from two channels and then determine the target channel using a channel selection scheme. Although VAD is performed on a short-term frame basis, we found that the channel selection performance improves with long-term signal information. Experiments using stereo recordings of real conversations demonstrate that the VAD accuracy averaged over both channels improves by 22% (absolute) indicating the robustness of the proposed approach to crosstalk compared to the single channel VAD scheme.

2.3.1 Introduction

Detection of speech and non-speech portions from conversations involving multiple speakers such as multi-channel audio recorded at meetings has become a critical initial

step for many spoken language processing applications including discourse analysis, turn taking or speaker interaction pattern analysis. With the increasing amount of dyadic or small group interaction data being generated requiring analysis, manual marking of speech segments becomes equally expensive and time consuming. There have been previous efforts in automatically detecting speech segments using multiple-state hidden markov model (HMM), where states either represent 'speech' and 'nonspeech' [73, 72] or correspond to 'speech', 'overlapped speech', 'crosstalk', and 'silence' [105]. One of the frequently faced problems is that the recorded signal in each channel of a multi-speaker (e.g., meeting) corpus often contains crosstalk [23] – the occurrence of the signal in the channel from sources other than the intended primary source for that channel. There can also be time segments where simultaneous voices from two or more participants get recorded resulting in overlapped speech. Modeling these various events using HMM states requires manually tagged training data and is also vulnerable to channel variation.

Laskowski et. al. [50] proposed an efficient cross-correlation-based voice activity detection (VAD) scheme, which requires no prior training and executes in one fifth real time. Fast multichannel VAD schemes without prior training yet robust to noise or channel variation are desirable since most of the data processing and analysis following VAD such as speech recognition are generally more time-consuming. In this paper, we consider only stereo recording and propose a simple but robust VAD scheme for two channel (stereo) speech signal with crosstalk. The proposed scheme consists of a standard VAD on the merged signal from two channels followed by a channel selection strategy. The goal of the channel selection is to determine the target channel at every short-time frame when voice activity is detected in the merged signal. There is no need for prior training in the proposed approach. Since microphones are close-talking, the channel corresponding to the target speaker's microphone will have higher energy compared to the same speech from other channel. Thus energy of the speech signal

in different channels could be a potential indicator for selecting the target channel during speech. In addition to this energy based channel selection, we found that the acoustic characteristics of the merged signal are closer to that of the primary channel signal compared to that of the cross-talk channel signal. We also used this property to perform channel selection. We found that a long-term analysis window around the target frames improves the channel selection and overall VAD performance. Experiments with recordings of real conversations from cross-lingual doctor patient interactions show that the VAD performance using the proposed method is further improved compared to the cross-correlation-based VAD proposed in [50].

We begin with the description of the dataset (section 2) used in our experiment. The proposed approach for joint VAD and channel selection is described in section 3. Experiments and results of VAD on stereo recording is explained in section 4. Conclusions are drawn in section 5.

2.3.2 Data: Cross Lingual Medical Interactions

As a part of the medical domain speech-to-speech (s2s) translation project, we have recorded conversations between an English doctor and a Spanish patient with an interpreter in between. The recording was done in a typical office room setting with some background noise from the air-conditioner. The doctors are students from USC's Keck School of Medicine while the Spanish speaking patients are standardized patients (actors trained to assess Doctor skills). Three close-talking microphones were used for the doctor, patient, and interpreter. The Microtrack II professional 2-channel mobile digital recorder was used for stereo recording with sampling frequency 48kHz. The microphones of the doctor and the patient were connected to the left channel of the recorder and the interpreter to the right. Since the doctors and patients do not know each others' language, they communicated through the interpreter and have practically

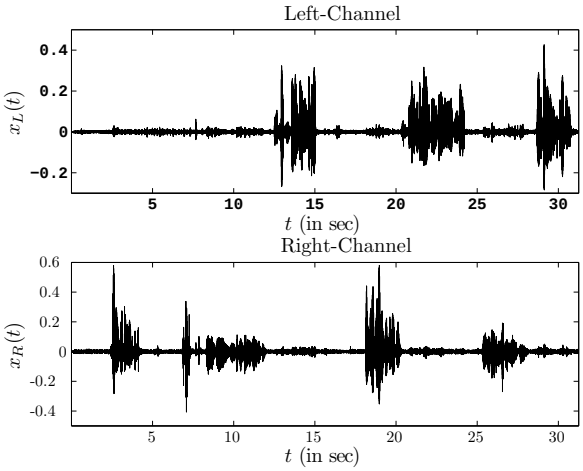

Figure 2.14: Sample signal shows a significant crosstalk. Right channel contains crosstalk over the following durations [12.5, 15.5], [20.5, 25], and [27, 31] sec. Similarly, crosstalks in left channel happen during [2.5, 12.5], [17.5, 20], and [25, 27] sec.

no overlap in their speech activity; therefore, the doctor and the patient are recorded on the same channel. From our entire collection, we have used 5 sessions of recordings for our experiments with an average duration of 30 minutes per session. During the collection, the patient and interpreter were sitting side by side while the doctor was sitting across approximately 1 meter away. This proximity and alignment of the participants resulted in a significant crosstalk. A sample recording for ∼30 seconds is illustrated in Fig. 2.14. The crosstalk in both channels because of the speech from the participants in other channels is clearly visible.

2.3.3 The proposed approach of VAD in stereo recording

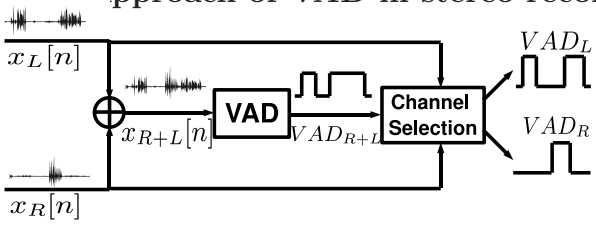

Figure 2.15: Schematic diagram of the VAD for stereo recording.

The proposed joint VAD and channel selection scheme is shown using a block diagram in Fig. 2.15. Let $x_R[n]$ and $x_L[n]$ denote the signal recorded in the right and the left channels respectively. Let us define $x_{R+L}[n] = x_R[n] + x_L[n]$. $x_{R+L}[n]$ is a single channel signal; hence, the standard VAD schemes can be used to detect the presence of speech in the merged signal at each frame (frame index m) with frame duration of N_w samples and frame shift of N_{sh} samples. The presence of speech in $x_{R+L}[n]$ at any frame indicates that there is speech either in the right or the left channel in the respective frame. In other words, when $\text{VAD}_{R+L}(m)$=1, we need to determine whether $\text{VAD}_R(m)$=1 or $\text{VAD}_L(m)$=1. This is done using the channel selection scheme which takes $\text{VAD}_{R+L}(m)$ and signals from two channels over a range of frame indices $[m-R,\ m+R]$ as inputs and produces the VAD for each channel (VAD_R and VAD_L). The purpose of performing VAD before channel selection is to determine the time segments when there is speech from the participants in the conversation. Unlike [50], we make a realistic assumption that the speech need not be present at every time instant and hence the channel selection should be performed only when there is speech in one of the channels. Below, we describe the channel selection scheme.

Suppose there are N samples observed from both right and left channels, i.e., $x_R[n]$ and $x_L[n], 0 \leq n \leq N-1$. Using the standard VAD on the combined signal, we know that there is speech from a speaker of the left or the right channel during the observation. Without loss of generality, let us assume that the speaker with the close-talking microphone corresponding to the right channel is speaking during this observation which implies that there is a leakage from the right channel to the left channel. Thus, we can write the observed signals in the two channels as follows:

$$\left.\begin{aligned} x_R[n] &= S[n] + N_1[n] \\ x_L[n] &= \alpha S[n-n_0] + N_2[n] \end{aligned}\right\} 0 \leq n \leq N-1 \tag{2.12}$$

where $N_1[n]$ and $N_2[n]$ are additive white noises and are assumed to be not only independent of each other but also independent of $S[n]$. We assume that $N_1[n]$ and $N_2[n]$ have zero mean and variance σ^2, $\forall n$. $S[n]$ is the speech of the participant in the right channel. $\alpha S[n-n_0]$ is the leakage in the left channel, where α denotes the level of leakage of the speech from the right to the left channel and n_0 denotes the delay in number of samples between the $S[n]$ and its leaked version. We also assume that $0< \alpha < 1$ based on the longer distance of the active speaker from the left microphone and the directionality of the close-talk microphones.

2.3.3.1 Energy-based channel selection

The power of the signal in the right channel ($\sigma^2_{x_R}$) is more compared to that of the left channel ($\sigma^2_{x_L}$). Let σ^2_s be the variance of the speech signal (assuming $S[n]$ is zero mean). Then,

$$\begin{aligned} \sigma^2_{x_R} &= \sigma^2_s + \sigma^2 \\ &> \alpha^2\sigma^2_s + \sigma^2 = \sigma^2_{x_L} (\because 0 < \alpha < 1) \\ \implies \sigma^2_{x_R} &> \sigma^2_{x_L} \end{aligned} \tag{2.13}$$

Thus, if the power of the right channel is more than that in the left channel (i.e., $\sigma^2_{x_R} > \sigma^2_{x_L}$), then the target channel at the respective frame is the right channel, i.e., $\text{VAD}_R(m)$=1. We refer this energy based channel selection scheme by ENERGY.

Let $\mathbf{F}^R_l(\omega) = |X^R_l(\omega)|^2$, $\mathbf{F}^L_l(\omega) = |X^L_l(\omega)|^2$, and $\mathbf{F}^{R+L}_l(\omega) = |X^{R+L}_l(\omega)|^2$ denote the Fourier spectra of $x_R[n]$, $x_L[n]$, and $x_{R+L}[n]$ at the l^{th} frame. Note that $X^{R+L}_l(\omega) =$

$X_l^R(\omega) + X_l^L(\omega)$, where ω denotes the angular frequency. Considering non-overlapping frames, we can write eqn. (2.13) using Parseval's theorem as follows:

$$\sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \mathbf{F}_l^R(\omega) d\omega > \sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \mathbf{F}_l^L(\omega) d\omega \tag{2.14}$$

$$\implies \sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \left| X_l^{R+L}(\omega) - X_l^L(\omega) \right|^2 d\omega$$

$$> \sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \left| X_l^{R+L}(\omega) - X_l^R(\omega) \right|^2 d\omega$$

This means that the short-time spectra of the combined signal will be closer to that of the primary channel as opposed to the crosstalk channel. This motivates us to explore the perceptual distance between short-time spectra and other spectral feature based on acoustic proximity measures for the task of channel selection.

2.3.3.2 Acoustic proximity measures for the channel selection

We have used two measures of acoustic and perceptual proximity (discussed below) between two signal segments to determine whether $x_{R+L}[n]$ is closer to $x_R[n]$ or $x_L[n]$.

2.3.3.2.1 Euclidean distance between MFCC features

The mel-frequency cepstral coefficients (MFCC) [107] capture the energy distribution over different frequency bands for a short-time speech segment. Let $\mathbf{c}_l^R$, $\mathbf{c}_l^L$, and $\mathbf{c}_l^{R+L}$ denote the MFCC vectors of $x_R[n]$, $x_L[n]$, and $x_{R+L}[n]$ respectively at the l^{th} frame. The distance between MFCC of $x_{R+L}[n]$ and $x_R[n]$ over the $2R+1$ observed frames is computed as

$$\delta_{R+L,R} = \sum_{l=m-R}^{m+R} ||\mathbf{c}_l^{R+L} - \mathbf{c}_l^R||_2^2 \tag{2.15}$$

where $||\cdot||_2$ denotes the L_2 norm of a vector. Similarly, $\delta_{R+L,L}$ is computed. If $\delta_{R+L,R} < \delta_{R+L,L}$, it indicates that the spectral characteristics of $x_{R+L}[n]$ are closer to that of $x_R[n]$ than that of $x_L[n]$ over the observed signal segment. Thus, when $\delta_{R+L,R} < \delta_{R+L,L}$, $\text{VAD}_R(m)$ is set to 1, otherwise $\text{VAD}_R(m)$ is set to zero. Decision for VAD_L is done in a similar way. We refer to this channel selection scheme by MFCC_0. We also consider a channel selection scheme without the zero-th coefficient of MFCC, which we refer only by MFCC.

2.3.3.2.2 Itakura-Saito distance

Itakura-Saito (IS) distance [77] is a measure of the perceptual difference between the spectra of two short-time signal segments. The distance between $x_{R+L}[n]$ and $x_R[n]$ over the $2R+1$ observed frames using frame-based IS distance is computed as

$$\delta_{R+L,R} = \sum_{l=m-R}^{m+R} d\left(\mathbf{F}_l^R(\omega), \mathbf{F}_l^{R+L}(\omega)\right) \tag{2.16}$$

$$\text{where, } d\left(\mathbf{F}_l^R(\omega), \mathbf{F}_l^{R+L}(\omega)\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{\mathbf{F}_l^R(\omega)}{\mathbf{F}_l^{R+L}(\omega)} - \log\frac{\mathbf{F}_l^R(\omega)}{\mathbf{F}_l^{R+L}(\omega)} - 1\right]d\omega$$

$\delta_{R+L,L}$ is computed in a similar way. Similar to the description in section 2.3.3.2.1, $\text{VAD}_R(m)$ is set to 1 when $\delta_{R+L,R} < \delta_{R+L,L}$, otherwise $\text{VAD}_R(m)$ is set to zero. VAD_L is determined similarly. We refer to the IS based channel selection scheme by IS.

Note that, under the signal model as defined in eqn. (2.12), the energy-based channel selection scheme (eqn. (2.13)) is independent of the noise variance σ^2. However, this may not be true for MFCC, MFCC_0 and IS channel selection schemes. Also, in practice, noise variance in both channels need not be identical; in such cases, MFCC, MFCC_0, and IS could be more robust compared to energy-based channel selecction scheme. Due to the non-linear function involved in the MFCC feature computation and

IS distance computation, they are not analytically tractable unlike the energy-based approach. Also, in practice, the two channel signal model defined in eqn. (2.12) is a simplification (e.g. it doesn't consider reverberation) and we would like to explore the utility of MFCC, MFCC_0 and IS for such cases through experiments.

2.3.4 Experimental Results

We manually labeled the speech segments in 5 sessions of stereo recording (overall $\sim$2.5 hours of two channel audio) to obtain reference VAD decisions which are used to evaluate the proposed approach. As a baseline, we chose the VAD decisions in each channel obtained by the ETSI AMR VADs option 2 [2] (AMRVAD2). The implementation was taken from [1]. We have downsampled the speech signal in both channels to 8kHz for VAD so that the sampling frequency satisfies the AMRVAD2 requirements. Based on the analysis in section 3.3.1.2, we used the VAD on the combined signal, $x_{R+L}[n]$, followed by ENERGY, MFCC, IS, and MFCC_0 schemes to obtain channel specific VAD decisions. For comparison, we also used the cross-correlation based channel selection scheme, similar to [50], which is denoted by CORR. This is done by finding the peak in the cross-correlation between signals of two channels and determining the delay between them.

VAD decisions are made every 10 msec (N_{sh}=80 samples) and the short-time frame duration is chosen to be 20 msec (N_w=160 samples). As described in section 3.3.1.2, we consider the signal segment over the frame indices $[m - R,\ m + R]$ to determine $\text{VAD}_R(m)$ and $\text{VAD}_L(m)$. We experiment by varying the value of R from the following set $\{1,\ 3,\ 5,\ 9,\ 15,\ 25,\ 50\}$; R=1 corresponds to the case when no signal from neighboring frame is considered and R=50 corresponds to the case when a signal segment of approximately 1 sec is considered with the current frame in the middle. Also, note that channel selection scheme is applied only to those frames for which VAD_{R+L}=1.

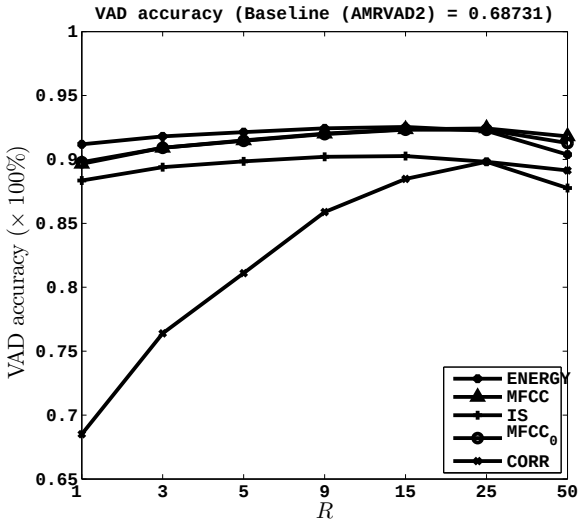

Figure 2.16: Stereo VAD accuracy obtained by different methods for varying R compared against the baseline accuracy of 68.73%.

We plot the VAD accuracy obtained by different channel selection schemes for varying R in Fig. 2.16. VAD accuracy is determined by the number of correctly detected speech and non-speech frames against the reference VAD decisions. It is clear that the VAD accuracies obtained using ENERGY, MFCC, MFCC_0 are similar and higher than those obtained by IS and CORR. Fig. 2.16 also indicates that the correlation based channel selection scheme is not realiable particularly over short frame lengths. This could be due to the fact that over short frame length a significant peak may not be observed in the cross-correlation estimate due to background noise; over long frame length, the estimate of the cross-correlation would be better and hence the effect of noise will be reduced. We observe that the VAD accuracy increases as R increases consistently for all the different channel selection schemes; when R goes above 15 or 25, the VAD accuracy starts to decrease. The highest VAD accuracy of 92.54% is obtained by ENERGY for R=15, which is $\sim$24% absolute improvement over baseline. The best VAD accuracies obtained by MFCC and MFCC_0 are 92.43% and 92.33% respectively for R=25 and R=15. This means that although ENERGY and MFCC do not exploit

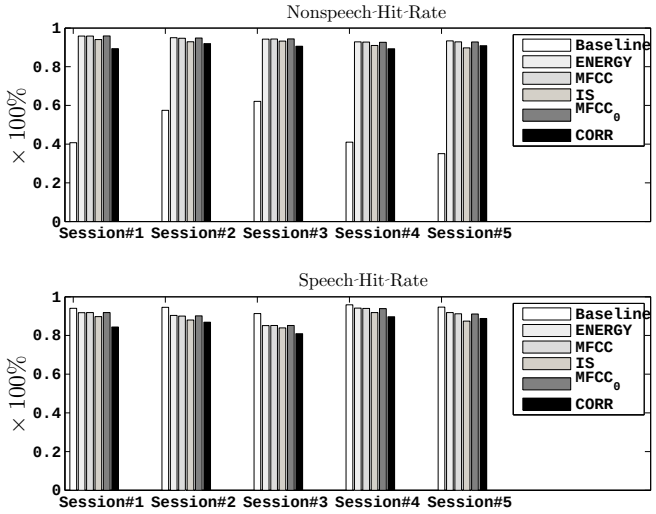

Figure 2.17: Speech and non-speech hit rate obtained by different channel selection schemes for five sessions considered at R=15.

identical acoustic properties of the signal, they yield similar VAD performance. Note that the VAD accuracy in individual channels obtained as a result of the channel selection scheme is affected by the VAD on the combined signal. The VAD accuracy on the combined signal using AMRVAD2 is $\sim$90%. Since channel selection is performed only in the frames where VAD_{R+L}=1, incorrect detection of a speech frame will affect the performance of VAD in both channels; but mis-detection of a non-speech frame will affect the performance of only one channel. A better VAD algorithm on the combined signal can result better VAD performance for the individual channels after channel selection.

In Fig. 2.17, we plot the speech hit rates (i.e., number of correctly detected speech frames) and non-speech hit rates for each session obtained by different channel selection schemes at R=15. It is clear that the performance of different schemes is consistent across sessions. It is also clear that the baseline non-speech hit rate is significantly poor compared to the proposed schemes including CORR for all sessions. Thus, the overall VAD accuracy improves because of the improvement in non-speech hit rate. In other words, the effect of crosstalk on VAD is significantly reduced by the proposed scheme.

For a comprehensive evaluation of the proposed stereo VAD schemes, we follow the testing strategy proposed by Freeman et al [51], where five different parameters reflecting the VAD performance are considered: (1) CORRECT, (2) FEC (front end clipping), (3) MSC (mid speech clipping), (4) OVER (carry over), and (5) NDS (noise detected as speech). Fig. 2.18 shows these five parameters obtained by various VAD schemes at R=15 over all five recording sessions. We observe that five evaluation parameters obtained by ENERGY, MFCC, MFCC_0 are similar to each other and are consistently better than the baseline and of those obtained by IS and CORR indicating the effectiveness of the proposed VAD scheme in the presence of crosstalk.

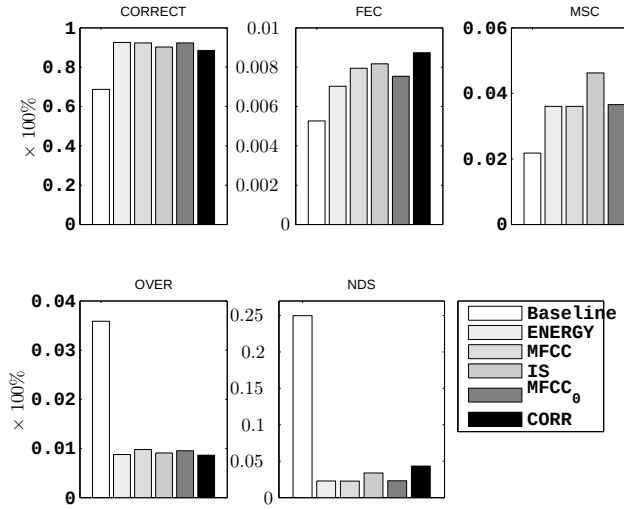

Figure 2.18: Different VAD performance evaluation measures obtained by the proposed schemes at R=15.

2.3.5 Conclusions

From the results of the VAD experiments, we find that consideration of a long-term analysis window yields better results compared to short-time frame. This is consistent with the findings in [47]. A long-term approach based VAD could be used on the combined signal as well to improve the stereo VAD accuracy further. It should be noted

that the proposed joint VAD and channel selection scheme can be easily extended to apply for audio recordings with more than two channels. We obtain ∼8% VAD error over 2.5 hours of stereo audio considered in our experiment. One potential reason for the remaining errors could be the overlapping speech in the two channels, which is not considered in the proposed two channel signal model. Although the overlapping speech is a small fraction in the stereo recording considered in this paper, in practice it could be significant portion of any natural interaction. Thus, considering overlapped speech within the framework may improve the VAD accuracy and is part of our future work.

Chapter 3:

High-quality bilingual subtitle document alignments with application to spontaneous speech translation [13, 11]

Works presented in this chapter have been carried out in collaboration with Prasanta Gosh

In this work, we investigate the task of translating spontaneous speech transcriptions by employing aligned movie subtitles in training a statistical machine translator (SMT). In contrast to the lexical-based dynamic time warping (DTW) approaches to bilingual subtitle alignment, we align subtitle documents using time-stamps. We show that subtitle time-stamps in two languages are often approximately linearly related, which can be exploited for extracting high-quality bilingual subtitle pairs. On a small tagged data-set, we achieve a performance improvement of 0.21 F-Score points compared

to traditional DTW alignment approach and 0.39 F-Score points compared to a simple line-fitting approach. In addition, we achieve a performance gain of 4.88 BLEU score points in spontaneous speech translation experiments using the aligned subtitle data obtained by the proposed alignment approach compared to that obtained by the DTW based alignment approach demonstrating the merit of the time-stamps based subtitle alignment scheme.

3.1 Introduction

Speech-to-speech (S2S) systems are used to translate conversational speech among different languages. In S2S systems, a critical component is the statistical machine translator (SMT). Due to the broad range of topics, domains, and different speaking styles that need to be potentially handled, an enormous amount of bilingual corpora that adequately represent this variety is ideally required to train the SMT. Therefore the S2S research and development efforts have not only focused on manually collecting multilingual data but also on automatically acquiring data, for example, by mining bilingual corpora from the Internet matching the domain of interest.

It is advantageous for the SMT of an S2S system to be trained on bilingual transcriptions of spontaneous speech corpora because they match the spontaneous speech style of ultimate S2S usage. A source of bilingual corpora that has gained attention recently is movie subtitles. Aligned subtitle documents in two languages can be used in SMT training. In this work, our efforts focus on extracting high quality bilingual subtitles from movie subtitle documents.

Corpora alignment research for training machine translators has been active since the early 90's. Past works have introduced a variety of methods for sentence alignment including the use of the number of tokens of each utterance [18], the length of

sentences [37], and the frequency, position and recency information under the dynamic time warping (DTW) framework [34].

Movie subtitle alignment as a source of training data in S2S systems is attractive due to the increasing number of available subtitle documents on the web and the conversational nature of speech reflected in the subtitle transcripts. Recently, there have been many attempts to align bilingual movie subtitle documents. For example, [59] were one of the first to describe a methodology to align movie subtitle documents. [55] posed this problem as a sequence alignment problem such that the total sum of the aligned utterance-similarities is maximized. [11] proposed a distance metric under a DTW minimization framework for aligning subtitle documents using a bilingual dictionary and showed improvement in subtitle alignment performance in terms of F-score [60]. Even though the DTW algorithm has been used extensively, there are inherent limitations due to the DTW assumptions. Notably, the DTW-based approaches have the disadvantage of not providing an alignment quality measure, resulting in the use of poor translation pairs depending on the performance of the alignment approach. Using such poor translation pairs results not only in degrading the performance but also in increasing the training and decoding time, an important factor in SMT design.

As a rule of thumb, increasing the amount of correct bilingual training data improves the SMT performance. Objective metrics for evaluating the performance of SMTs include the BLEU score [70]. [87] reported BLEU score improvements using subtitle data with only 49% accurate translations, demonstrating the usefulness of subtitle data. It should be noted that Sarikaya *et al.* included an additional step to their scheme by automatically matching the movies first, resulting in a potentially noisy step that can cause performance degradation. This step can be avoided since many subtitle websites offer deterministic categorization of subtitle documents with respect to the movie title.

Importantly, their approach has not used any information from the sequential nature of bilingual subtitle documents alignment as done in DTW approaches.

Timing information has been considered in subtitle documents alignment. Tiedemann [98, 99, 97] synchronized subtitle documents by using manual anchor points and anchor points obtained from cognate filters. In addition, an existing parallel corpus was used to learn word translations and estimate anchor points. Then, based on the estimated anchor points, subtitle documents were synchronized to obtain bilingual subtitle pairs. However, in many cases a parallel corpus is either not available or there is a domain mismatch, so in such cases anchor point estimation using parallel corpus is not a feasible option. [43] introduced a cost function to align subtitle documents using subtitle durations and sentences lengths under the DTW framework to find the best alignments. However, this approach fails when the subtitle documents contain many-to-one and one-to-many subtitle pairs because they tend to skew the sentence length and subtitle timing duration. Even when there are only one-to-one subtitle pairs, it requires that the subtitles have approximately the same length which might not be true for all language pairs. Also, time shifts and offsets [43] can distort the subtitle durations. [106] proposed an approach that uses time differences, and the approach was applied only for subtitle documents having the same starting and ending time-stamps. They reported comparable performance to subtitle alignment works using lexical information. In addition, they reported performance gains by incorporating lexical information.

Time-stamps can be crucial and important in aligning subtitle document pairs. In this work, we aim to study the properties and the benefits of the timing information and matching bilingual subtitle pairs using time-stamps. We propose a two-pass method to align subtitle documents. The first pass uses the Relative Frequency Distance Metric (RFDM) [11] under the DTW framework. Using the DTW approach and the lexical information, we identify bilingual subtitle pairs. It is crucial at this point to find

pairs that are actual translations of each other and that have timing information describing the deterministic relation between the time-stamps. The identification and usage of these pairs is incorporated in the proposed approach. The second pass uses timing information to align subtitle documents. In particular, we assume that there exists an approximately linear mapping between the time-stamps of the bilingual subtitle documents that can align the bilingual subtitle pairs. This assumption is verified experimentally for most of the bilingual subtitle documents available in our bilingual subtitle sets. This approach results in high quality translation pairs and, in a small set with tagged mappings, significant improvement in the alignment accuracy is obtained compared to that in our prior work [11]. Also, the performance of this method is demonstrated by training and testing an SMT using downloaded subtitle documents from the web (http://www.opensubtitles.org) on a large scale.

This work is structured as follows. In section 2, we present the theory and implementation used in this work. In section 3, we describe the experimental results and the evaluation methodology used in our approach. Finally, in section 4, we summarize the results of this work.

3.2 Theory and methodology

We start by formulating the subtitle alignment problem under the DTW framework. Next, we formulate the time-stamp-based subtitle alignment method. Finally, we describe the methodology used to align the subtitles under the proposed two-pass approach. The general diagram of the two-pass approach is shown in Fig. 3.1

3.2.1 First step: DTW using lexical information

We follow the definition and approach as followed by [11]. We define the utterance fragments with starting and ending time-stamps as subtitles and the sequence of subtitles of

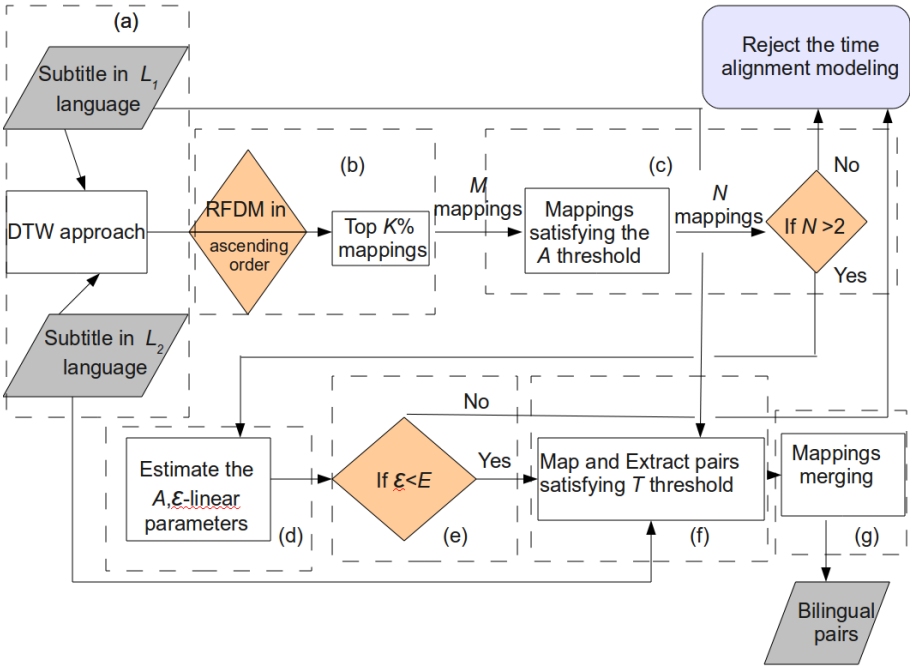

Figure 3.1: Two-step bilingual subtitles document alignment approach.

a movie as a subtitle document. The first part of the movie subtitle alignment problem is defined as follows:

Say the subtitles documents in two languages, L_1 and L_2 are to be aligned. We denote the i^{th} subtitle in the L_1 subtitle document as $S_i^{L_1}$ and the j^{th} subtitle in the L_2 subtitle document as $S_j^{L_2}$. Also, let N_1 and N_2 be the number of subtitles in the L_1 and L_2 subtitle documents respectively. We try to estimate the mappings $m_{ij}^{\star}$ that minimize the global distance as follows [11]:

$$\{m_{ij}^{\star}\} = \arg\min_{m_{ij}} \sum_{i,j} m_{ij} \mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right) \tag{3.1}$$

where $m_{ij} = 1$, if $S_i^{L_1}$ aligns with $S_i^{L_2}$ and $m_{ij} = 0$ otherwise
and $\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$ is a distance measure between $S_i^{L_1}$ and $S_j^{L_2}$.

The above-mentioned optimization problem can be solved efficiently using the DTW algorithm under the following assumptions:

(i) Every subtitle in the L_1 document must have at least one mapping with a subtitle in the L_2 document and vice versa.

(ii) The estimated mappings must not cross each other. Thus, if $m_{ij} = 1$ is a correct match, then $m_{i+k,j-l} = 0,\ k = 1, 2, ..., N_1 - i$ and $l = 1, 2, ..., j - 1$ must be satisfied.

(iii) Finally, we assume $m_{1,1} = 1$ and m_{N_1,N_2}=1, which implies that the first and last subtitles match (i.e., $S_1^{L_1}$ matches with $S_1^{L_2}$ and $S_{N_1}^{L_1}$ matches with $S_{N_2}^{L_2}$).

The DTW block is shown in Fig. 3.1 in dashed rectangle (a). The details of the DTW algorithm used in this step is described in .3. The inputs are two bilingual subtitle documents and the output is a list of aligned subtitles with their time-stamps. In the next section, we discuss the distance metric used by the DTW.

3.2.1.1 Distance Metric

Following [11], we define the Relative Frequency Distance Metric (RFDM) between subtitles across the two languages as follows:

Consider the subtitle $S_i^{L_1}$ and denote the words in that subtitle by W_i. Also, the words of the subtitle $S_j^{L_2}$ are translated using a dictionary and the resulting bag of words of the translated subtitle is denoted by $\mathcal{B}_j$. Note that both $\mathcal{B}_j$ and W_i contain words in the language L_1. First, we compute the unigrams distribution of the words in

the L_1 subtitle document. Using the unigrams distribution of words in the L_1 subtitle document, the RFDM is defined as:

$$DM\left(S_i^{L_1}, S_j^{L_2}\right) = \left(\sum_{k \in W_i \cap B_j} p_k^{-1}\right)^{-1} \tag{3.2}$$

where p_k is the relative frequency of the word k in the L_1 subtitle document. RFDM has the property that it gives high-quality anchor points of subtitle pairs. The lower the RFDM score, the higher the similarity of the subtitles is. In particular, low RFDM occurs when there are infrequent words that match in both subtitles. For example, the sum of the inverse probability of infrequent words will be high and, thus, the inverse of the sum will be low. Hence, infrequent words in the text play the important role of aligning subtitle documents. Finally, RFDM is used as a distance metric to obtain the best mappings $\left\{m_{ij}^{\star}\right\}$.

3.2.1.2 Experimental procedure

To experiment with parallel subtitles alignment using this first step, we downloaded 42 subtitle Greek-English pairs (http://www.opensubtitles.org/) and we randomly aligned 40 english utterances with the corresponding utterances, totalling 1680 pairs. As a preprocessing step, we cleaned subtitles from noisy symbols (i.e. non-alphanumeric symbols) similar to what one would do for cleaning text for statistical machine translation purposes; we removed all the time stamps and the reference numbers. Finally, we removed all punctuation and capitalized all letters.

We mined English translations of Greek words by querying the Google dictionary ("http://www.google.com/dictionary"). Using translated words, we computed the RFDM. For comparison, we also computed the F-measure [55]. In addition, we inverted

the F-measure (called IF-measure) and used it as $\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$ in .3 to obtain the best alignments.

We used the manual mappings as references for evaluating the movie subtitle alignment task. We computed the precision, the recall, and reported the F-score [60], averaged over all movies. To check the effect of stemming on this problem, we repeated the above experiments with stemming [62]. Stemming was performed on the English words and on the English BOW obtained after translating Greek utterances.

3.2.1.3 Results

In table 1, we report the F-score of the movie subtitle alignment task using the F-measure, the IF-measure and the RFDM. We compare the metric proposed by Lavecchia et al. in [55] where they considered the movie subtitle alignment task as a maximization problem using the F-measure, unlike minimization of RFDM. From table 1, it can be seen that formulating the alignment problem as the minimization of the total distance using the RFDM measure is advantageous over formulating it as a maximization of the total similarity using F-SCore. The RFDM gives 8-11% absolute improvement in terms of the F-score compared to that of the F-measure.

Metric	No Stemming	Stemming
F-measure	0.609	0.607
‘ IF-measure	0.681	0.688
RFDM	0.711	0.713

Table 3.1: The F-score of subtitle alignment using different metrics

The F-scores obtained for metrics computed by stemming do not differ much from those without stemming. One possible reason could be that stemming introduced some noise, by mapping words of different meaning to the same root word. Due to such noise, any benefit from stemming might have been compensated.

From this point on, when we refer to DTW experiments, we mean that we minimize the DTW distance using the RFDM distance metric.

3.2.2 Second step: alignment using timing information

We select a subset of the best DTW output mappings $\left\{m_{ij}^{\star}\right\}$ and estimate a relation among the bilingual subtitles.

In this work, we argue that one can relate the time-stamps of most bilingual subtitles using a linear relation. We hypothesize that this linearity assumption stems from the fact that movies are played in different regions and versions with varying frame rates (slope) and varying offset times (intercept).

For this purpose, consider the scenario of aligning subtitle documents in two languages, say L_1 and L_2. Assume L_1 is the source language and L_2 is the target language. Also, assume that we know a-priori M actual one-to-one matching pairs, for example, subtitles which are bilingual translations of each other. Moreover, consider the i^{th} one-to-one pair. We denote the starting and ending time-stamps of the i^{th} subtitle in L_1 by x_{1i} and x_{2i} respectively. The starting and ending time-stamps of the matching subtitle in the L_2 subtitle document are denoted by y_{1i} and y_{2i}. Hence, using the time-stamps of M pairs, we define the set $P = \{\{x_{1i}, y_{1i}\}, \{x_{2i}, y_{2i}\} : 1 \leq i \leq M\}$. In addition, we use the following definition:

Definition 1 *The absolute error, $\mathcal{E}$, of a set of N pairs given a linear function $f(x) = mx + b$ is defined by:*

$$\mathcal{E} = \frac{1}{2N} \sum_{i \in I} |mx_{1i} - y_{1i} + mx_{2i} - y_{2i} + 2b|$$

As discussed in the previous paragraph, the end goal is to approximate the relation of the starting and ending time-stamps of bilingual subtitles with an approximately

linear function. Under the assumption of linear mapping, the time-stamps are related by $f(\frac{x_{1i}+x_{2i}}{2}) = \frac{y_{1i}}{2} + \frac{y_{2i}}{2}$, where f is a linear function. Since in practice the relation is not exactly linear, due to factors like human error in tagging, we allow an absolute error bound for all the bilingual pairs. Thus, we model the relation between time-stamps of subtitles in L_1 and L_2 with an α, ϵ-linear function of order N which is defined next.

Definition 2 *A function $f(x) = mx+b$ is called an α, ϵ-linear function of order N if for a set of pairs $P = \{\{x_{1i}, y_{1i}\}, \{x_{2i}, y_{2i}\} : 1 \leq i \leq M\}$ there is a set $I \subseteq \{i : 1 \leq i \leq M\}$ of order $|I| = N$ pairs with $3 \leq N \leq M$ such that:*

(i) $\frac{1}{\alpha} < \left|\frac{y_{2i}-y_{1i}}{x_{2i}-x_{1i}}\right| < \alpha, \quad \forall i \in I$ *and* $\alpha > 1$

(ii) $\mathcal{E} \leq \epsilon$, *where* $\mathcal{E}$ *is the absolute error of* I *given the linear function* $f(x)$

Def. 2 uses a linear function f to relate a subset of the set of pairs, P, (the starting and ending time-stamps in the source language and the corresponding time-stamps in the target language) under two conditions. Initially, we have M pairs (in practice returned by the DTW step). Then, a subset of N out of M pairs and a linear function f based on the α and ϵ parameters are defined. The α parameter controls the allowed duration divergence of bilingual subtitles at subtitle level. The ϵ parameter establishes the connection between the linear function f and the N pairs by imposing a maximum absolute error between the linear function and the points.

In the ideal case, time-stamps are ideally scaled and shifted from source to target time-stamps, no noise is introduced and there are N one-to-one pairs. Any two pairs selected will fall on a line with the same slope, α, and $\epsilon \to 0$. Thus, if we could extract the N noise-free one-to-one pairs, then, the relation would be simply a straight line connecting the middle points of the pairs. In other words, the lower the absolute error, the closer is the relation of the pairs to a line, thus, the more approximately linear their relation is. Hence, ideally, we want ϵ as small as possible.

On the other hand, in the practical case, humans will transcribe the movies separately. On top of the ideal time scaling and shifting, noise will be introduced to the time-stamp points. Hence, the absolute error is used to reflect the linearity of the pairs selected. Using the absolute error as a measure to reflect the linearity of the map offers a great advantage. The absolute error, $\mathcal{E}$, is just an average of N points, thus, $\mathcal{E}$ is robust to M and N variations, making the absolute error comparable across aligning different bilingual subtitle documents. In addition, in practice, it is crucial to select N reliable points to estimate the linear function, rather than considering all M points. At the global level, a movie's duration could be scaled by a few minutes or seconds. However, at the local level (subtitle level), this duration change is in the order of milliseconds and we expect the bilingual subtitles to have similar durations. For this purpose, α is used to filter bilingual subtitles with large duration divergence.

In summary, modelling the subtitles alignment problem using α, ϵ-linear functions offers various advantages compared to the DTW-based modeling approach [11]. First, α serves as a quality measure to accept or reject the pairs used to estimate the relation. Then, the absolute error, $\mathcal{E}$, is employed to filter the sets of N pairs that cannot describe a linear relation. Consequently, α and ϵ serve as measures for the quality of the alignments. In addition, alignment using α, ϵ-linear functions depends only on timing information rather than on the semantic closeness of the utterances which is more complicated to model.

Based on Def. 2, once the α is set, one can find no or infinitely many m's and b's that satisfy the three conditions. However, we seek m^* and b^* that minimize the squared-error of the pairs considered, so that the total squared error is minimum for the N pairs. Such a function is defined next.

Definition 3 *A function* $f^*(x) = m^*x + b^*$ *is called an optimal* α, ϵ*-linear function of order* N *if for a set of pairs* $P = \{\{x_{1i}, y_{1i}\}, \{x_{2i}, y_{2i}\} : 1 \leq i \leq M\}$ *and* $I \subseteq \{i : 1 \leq i \leq M\}$ *of size* $|I| = N$ *the following are satisfied:*

(i) The function f^* *is an* α, ϵ*-linear function of order* N*.*

(ii) f^* *minimizes* $MSE = \sum_{i \in I} \left(\frac{y_{1i}}{2} + \frac{y_{2i}}{2} - f^* \left(\frac{x_{1i}}{2} + \frac{x_{2i}}{2} \right) \right)^2$.

The optimal function parameters, m^* and b^* are estimated using the least squares line-fitting method. The difference between the least squares line-fitting and this method is that we are using a subset of high-quality mappings to estimate the line in order to control the quality of the linear relation. Thus, the relation is robust to errors either from bad estimates of the DTW step or from additional noise. For the sake of completeness, we show the formula for estimating the optimal estimates, m^* and b^* along with the proof in .4 .

3.2.2.1 Implementation

An overall diagram of the proposed implementation described in this section is shown in Fig. 3.1.

3.2.2.1.1 Select one-to-one mappings As discussed in the previous section, the end goal of this approach is to estimate a relation between the subtitles in the L_1 and L_2 documents based only on the time-stamps under the assumption that they are approximately related by a linear function. Initially, we need to extract a set of reliable points that best describe the relation between the subtitles in L_1 and L_2 subtitle documents. For this purpose, we assume that the most reliable mappings are the $K\%$ one-to-one pairs with the lowest RFDM returned by the DTW approach. By one-to-one pairs, we mean the source subtitles each of which is related with exactly one subtitle in the target subtitle document.

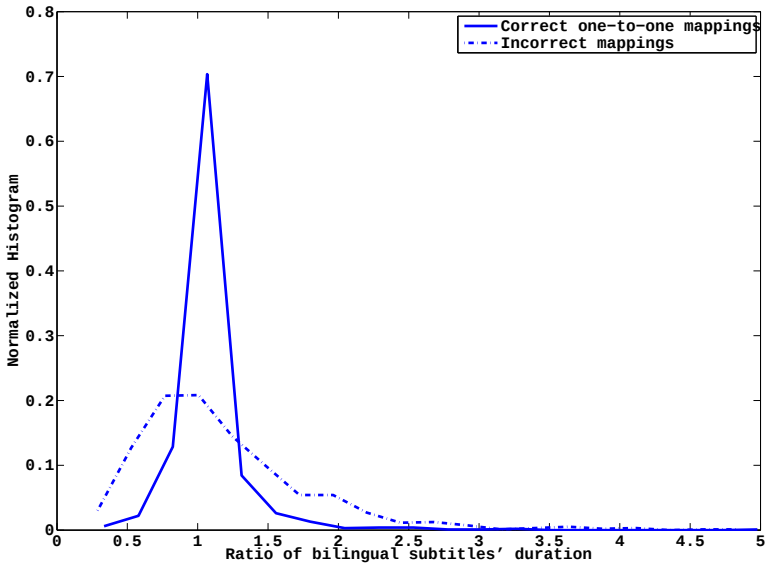

Figure 3.2: This figure shows the distribution of the ratio of the pair durations for correct and incorrect subtitle mappings.

This step is shown in the dashed rectangle (b) of Fig. 3.1. As shown in the diagram the input is the DTW-step output and the output is a list of ranked RFDM values.

3.2.2.1.2 Duration ratio bound After keeping only the one-to-one mappings, M mappings are left. At this point, our goal is to find one α, ϵ-linear function of order N which could model the subtitles alignment problem using time-stamps. In practice, we optimize α on a development set and denote this value as A. Thus, A acts as a bound to accept only the reliable mappings to be used in the A, ϵ-linear function parameters estimation. To justify the usage of this bound, we study and present its relation with correct and incorrect mappings.

Fig. 3.2 shows the empirical distribution of the duration ratio, $\left|\frac{y_{2i}-y_{1i}}{x_{2i}-x_{1i}}\right|$, for correct mappings along with the empirical distribution for incorrect mappings. The distribution of correct mappings shows that the ratio of pairs duration is mostly in the range $\frac{1}{2} <$

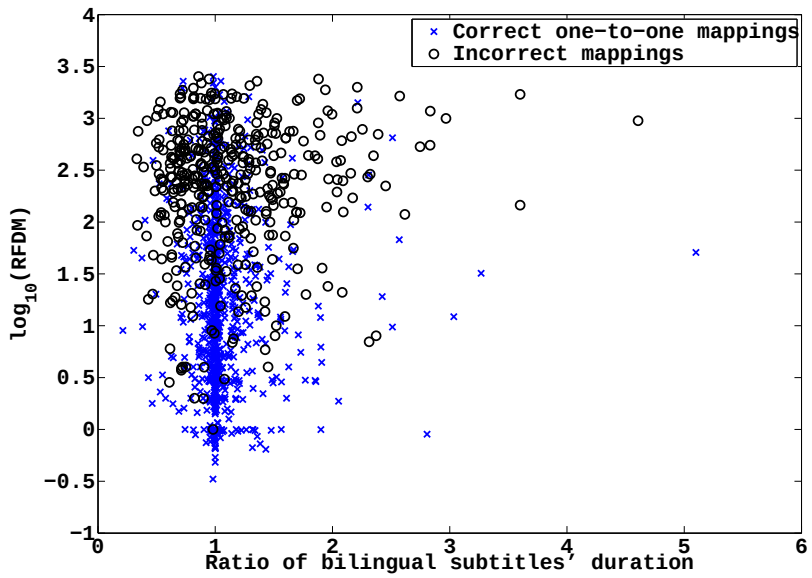

Figure 3.3: This is the scatter-gram of the correct and incorrect mappings with respect to the log(RFDM) value and the duration ratio.

$\left|\frac{y_{2i}-y_{1i}}{x_{2i}-x_{1i}}\right| < 2$. Thus, it is reasonable in practice to impose this constraint on the duration ratio of the mappings to filter out the incorrect mappings.

Fig. 3.3 is a 2-dimensional scatter-gram showing how the correct and incorrect mappings are distributed with respect to the log(RFDM) value and the duration ratio. As Fig. 3.3 suggests, mappings with low RFDM and duration ratio close to 1 justify the fact that they are important in selecting one-to-one mappings. Thus, DTW will return $K\%$ reliable mappings and A will play the role of detecting the outlier points by imposing the constraint of the property (i) of the α, ϵ-linear functions (Def. 2). Hence, the thresholds K and A are important in filtering incorrect mappings while estimating the A, ϵ-linear function parameters.

The duration ratio bound block is shown in the dashed rectangle (c) of Fig. 3.1. This block filters out mappings with duration ratio higher than A. The input to this

block are the ranked RFDM mappings and the output is the subset of there mappings with duration ratio $\frac{1}{A} < \left|\frac{y_{2i}-y_{1i}}{x_{2i}-x_{1i}}\right| < A$.

3.2.2.1.3 Line parameters estimation As a consequence of the previous step, for a fixed $A = \alpha$, the N pairs that satisfy $\frac{1}{A} < \left|\frac{y_{2i}-y_{1i}}{x_{2i}-x_{1i}}\right| < A$ are used to estimate the optimal slope, m^*, and intercept, b^*, of the A, ϵ-linear function (of order N will be omitted but implied from this point onwards) using the results of .4. Moreover, the absolute error is computed using the N pairs and the function $f^*(x) = m^*x + b^*$.

The line parameters estimation block takes as input the mappings with duration ratio less than A and outputs the optimal slope, m^*, intercept, b^*, the absolute error, $\mathcal{E}$, of the A, ϵ-linear function and the filtered mappings. The line-parameters estimation block is shown in the dashed rectangle (d) of Fig. 3.1.

3.2.2.1.4 Absolute error threshold Now, we need a measure to assess the level of linearity of the mapping. For this purpose, we define a fixed threshold, E. Due to the fact that $\mathcal{E}$ is robust to M and N variations (as discussed in 3.2.2), E is used as an upper bound to the check if the absolute error, $\mathcal{E}$, is "low" enough. Hence, by assumption, we accept the A, E-linear modeling, if $\mathcal{E} \leq E$. If this condition is not satisfied, the alignment cannot be modeled with an A, E-linear function of order N. In this case, one might choose another set of N pairs or use only the DTW approach if there is no approximately linear relation between the time-stamps.

The absolute error threshold block is shown in the dashed rectangle (e) of Fig. 3.1. The input of this block are the A, E-linear function parameters and filtered mappings and the output is a decision if the A, E-linear function can model the subtitles relation. Also, this block output the A, E-linear function parameters.

3.2.2.1.5 Time-stamps mapping With the A, E-linear function and the optimal slope, m^* and intercept b^* in place, we relate all starting time-stamps by translating the L_1 subtitle document time-stamps into the L_2 subtitle document time-stamps. In particular, assume x_1 is a starting time-stamp in the L_1 document. Then, the assigned starting time-stamp in the L_2 document is the point y_1 that minimizes the distance $D_1 = |y_1 - f(x_1)|$. Similarly, we relate all ending time-stamps in the L_1 document with ending time-stamps in the L_2 document. Assume, x_2 is an ending time-stamp in the L_1 document; then the assigned ending time-stamp in the L_2 document is the point y_2 that minimizes the distance $D_2 = |y_2 - f(x_2)|$. Also, we seek additional subtitle pairs by mapping y_1 with the starting time-stamp of x_1 that minimizes $D_3 = \left|x_1 - f^{-1}(y_1)\right|$ and by mapping y_2 with the ending time-stamp of x_2 that minimizes $D_4 = \left|x_2 - f^{-1}(y_2)\right|$. Note at this point that the pairs might not be one-to-one because the closest distance might suggest to merge two subtitle pairs. Next, we filter out mappings which do not satisfy ($D_1 < T$ and $D_2 < T$) or ($D_3 < T$ and $D_4 < T$). T is chosen empirically by maximizing the performance on a development set. The last step is important in checking for possible subtitle pairs that might not be modeled by the estimated relation.

The time-stamp mapping block is shown in the dashed rectangle (f) of Fig. 3.1. This block takes as input the A, E-linear slope, intercept and the subtitles documents, maps the subtitles based on the closest translated time-stamps, filters the mappings with distance greater than T and, finally, the outputs a subset of the mappings by filtering non-matching subtitles based on the approach described above.

3.2.2.1.6 Mappings merging Finally, we need a method to merge many-to-one, one-to-many, and many-to-many mappings because, in practice, there may not be a clear pair boundary between bilingual subtitles in L_1 and L_2 subtitle documents. The

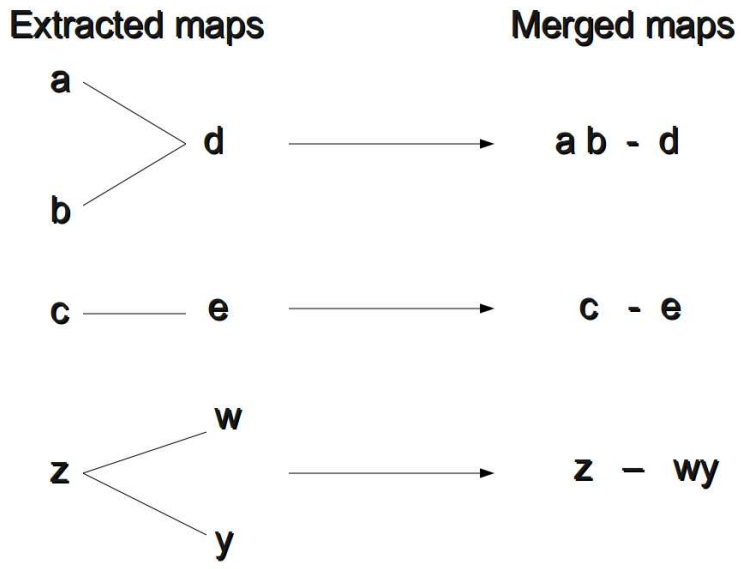

Figure 3.4: Rules for merging extracted maps.

goal is to identify many-to-one, one-to-many, and many-to-many mappings and merge them. Fig. 3.4 shows the fundamental rules used to merge two-to-one, and one-to-two mappings. For example, if subtitles a and b in the L_1 subtitle document are mapped to subtitle d in the L_2 subtitle document, we merge a and b subtitles and map them to d subtitle. This merging defines a two-to-one mapping. Similarly, the other rules define one-to-one and one-to-two mappings. To merge the subtitles in L_1 and L_2 subtitle documents, we apply recursively the rules shown in Fig. 3.4 for all subtitles in L_1 and L_2 documents until no subtitles can be merged. Fig. 3.5 shows an example of a three-to-three mapping merging. The above-mentioned basic rules are applied recursively until only the one-to-one rule can be applied. In this example, first we merge f and g subtitles in the L_1 subtitle document using the rule for merging two-to-one mappings.

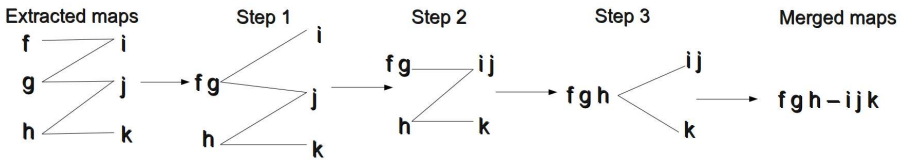

Figure 3.5: Illustrative example of the mappings merging algorithm.

We continue in this fashion until f,g and h subtitles in the L_1 subtitle document are mapped to i, j and k subtitles in the L_2 subtitle document as shown in Fig. 3.5 .

While Fig. 3.4 shows a closer look into how the mappings merging rules are applied, the integration of the mapping merging block into the algorithm is shown in the dashed rectangle (g) of Fig. 3.1. As shown in the diagram the input are the filtered aligned mappings and the output are the aligned merged mappings.

3.3 Experimental results

In this section, we describe the data collection and the experimental results. The experiments are divided into two sections: the pilot and the full-scale experiments. The pilot study using a small set of tagged bilingual mappings was used to understand the parameter trade-offs related to performance. Moreover, the pilot experiments section serves as a development set to optimize the parameters of the time-alignment approach. Finally, the full-scale experiments use the optimal parameters obtained by the pilot study and expand the experiments by aligning a large set of untagged bilingual subtitle document pairs. The aligned data are used to train a SMT system. Finally, the SMT performance is tested on the extracted bilingual sets and the BLEU score performance is reported.

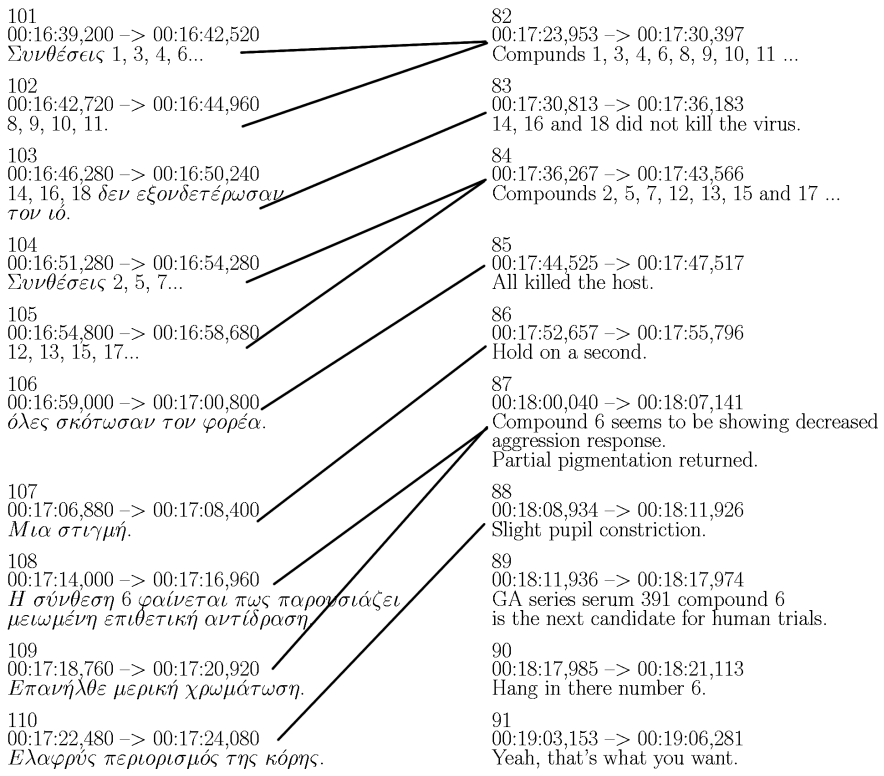

Figure 3.6: This is an illustrative example of the reference mappings from the movie "I am Legend".

3.3.1 Pilot experiments

3.3.1.1 Experimental setup

For the pilot experiments, we used the 42 Greek-English subtitle document pairs described in [11]. In each subtitle document pair, a set of 40 consecutive English subtitles were paired with the corresponding Greek subtitles and we ended up with 1680 tagged pairs. The English subtitle documents have 1443 subtitles on average per movie with standard deviation 369. On the other hand, the Greek subtitle documents contain 1262 subtitles on average with standard deviation 334. The difference in the average number of subtitles indicates that subtitles in bilingual subtitle document pairs may not always

have one-to-one correspondence. A typical example of an aligned bilingual subtitle is shown in Fig. 3.6, obtained from the movie "I am Legend".

All subtitle documents are preprocessed and filtered from non-alphanumeric symbols similar to what one would do for cleaning text for statistical machine translation purposes. Then, the time-stamps and subtitle numbers are removed resulting in a list of Greek subtitles and a list of English subtitles per subtitle document. Each subtitle time-stamp is saved separately as well. For all Greek words available, a system was built to mine all the translations returned by the Google dictionary [1]. Using the dictionary, each Greek subtitle is converted from Greek into a bag of words in English. Then, the RFDM is computed for all subtitle pairs. The best mappings are extracted using the DTW approach described in section .3. The parameters used in the DTW approach are the same parameters used by [11] since the data-sets are identical. Lastly, the method used to merge one-to-one, many-to-one, and one-to-many subtitle pairs is applied to also merge the subtitles of the DTW approach as described in section 3.2.2.1.

The mappings obtained by the DTW approach are used to estimate the A, E-linear function and, in turn, use the function to align the subtitles. Initially, the pairs are ranked in ascending order of RFDM values. For various experimental values of $K\%$=[0.01 0.02 0.03 0.05 0.07 0.1 0.15 0.2 0.4 0.6], the $K\%$ lowest RFDM one-to-one mappings are extracted for each bilingual subtitle pair. For the i^{th} bilingual subtitle document pair, keeping only one-to-one mappings results in M_i mappings. Next, by varying A=[1.05 1.1 1.15 1.22 1.3 1.42 1.5 1.7 1.9 2.2], a subset of the one-to-one mappings of order N_i is used to estimate the A, E-linear function of order N_i for each bilingual subtitle document pair. Then, for different values of E=[0.1 0.2 0.3 0.4 0.5 0.6], the A, E-linear relation is accepted or rejected if $\mathcal{E}_i \leq E$. The starting and ending time-stamps are mapped using the closest distance rule as described in section 3.2.2.1.

[1] "http://www.google.com/dictionary"

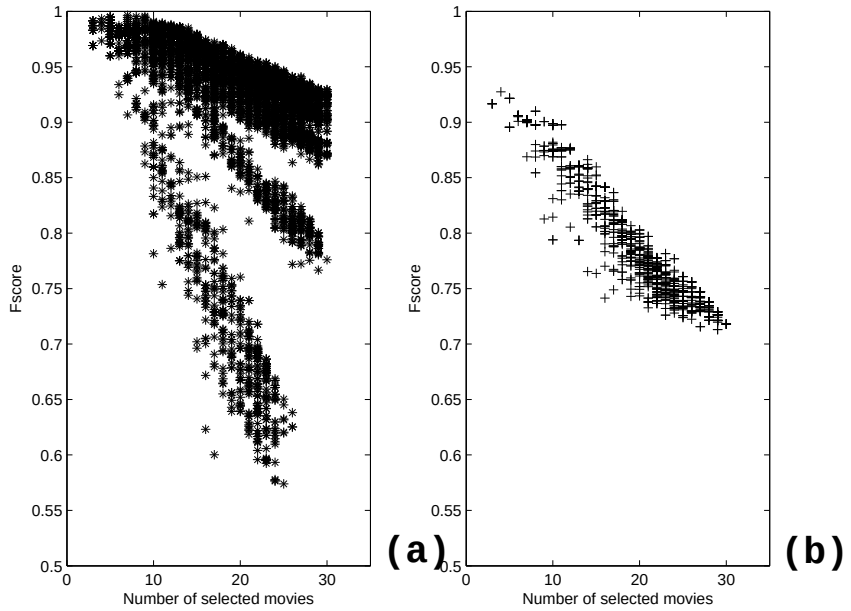

Figure 3.7: Figure (a) shows the averaged F-Score of the time-alignment approach vs the number of movies for various K,A,E and T parameter values. Figure (b) represents the averaged F-Score using the DTW approach for the different number of movies considered when varying the K,A,E, and T parameter values.

Finally, for different values of T=[0.2 0.5 0.8 1 1.5 1.8 2 2.5 3 5], outliers are filtered. The final mappings are obtained using the method to merge one-to-one, many-to-one, and one-to-many subtitle pairs as described in section 3.2.2.1. For each combination of the parameters K, A, E, and T, we compute the balanced F-Score [60, p. 156] averaged over all bilingual subtitle document pairs and the number of considered movies.

3.3.1.2 Results and discussion of pilot study

In this section, we aim to understand the trade-offs among the time-alignment approach parameters.

Fig. 3.7(a) shows the averaged F-Score (vertical axis) and the corresponding number of movies (horizontal axis) for different K, A, E and T parameter values. Clearly, Fig.

3.7(a) indicates that we can get an F-Score close to 1 for some K, A, E, and T values. On the other hand, Fig. 3.7(b) indicates that the F-Score of the DTW-based approach is much lower than that of the time-alignment case, considering the same number of movies. For example, when we consider the parameters aligning bilingual subtitle documents of 30 movies, the DTW-based approach F-Score is less than 0.75 as opposed to the time alignment approach in which the F-Score is close to 0.95. Furthermore, Fig. 3.7(a) suggests that there is a trade-off between the quality of the alignments (i.e. F-Score) and the number of movies used. Thus, one should consider the amount of data and the quality of the bilingual subtitle pairs needed. Based on the quality and amount of data needed, the appropriate K, A, E, and T values can be assigned. To understand the importance of the α, ϵ-linear functions and the associated parameters K, A, E, and T in relating the time-stamps, we also computed the F-Score using a linear relation estimated by the results in .4 using all the DTW output mappings. For this case, the resulting F-Score was 0.56 which is even below the DTW-based approach.

Fig. 3.8 is a 5-Dimensional diagram representing the F-Score as intensity against the values of K, A, E, and T parameters. Similarly, intensity in Fig. 3.9 represents the number of movies aligned for each set of threshold values and, thus, is an indicator of the amount of parallel data extracted. An important parameter is the absolute error threshold, E, used to accept or reject the A, E-linear function alignment for the corresponding movie. Decreasing the absolute error threshold, E, the F-Score increases but at the same time, as Fig. 3.9 suggests, the number of movies aligned decreases. In addition, the choice of the duration ratio threshold, A, becomes less important in filtering the incorrect DTW mappings when a low error threshold is used. This happens because the subtitle pairs, kept with low error, have approximately linearly-related time-stamps obtained by the correct DTW mappings. In spite of giving high F-Scores, the number of movies aligned is much less as E decreases. On the other hand, as

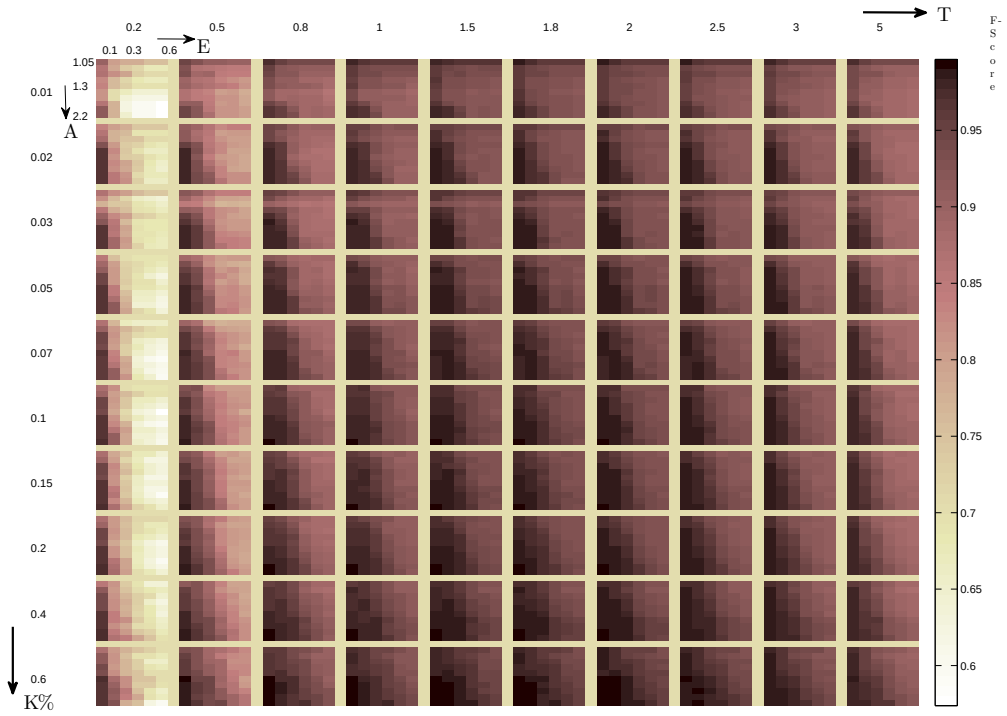

Figure 3.8: This figure shows the F-Score of the time alignment approach for various values of K,A,E, and T parameters.

the threshold E increases and as the threshold on duration ratio, A, approaches 1, the performance decreases but the number of movies modeled increases. The trade-off between A and E is important to consider in aligning subtitle documents. In practice, it is preferable to allow an absolute error threshold, E, greater than 0.4 and a duration ratio threshold, A, less than 1.6 since they maintain not only high F-Scores but also more bilingual data compared to the case with low E and high A. Intuitively, one can think that it is preferable to select accurate mappings at an earlier stage so that we can better estimate the A, E-linear function parameters. Allowing inaccurate mappings results in a higher absolute error, $\mathcal{E}$ and, thus, subtitle document pairs are dropped by the E threshold. Hence, the amount of bilingual data is reduced. If the quality of the

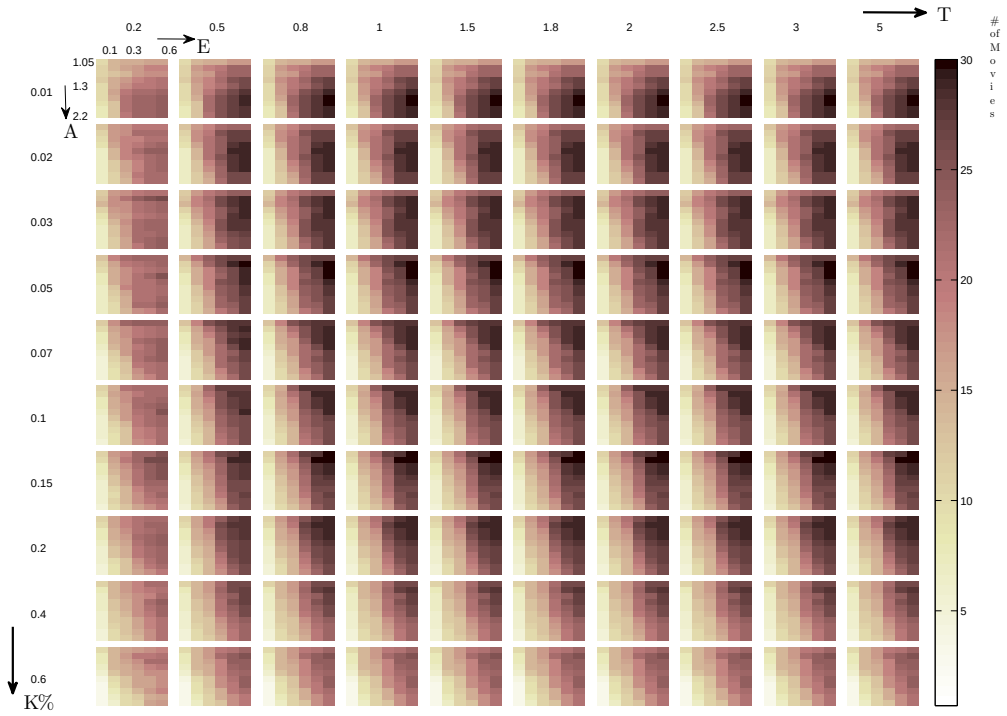

Figure 3.9: The intensity in this figure shows the number of movies modeled by the time alignment approach for various values of K,A,E, and T parameters.

alignment is more important than the size of the corpus, then a low E and A should be considered.

Moreover, Fig. 3.8 suggests that increasing K, increases the F-Score as well. However, the F-Score increase rate is almost flat when $K > 0.1$. On the other hand, increasing K above 0.2 reduces the number of movies aligned using A, E-linear functions and, in turn, decreases the bilingual data. The rationale behind this fact is that K increases the number of DTW mappings used. Since we are choosing the mappings based on the RFDM score in increasing order, the more DTW mappings considered, the higher the RFDM score of the mappings considered in which we are less confident about their accuracy according to the RFDM score. Since the threshold $\frac{1}{A} < \left|\frac{y_{2i}-y_{1i}}{x_{2i}-x_{1i}}\right| < A$ might not

always filter the misaligned mappings as Fig. 3.2 suggests, it will be preferable to choose the most reliable mappings with the lower RFDM score. Including possibly misaligned mappings, i.e. high RFDM score mappings, increases the error and, thus, reduces the number of subtitles accepted by the E threshold. However, if $K\%$ is high and the E is low, it suggests that the $K\%$ of the DTW mappings can be related with an almost linear relation and, thus, for those subtitle pairs the estimation of the A, E-linear function parameters is accurate resulting in a higher F-Score. Hence, another trade-off to consider that affects the quality and the size of the extracted bilingual corpus is between the thresholds K, E and A.

Finally, the value of absolute error of the starting and ending times differences threshold, T, takes place after accepting or rejecting the alignment of each bilingual subtitle movie. Fig. 3.9 shows the number of movies aligned is the same across all values of T for a specific value of K, E, and A. Hence, T does not affect the number of movies considered. However, Fig. 3.8 suggests that choosing a very low value of T reduces the F-Score. In this case, the F-Score is reduced because recall is reduced and precision remains close to 1 as T decreases below 1. On the other hand, as T increases above 3, the precision decreases and the recall remains close to 1 resulting into a lower F-Score. Fig. 3.8 suggests that $1 \leq T \leq 3$ maximizes the F-Score.

The absolute error, $\mathcal{E}$, plays an important role in deciding if the A, E-linear function can model the time-stamps' relation. Thus, it is interesting to study the relationship between the absolute error, $\mathcal{E}$, and the quality of the mappings. For this reason, we set $K = 0.6$ and $A = 1.5$ which are the optimal parameters for maximizing the F-Score when 24 subtitle document pairs are selected. Using these parameters, we compute the absolute error, $\mathcal{E}$, of the A, E-linear function. Fig. 3.10 suggests that there is a trade-off between the quality of the alignments and the absolute error, $\mathcal{E}$. In practice, a low absolute error results in a higher F-Score, precision, and recall. In particular, for

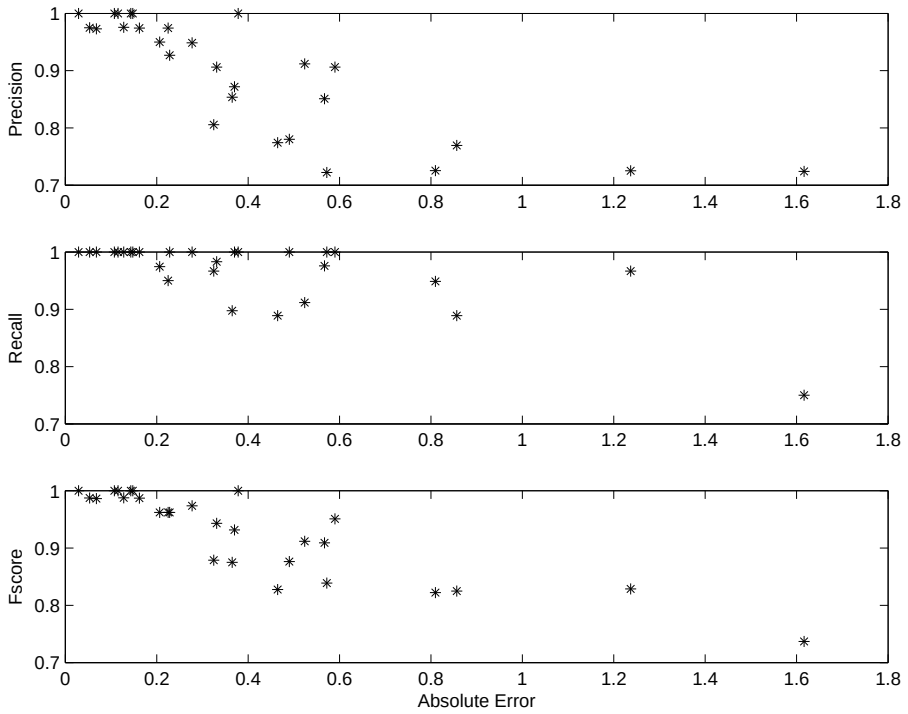

Figure 3.10: The first, second, and third sub-figures show the Precision, Recall, and F-Score vs the Absolute error respectively. Points with an error more than 1.65 are not shown in this figure. Absolute error beyond 1.65 greatly reduces the F-Score.

absolute error, $\mathcal{E}$, less than 0.2, we get almost perfect mappings with F-Score close to 1 due to aligning movies with almost linearly related time-stamps. Fig. 3.10 also justifies the fact that reducing the error threshold, E, increases the F-Score but, on the other hand, decreases the number of movies aligned because fewer subtitle document pairs will satisfy the E threshold.

After analyzing the trade-offs between the various parameters, we choose two sets of parameters for the full-scale experiments. The first set of parameters is fixed to $K = 0.6$, $A = 1.5$, $E = 0.6$ and $T = 2$. This set is denoted by TA-1. For TA-1 pilot experiments, the F-Score is 0.95, precision is 0.92, and recall is 0.98. The number of movies modeled by TA-1 parameters is 24 movies. The corresponding DTW approach F-Score for the

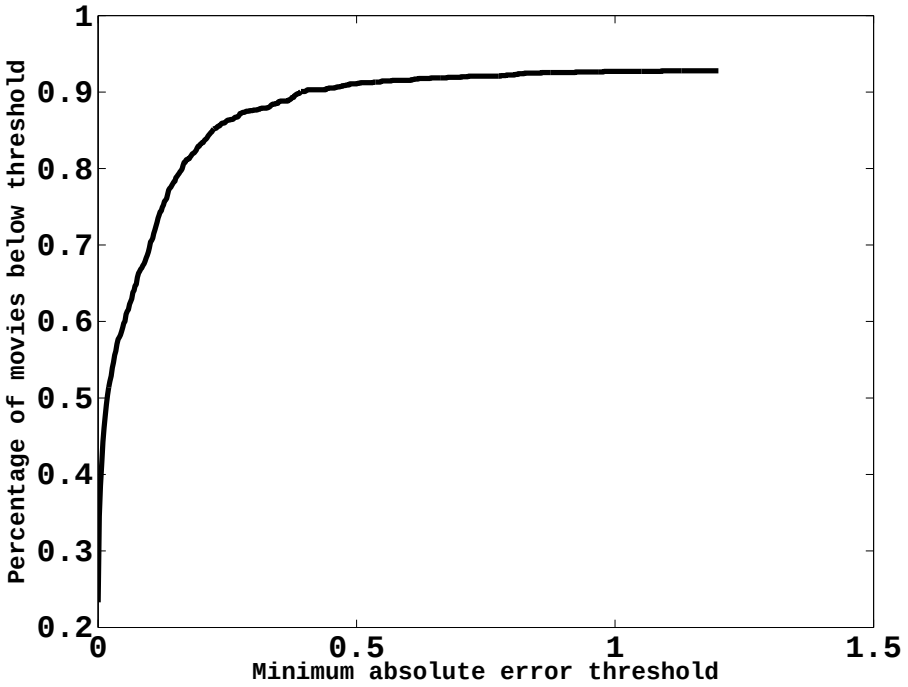

Figure 3.11: The figure shows the percentage of the movies having at least one subtitle document pair with error less than the error threshold.

24 movies considered is 0.75. The second set of parameters produces alignments of less quality than TA-1 but more data. In particular, the second set of parameters is fixed to $K = 0.15$, $A = 1.1$, $E = 0.5$, and $T = 1.5$. This set is denoted by TA-2. For TA-2 pilot experiments, the F-Score is 0.93, precision is 0.92, and recall 0.94. The number of movies aligned is 30. The corresponding DTW approach F-Score for the 30 movies considered is 0.72.

3.3.2 Full-scale experiments

3.3.2.1 Experimental setup

For the full-scale experiments, we downloaded Spanish-English and French-English subtitle document pairs (http://www.opensubtitles.org/). For the Spanish-English subtitle

document pairs, we collected 1758 Spanish subtitle documents and 1936 English subtitle documents. Note that these come from 699 unique movies. By combining all possible document pairs of movies, we end up with 4921 Spanish-English subtitle pairs including repeated subtitle documents for some movies.

On the other hand, for the French-English subtitle document pairs, we collected 1745 French and 2145 English movie subtitle documents out of 641 unique movies. By combining all possible document pairs, we end up with 5967 French-English subtitle document pairs including repeated subtitle documents for some movies.

For the above-mentioned subtitle documents, the non-alphanumeric symbols were filtered for all the subtitle documents. In addition, for all Spanish and French words available in the Spanish and French subtitle documents, we queried the Google dictionary and saved all the available English translations. Then, the bilingual subtitle documents pairs are aligned using the DTW procedure as described in section 3.2.1 and the DTW mappings are obtained. Using the DTW mappings, the subtitle document pairs are aligned using the time-alignment algorithm described in section 3.2.2. The time-alignment approach was run twice using the TA-1 and TA-2 parameters.

Since there are multiple subtitle document versions for each movie available, we can use the quality measures of the proposed approach to find the subtitle documents pair for each movie that maximizes the performance. Thus, among the multiple subtitle document pairs per movie, we select the subtitle document pair giving the lowest absolute error, $\mathcal{E}$. Because the DTW baseline has no quality tests to accept or reject alignments, we randomly pick a subtitle document pair for each movie to align. Fig. 3.11 implies that there are approximately 95% of the movies having at least one bilingual subtitle document pair with absolute error $\mathcal{E} < 1$. Hence, for the proposed approach, we align the subtitle pairs for each movie with the lowest error. The parameters used in Fig.

3.11 are $K = 0.15$ and $A = 1.1$ which are the parameters of TA-2. Using the K and A parameters of TA-1 yields similar results.

Finally, using parallel data from a corpus from aligned movie subtitles, we train the SMT models on each language pair separately. Experiments using the SMT trained on the TA-1, TA-2, and DTW corpora are denoted by TA-1, TA-2 and DTW respectively. Moreover, 2000 randomly picked utterances for tuning and 2000 randomly picked utterances for testing were used to evaluate the performance from the DARPA TRANSTAC English-Farsi data set. Only the English utterances were extracted and manually translated to Spanish and French for evaluating the performance. TRANSTAC is a protection domain corpus (e.g. dialogs encountered at military points). The randomly picked subset includes conversations of a spontaneous nature; for example, there are spontaneous discussions on various topics such as medical assistance related conversations, etc. Tuning and evaluation on this set is denoted by TRANSTAC. In addition, the development and test sets of the News Commentary corpus[2] have been used to evaluate the experiments. We refer to the NEWS development and test set as NEWS-TEST.

The SMT requires language models of the target language to translate the source utterances. In each experiment, the training set of the target language is used to train the language models for each experiment as well. The trigram language models were built using the SRILM toolkit [92] and smoothed using the Kneser-Ney discount method [53]. We compared the performance of various combinations and sizes of the training sets using BLEU score [70] on the TRANSTAC and NEWS test sets.

3.3.2.2 Results and discussion

Figs. 3.12 and 6.3 compare the performance of the SMT models obtained by training on the corpora extracted by the time-alignment approach and that extracted by the

[2]Made available for the WMT10 workshop shared task http://www.statmt.org/wmt10/

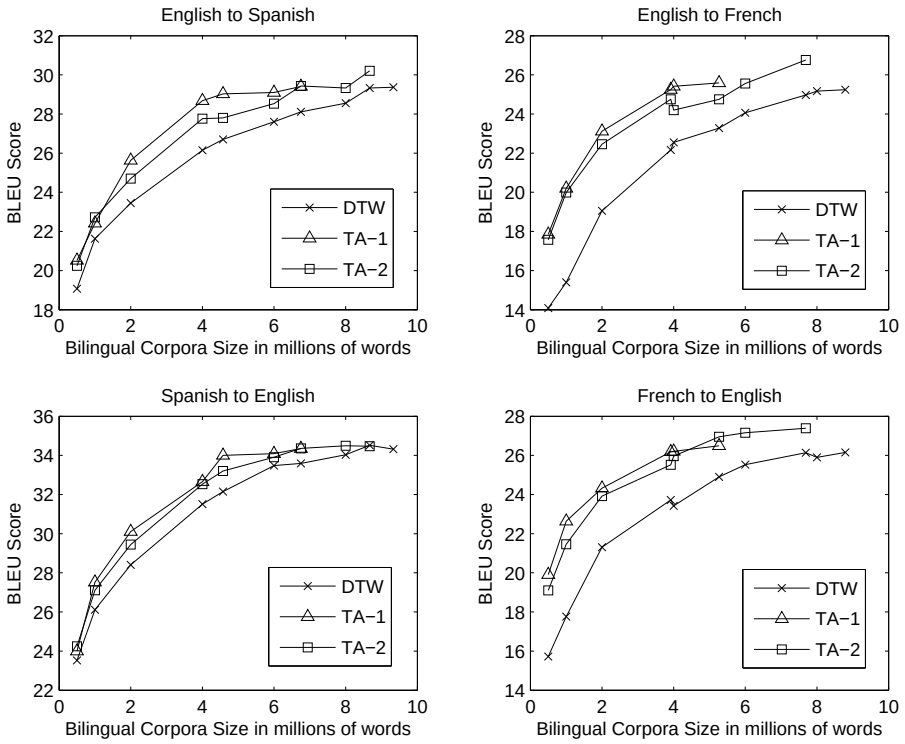

Figure 3.12: This figure compares the performance of the SMT models trained on the corpus created using the DTW-based approach and the models trained on the corpora extracted by the time-alignment approach with parameters TA-1 and TA-2 when the TRANSTAC development and test sets are considered. The experiments were repeated for various bilingual corpora sizes. The comparison is extended for the language pairs between English-Spanish, English-French, and vice versa.

DTW-based approach in the TRANSTAC and NEWS-TEST domains. In addition, the comparison is extended into four language pairs, namely, English to Spanish, English to French, and vice versa.

In Fig. 3.12, the goal is to compare the quality of the alignments in a spontaneous speaking style domain and, hence, the TRANSTAC domain is used for tuning and evaluating. The figure shows the performance gains of the models trained on the TA-1 and TA-2 corpora over the models trained on DTW-based approach corpus. In particular,

the performance of TA-1 and TA-2 corpora is very close in terms of BLEU score; however, the parameters used in TA-2 could extract a larger bilingual corpus as shown in Fig. 3.12. In these experiments, the time alignment approach corpora consistently outperforms the DTW-based approach corpus across different language pairs and different bilingual corpus sizes by up to 2.53 BLEU score points for the English-Spanish experiments and by up to 4.88 BLEU score points for the English-French experiments. The improvement stems from the fact that the TA-1 and TA-2 corpora approaches have been shown in section 3.3.1.2 to deliver F-Scores close to 96% as opposed to the DTW-based approach corpus which is expected to deliver F-Scores of 71% [11]. Thus, for a fixed amount of subtitle pairs, the F-Score improvement of the alignment is translated into SMT performance boost showing the importance of the time-alignment based approach.

In Fig. 6.3, the goal is to compare the quality of the alignments in the broadcast news domain by using the NEWS test set. Similar to the TRANSTAC test set results, Fig. 6.3 indicates that SMT models trained on TA-1 and TA-2 outperform those trained on the corpus created using the DTW-based approach. We observe performance improvements of up to 1.2 BLEU score points for the English-Spanish experiments and by up to 2.65 BLEU score points for the English-French experiments. The performance improvement is consistent along all of the different bilingual corpus sizes. These experiments suggest that the time-alignment approach is superior to the DTW-based approach across different domains in terms of the SMT performance. We note that the F-Score improvement delivered by the time-alignment approach is reflected even in domains not matching the subtitles speaking style such as in the NEWS-TEST domain.

For readers interested in the performance of subtitles for the TRANSTAC domain compared to additional corpora could refer to .5.

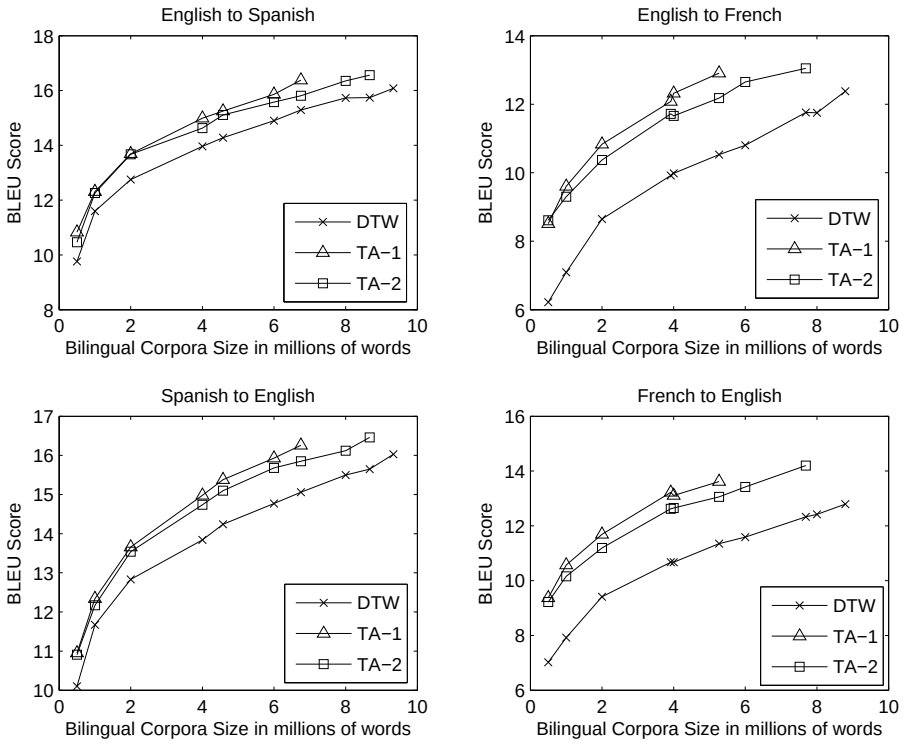

Figure 3.13: This figure compares the performance of the SMT models trained on the corpus created using the DTW-based approach and the models trained on the corpora extracted by the time-alignment approach with parameters TA-1 and TA-2 when the NEWS-TEST development and test sets are considered. The experiments were repeated for various bilingual corpora sizes. The comparison is extended for the language pairs between English-Spanish, English-French, and vice versa.

3.4 Conclusions

In this work, we presented a novel method to align subtitle documents using timing information. We modeled the time-stamps relation using the so-called α, ϵ-linear functions. In section 3.2.2, we presented some properties of the α, ϵ-linear functions. Then, we took advantage of the properties of the α, ϵ-linear functions to find an approximately

linear relation between the time-stamps of two subtitles. In addition, we analyzed practical aspects of the α, ϵ-linear functions in modeling the subtitles alignment problem and showed the advantages over the standard DTW-based scheme.

Moreover, we have shown the importance of various parameters and the trade-offs associated with the quantity and quality of the data. The threshold K plays the role of selecting the most reliable mappings from the DTW output with respect to the RFDM score. Additionally, we have shown the importance of the threshold A in filtering unreliable mappings with low RFDM score. On the other hand, the threshold E has been used to exclude subtitles with error $\mathcal{E} \geq E$ in order to filter out subtitle pairs that cannot be modeled by the α, ϵ-linear function. In this work, the error was estimated using the absolute norm. For the sake of completeness, we have experimented with various norms, i.e. Euclidean norm, to compute the error, but no significant improvement was observed in terms of the F-Score. Furthermore, a threshold on the start and end time differences has been imposed to filter out the outliers. Lastly, a method has been presented to join many-to-many subtitle pairs.

In the experiments section, we analyzed and compared the performance of the time-alignment approach and the DTW-based approach in a small set and showed improvement in terms of F-Score. Moreover, we showed that the F-Score improves by 39% (absolute) using the α, ϵ-linear function approach compared to a line fit obtained by linear regression on all DTW output pairs, indicating the importance of using the α, ϵ-linear function modeling. In addition, we have shown that our approach outperforms the DTW-based approach by up to 4.88 BLEU score points (approx. 25% relative improvement) on spontaneous speech Statistical Machine Translation (SMT) experiments. Overall the proposed approach can provide high quality bilingual translations by linearly relating the subtitle time-stamps and choosing parameters to set the quality and the quantity of the data extracted.

Chapter 4:

Parallel bilingual speech data extraction

Works presented in this chapter have been carried out in collaboration with Prasanta Gosh

4.1 Bilingual audio-subtitle extraction using automatic segmentation of movie audio [12]

Extraction of bilingual audio and text data is crucial for designing Speech to Speech (S2S) systems. In this work, we propose an automatic method to segment multilingual audio streams from movies. In addition, the audio streams are aligned with the corresponding subtitles. We found that the proposed method gives 89% perfectly segmented bilingual audio and 6% partially segmented bilingual audio. In addition, the mapping of the audio to the corresponding subtitles has accuracy 91%.

4.1.1 Introduction

One of the critical areas of research related to the development of Speech-to-speech (S2S) translation systems has been on techniques to identify and acquire parallel data. It is hence not surprising that given this heavy dependence on bilingual data for system design, breakthroughs in S2S system performance and capabilities have accompanied the increasing availability of bilingual data for training the S2S systems. The critical role of the bilingual data has led the S2S community to acquire bilingual data using both manual efforts and automatic algorithms. Such data include spoken utterances translated by interpreters and translation of speech transcriptions.

Manually translated speech corpora that have been used for speech translation include the Europarl [54] and the news commentary corpus[1]. It should be noted that many of these data do not adequately represent the conversational aspects of human interaction. Various methods have been also proposed to extract bilingual parallel corpora automatically from available, incidental, resources. Some of these methods have focused on aligning movie subtitles. Past works on movie subtitles have demonstrated the importance of such data for S2S systems. Sarikaya *et al.* [87] showed BLUE score [70] improvements on a large-scale S2S system. In our past work [11], we focused in linking the speech transcriptions of movies as shown in Fig. 4.1. However, parallel speech transcriptions limit the information that can be contained in the training data of S2S systems.

In addition to manually translated speech transcriptions, the S2S community extensively has used manually-translated speech audio. Such audio and text corpora examples include DARPA TRANSTAC domain data [88] and Basic Travel Expression Corpus (BTEC) [95]. However, very little work has been done to automatically acquire and align bilingual speech utterances. In this work, we focus on segmenting parallel

[1]Made available for the workshop shared task http://www.statmt.org/wmt10/

audio from movies and aligning the segments with the corresponding subtitles as shown in Fig. 4.1. Thus, the goal is to find parallel audio segments from a movie containing multilingual audio streams which can have many potential S2S uses, for example, it can be a rich source for analyzing various acoustic cues across languages that can potentially bring more naturality in S2S translation.

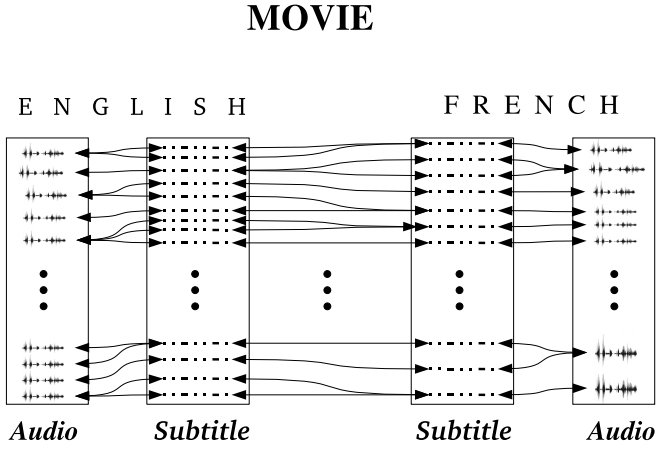

Figure 4.1: An illustration of bilingual audio streams and subtitles alignment and segmentation between English and French.

This work is structured as follows. In section 2, we present the collected data used in this work. In section 3, we describe and analyze the features related to this work. In section 4, we explain the proposed approach. Section 5 discusses the results of our approach and, finally, in section 6, we summarize the results of this work and provide some future directions.

4.1.2 Data collection

For the purpose of these experiments, we collected 5 movies containing audio and subtitles in English and French. Since movies usually contain audio in multiple channels, we down-mix all channels to one channel. We manually tagged 2 hours and 30 minutes

of parallel audio data from these 5 movies and extracted 1050 parallel English-French speech segments. Then, we marked the bilingual speech segments not only at points that speech exists in both languages but also at bilingual speech segments that are translations of each other. In addition, we manually mapped the speech segments to the corresponding subtitles and merged subtitles, if necessary. An overview of the the manual audio segmentation and subtitle alignment tagging is shown in Fig. 4.7.

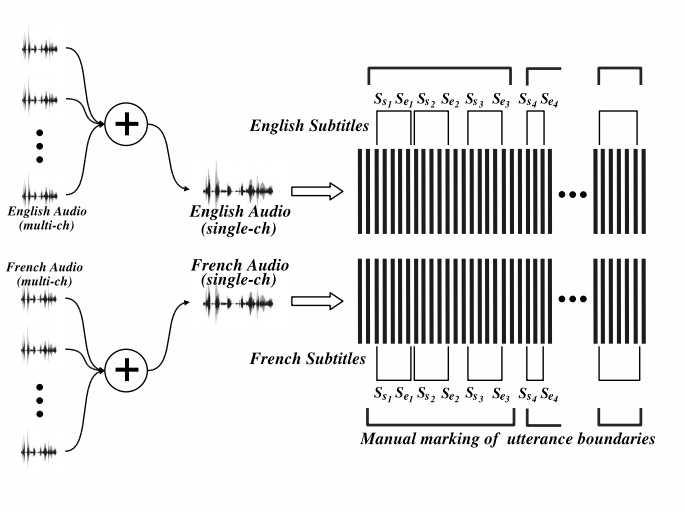

Figure 4.2: An illustration of the manually tagged bilingual audio streams for English and French.

4.1.3 Proposed features

To design the features for identifying bilingual speech segments in movies, we need to understand some important properties movies have. Since the movie audio has to match the video, the movies have parallel audio streams in various languages of exactly the same size. In addition, the speech signal in different languages is approximately at the same time locations since the audio has to match the video scene in the movie. Moreover, if speech is not present the background sound is approximately the same for all language streams so that the audio experience is not altered. Also, the subtitles are

close to the corresponding speech segments so that they match the spoken dialogues with the video scenes. Thus, using the properties described, we define the Long Term Spectral Distance, Subtitles time distance and subtitle time-stamps to segment the bilingual speech regions.

4.1.3.1 Long Term Spectral Distance

The Long Term Spectral Distance (LTSD) can be used to capture the acoustic distance between two segments of audio of R short-term frames. In our problem, the motivation for using LTSD is to find regions of acoustic similarity in the speech streams in a longer term basis. Firstly, the audio streams of both languages, say L_1 and L_2, are segmented into short-term frames. The frame streams are denoted by $F_{L_1}(m)$ and $F_{L_2}(m)$, where m is the frame index. For each frame, we compute the LTSD by:

$$LTSD(m) = \sum_{i=m-R}^{m+R} D(F_{L_1}(i), F_{L_2}(i))$$

where R is the window used to compute the long-term distance.

$D(F_{L_1}(i), F_{L_2}(i))$ represents the spectral distance between the two frame streams. This distance takes high values when the audio streams differ. For example, the LTSD value is low when the frame streams contain only background noise, since the spectrum is expected to be the same in the two bilingual audio streams. The LTSD takes high values when there is speech in the audio streams. In this case, since both stream contain speech in different languages, the spectral distance is expected to be high. Fig. 4.3 indicates that the LTSD values differ significantly for parallel speech and non-speech regions.

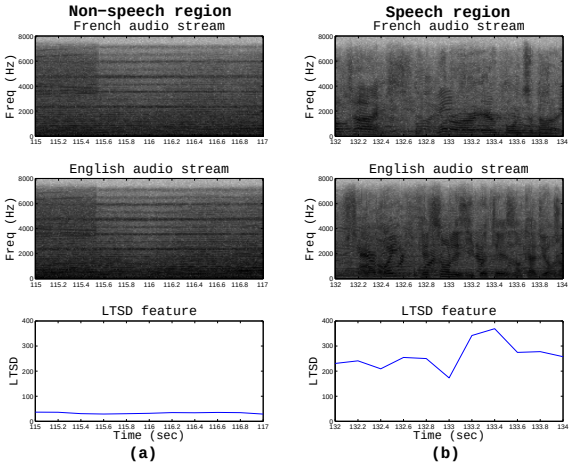

Figure 4.3: Fig. 4.3(a) shows the spectrogram of French and English non-speech audio regions along with the value of LTSD. Fig. 4.3(b) shows the spectrogram of French and English speech audio regions along with the value of LTSD.

4.1.3.2 Spectral distance

To approximate the acoustical and perceptual proximity of the two audio streams, we use the mel-frequency cepstral coefficients (MFCC) [107], excluding the zero-th coefficient to make the spectral distance independent of energy levels. Thus, the remaining MFCCs capture the spectral variability over different frequency bands for each short-time frame. We denote the MFCCs coefficient vectors of the i^{th} frame by $c_{L_1}(i)$ and $c_{L_2}(i)$ for the frame streams in L_1 and L_2 respectively.

Thus, the spectral distance for the i^{th} frame is defined by

$$D(F_{L_1}(i), F_{L_2}(i)) = ||c_{L_1}(i) - c_{L_2}(i)||^2$$

4.1.3.3 Analysis

In this section, we present details of pilot experiments conducted to check the effectiveness of the LTSD for segmenting bilingual audio speech. Fig. 4.4 suggests that,

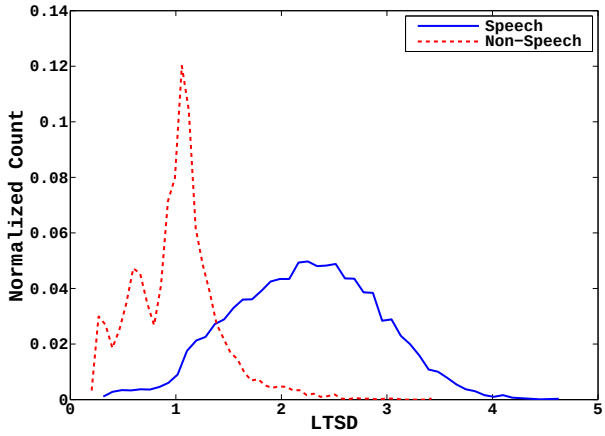

Figure 4.4: Distribution of LTSD for speech and non-speech frames

at the frame level, the LTSD can distinguish speech and non-speech frames with accuracy 87.83% at the equal error rate point (EER) indicating that the LTSD can play an important role in discriminating speech and non-speech regions.

Furthermore, based on the LTSD, we segmented the signal by using a threshold obtained at the EER and manually evaluated the resulting segments. The segments are tagged as correct if they contain one or more bilingual spoken utterances that are all translations of each other. If the spoken utterances are partially translations of each other or do not match at all, they are tagged as wrong. Table 4.1 shows the percentage of correct segments for two different segmentation schemes. The first one is based on the LTSD threshold at the EER where an uptrend in LTSD crossing the threshold is marked as starting point of the segment and a downtrend in LTSD crossing the threshold is marked as ending point of the segment. The second segmentation scheme is based on the starting and ending points of the subtitle time-stamps but as the results indicate the subtitles may not provide exact beginning and ending times for the corresponding speech utterance. The LTSD performs better in segmenting the audio signal for parallel speech segments, however, the subtitles provide information

	Accuracy
LTSD	41.39%
Subtitles	37.88%

Table 4.1: Table shows the percentage of segments that gave perfect segmentations using subtitles and the LTSD feature.

about the text and its approximate location in the audio signal. A disadvantage of segmenting the signal based on LTSD is that many spurious segments occur with a very small duration which penalizes the performance of the LTSD segmentation. However, the relative advantages and disadvantages of both LTSD and subtitle motivated us to improve the performance by combining them both and gain additional information in aligning and segmenting the parallel audio streams.

4.1.3.4 Subtitles time distance

Another feature used in this work is the distance between the starting and ending points of two consecutive subtitles. We refer to the subtitle time distance by STD. If we denote the l^{th} subtitle starting and ending points by S_{s_l} and S_{e_l}, as shown in Fig. 4.7, then the STD between the l^{th} and $(l+1)^{th}$ frame is defined as:

$$STD(l) = S_{s_{l+1}} - S_{e_l}$$

4.1.4 Cross-language automatic audio segmentation and alignment

Initially, we use the manually labeled data to identify subtitles that have to be merged. Each segment, from S_{e_l} to $S_{s_{l+1}}$, is a candidate for merging or splitting the parallel audio streams. We use the K-nearest neighbor (K-NN) classifier to take this decision based on the manually tagged data. The features used are the STD feature and the

minimum $LTSD$ value in the segment from S_{e_l} to $S_{s_{l+1}}$. A low $LTSD$ feature might suggest that there is a dip in the spectral distance and, thus, a possible cut.

If the K-NN classifies a segment from S_{e_l} to $S_{s_{l+1}}$ as split, we cut the stream at the lowest $LTSD$ point, otherwise, the subtitles are merged. For practical purposes, if the lowest LTSD time point is beyond 2 seconds from the closest subtitle time-stamp, we cut at the time point corresponding to the minimum LTSD within those 2 seconds. Finally, we map the starting point of the segment to the starting point of the closest starting time-stamp of a subtitle and the ending point of the segment to closest ending time-stamp of a subtitle. If the starting and ending points do not correspond to only one subtitle, we merge the boundary subtitles and all in between subtitles.

4.1.4.1 Experimental evaluation setup

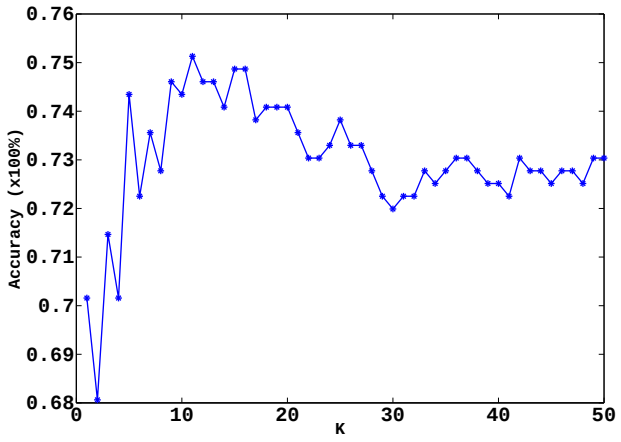

Figure 4.5: K-NN accuracy on the development set for various K.

For experimental purposes, we tagged a small development set to optimize the parameters of the proposed approach. Based on the development set, we found that the EER is minimized when $R = 40$, so we use this value in our experiments. Moreover, Fig. 4.5 shows that on the development set the K-NN performs best when $K = 11$ and, thus,

this value is used in the experiments. Also, the distance used by the K-NN classifier is the mahalanobis distance.

Moreover, we split the data into a training and test set. The proposed approach is applied to the test set to decide if the audio stream is to be split at some point between S_{e_l} and $S_{s_{l+1}}$. Finally, we provided the segments to bilingual human evaluators. The manual evaluation was done to rate the quality of each speech segment in three categories. They report the parallel audio segment as "full", if the parallel audio segment contains bilingual audio streams that are translations of each other. If a subset of the parallel audio is a translation of each other, they rate it as "Partial" and if the parallel audio segments do not match or do not contain speech, they are rated as "None". In addition, they rated the audio alignment with subtitles. They rated it "Full", if the subtitles match the audio stream, "Partial" if the audio stream matches partially the assigned subtitles and "None" if the don't match at all. We refer to the bilingual audio segmentation scheme as "Audio segmentation" and to the alignment of the resulting segmented audio with subtitles as "Subtitle alignment".

4.1.5 Results and discussion

Table 4.2 shows that 89% of the segments contain bilingual spoken utterances and only 6% of the segments contain partial bilingual utterances. On the other hand 91% of the alignment from audio to subtitles is accurate and only 6% contains partial speech or subtitles.

It is important to note that subtitle distance duration plays an important role in deciding if the subtitles are to be merged or split. Additional information is provided by the minimum LTSD at points between consecutive subtitles. LTSD detects dips in the spectral distance and, thus, possible candidate points for splitting the audio streams. Also, the subtitle time-stamps provide an approximate location on where there is actual

	Full	Partial	None
Audio segmentation	89.29 %	5.8%	4.91%
Subtitle alignment	91.42 %	6.44%	2.15%

Table 4.2: Table shows the percentage of the audio segments and subtitle alignments rated as “Full”, “Partial” and “None”.

speech and by searching for a dip in LTSD, we are able to detect exact boundaries of parallel speech starting and ending points. The strong performance of the approach shows that it can be used to produce quality segments of the bilingual audio streams.

Furthermore, it is interesting to study the duration of the resulting segments. Fig. 4.6 shows the normalized histogram of the segments duration. The histogram indicates that more than 80% of the segments have duration less than 10 seconds. Also, the median duration is 4.02 seconds. The resulting segments duration along with the high accuracy of the alignment and segmentation make this approach ideal for using the segmented data in training S2S applications and aligning the speech segments even at the phoneme level, for example, using a force alignment technique.

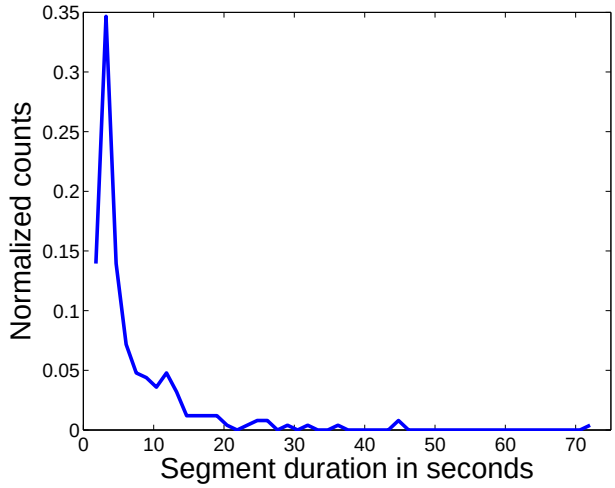

Figure 4.6: Distribution of the duration of the resulting segments.

4.1.6 Conclusions and future work

The goal of this work is to segment bilingual movie audio streams. We proposed the Long Term Spectral Distance (LTSD) feature and enhanced it with information from subtitles to segment the bilingual audio stream. 89% of the resulting segments contained perfect bilingual speech segments. In addition, we aligned the speech segments with subtitles with accuracy 91% on our test set.

For future work, we want to use language information to assess the choice of splitting points, for instance, identifying utterances boundaries. In addition, we aim at detecting noise-free segments and use an alignment method to align the subtitles with the speech signal at a lower level, for example, at the phoneme level.

4.2 Classification of clean and noisy bilingual movie audio for speech-to-speech translation corpora design [14]

Identifying suitable sources of bilingual audio and text data is a crucial part of statistical Speech to Speech (S2S) research and development. Movies, often dubbed in other languages, offer a good source for this purpose; but not all data are directly usable because of noise and other audio condition differences. Hence, automatically selecting the bilingual audio data that are suitable for analysis, and training S2S systems for specific environments becomes crucial. In this work, we extract bilingual speech segments from movies and aim at classifying segments as clean speech or speech with background noise (i.e. music, babble noise etc.). We examine various features in solving this problem and our best performing method delivers accuracy up to 87% in discriminating clean and noisy speech in bilingual data.

4.3 Introduction

Due to the statistical nature of Speech-to-speech (S2S) translation systems, bilingual data have played a significant role in their research and development, for example, bilingual parallel audio have been shown to be important for the translation of paralinguistic cues [101, 102, 96]. Researchers have focused on both manual and automatic data collection approaches for the design of S2S translators. Such bilingual data not only include spoken utterances in the source language along with their interpretation in the target language but also text translation of speech transcriptions. Automatically acquired data could contain speech segments that are not suitable due to low Signal to Noise Ratio (SNR) levels and, thus, making this type of data difficult to use in a S2S system. For this reason, additional research is needed to automatically distinguish low

SNR from high SNR bilingual speech signals which that are suitable for S2S translation design.

Examples of manually obtained bilingual speech transcriptions include the Europarl [54] and the news commentary corpus[2]. In addition to manually collected data, many approaches have been proposed to automatically collect and align bilingual transcriptions. A key component of the automatic algorithms was to model the variability and noise in the alignment of bilingual transcriptions. Such algorithms have been often used to align movie subtitles. For example, Tsiartas et al. [11] focused on alignment speech transcriptions of movie subtitles. Sarikaya et al. [87] selected subsets of bilingual subtitle transcriptions by removing noisy pairs and showed BLUE score [70] improvements on a large-scale S2S system.

Bilingual text transcriptions lack additional information that resides in the audio that may contain important linguistic (e.g., prosody) and paralinguistic (e.g., affect) information for modeling speech translations. Hence, beyond text bilingual data, researchers have been collecting audio bilingual data such as, for example, the DARPA TRANSTAC domain data [88] and Basic Travel Expression Corpus (BTEC) [95]. In addition to manually collected audio bilingual data researchers have proposed approaches to extract bilingual audio data automatically from existing sources. In our past work [12], we had proposed a method to segment bilingual audio from movies and align the segments with the corresponding subtitles.

However, the aforomemntioned method did not distinguish between the quality of bilingual speech (clean or noisy) but instead focused if detecting just the presence of speech. Movie data contain a wide variation in the audio quality and, hence, automatic data selection becomes critical. For this classification task, we use movies that are dubed in at least two languages and propose an approach to classify the bilingual parallel audio

[2]Made available for the workshop shared task http://www.statmt.org/wmt10/

as noisy segments of speech (i.e, background music, gunshots etc) or clean speech. To solve this problem, we exploite the fact that the noise in the two channels is acoustically correlated but the speech signals are not correlated since they are in two different languages. For this purpose, we design a set of diverse features and evaluate their performance on a data set annotated by humans.

This paper is structured as follows. In section 2, we present the collected data used in this work. In section 3, we describe the proposed features used in discriminating low and high SNR speech regions. In section 4, we present the experimental setup. Section 5 discusses the experiments and results of our approach and, finally, in section 6, we summarize the results of this work and provide some future directions.

4.4 Data collection

For the purpose of these experiments, we collected 5 movies containing audio and subtitles in English and French and we down-mix all channels to one channel for each language. Then, we use the approach proposed in [12] to segment the parallel streams of audio into multiple aligned bilingual speech segments. This generates a corpus of 490 bilingual segments. An overview of the the audio segmentation and alignment tagging is shown in Fig. 4.7. These were manually tagged into clean and noisy bilingual speech audio. In the next stage of annotation, we classified segments that contained no background noise as clean bilingual speech and segments with even some noise as noisy bilingual speech. Overall, we obtained 27% clean English-French speech segments with the other 73% tagged as noisy speech segments.

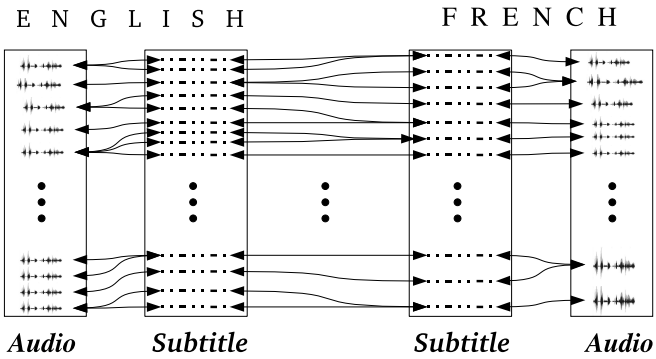

Figure 4.7: An illustration of the automatically segmented bilingual audio streams for English and French. S_{s_i} and S_{e_i} denote the begin and end sample indices for the i^{th} segment

4.5 Proposed features

In this section, we aim to design the features that capture information that discriminate bilingual noisy and clean speech segments based on the SNR levels. To design these features, we need to understand some important properties of the bilingual speech audio. Firstly, the bilingual segment pair contains speech in two different languages in two separate signals. The speech signal may or may not contain noise. Noise can be background music, background babble noise and in general any non-speech audio signal including noise that can be much smaller in duration than the speech segment. Acoustically, noise is similar in both audio streams. In some cases, noise in one audio stream can be a shifted, scaled and maybe filtered version of the noise in the other audio stream. Using the above-mentioned properties, we construct features that capture the spectral correlation (due to the acoustic similarity) between the two audio streams. In addition, we use the first audio stream to predict the second audio stream, thus,

estimate the noise and measure the energy ratio between the estimated noise and both audio streams.

4.5.1 Spectral correlation (SC)

In order to approximate the acoustic and perceptual proximity of the bilingual audio streams, we use the mel-frequency cepstral coefficients (MFCC) [107] to represent the audio signal. Suppose for the i^{th} segment there are R^i frames. To define the spectral correlation, we first concatenate R^i consecutive frames' L-dimensional MFCC feature vector (excluding the DC coefficient). Thus, for each segment, say segment i, we have two vectors (one for each language) of dimension $I^i = R^i L$. Hence, we define the two I^i MFCC feature vectors as $C_{L_1}(i)$ and $C_{L_2}(i)$ for the i^{th} segment. Using these two vectors, we compute the correlation coefficient. The reason we are using the spectral correlation is to capture the spectral similarity of the streams while keeping the feature robust to any scalings and short-term shifting of any of the two audio streams.

Hence, the Spectral Correlation (SC) of the i^{th} segment is defined as:

$$SC(i) = \frac{\sum_{j=1}^{I^i} \left(C_{L_1}(i,j) - \bar{C}_{L_1}(i)\right)\left(C_{L_2}(i,j) - \bar{C}_{L_2}(i)\right)}{\sqrt{\sum_{j=1}^{I^i} \left(C_{L_1}(i,j) - \bar{C}_{L_1}(i)\right)^2 \sum_{j=1}^{I^i} \left(C_{L_2}(i,j) - \bar{C}_{L_2}(i)\right)^2}}$$

where the mean is defined as: $\bar{C}_{L_1}(i) = \sum_{j=1}^{I^i} \frac{(C_{L_1}(i,j))}{I^i}$

Thus, by definition, the closer the acoustic similarity of the two vectors is, the closer $SC(i)$ will be to 1. Thus, a high value of $SC(i)$ indicates the presence of acoustically similar noises in the two audio streams.

4.5.2 Noise to Speech and Noise Ratio (NSNR)

The Noise to Speech and Noise Ratio (NSNR) aims to capture the ratio between the noise that is common in the two channels and the amount speech that is present. For each segment separately, we denote the audio of language L_1 and L_2 as $\mathcal{S}_{L_1}$ and $\mathcal{S}_{L_1}$ respectively. Moreover, we assume the following signal model for the audio signals $\mathcal{S}_{L_1}$ and $\mathcal{S}_{L_1}$: $\mathcal{S}_{L_1} = X_{L_1} + N$ and $\mathcal{S}_{L_2} = h * X_{L_2} + h * N$ where * represents the convolution operator. X_{L_1} and X_{L_2} are the speech signals in the audio streams of language L_1 and L_2 respectively. N is the noise in the audio stream of L_1 and h is a filter. In addition, we assume X_{L_1}, X_{L_2} and N are uncorrelated. To verify this uncorrelated assumption, we computed the correlation coefficient between such signals and we found that the correlation coefficient is very close to 0 (The average correlation coefficient is of the order 10^{-4}). We define NSNR for the i^{th} segment as:

$$\begin{aligned} NSNR &\triangleq \frac{|E\{(h * \mathcal{S}_{L_1})\mathcal{S}_{L_2}\}|}{E\{(h * \mathcal{S}_{L_1} + \mathcal{S}_{L_2})^2\}} \\ &= \frac{\left|E\left\{UC + (h * N)^2\right\}\right|}{E\{(SSN + 4UC - 2(h * X_{L_1})(h * X_{L_2})\}} \\ &= \frac{\left|E\left\{(h * N)^2\right\}\right|}{E\{SSN\}} \end{aligned}$$

where $SSN = (h * X_{L_1})^2 + (h * X_{L_2})^2 + (2h * N)^2$

$UC = (h * X_{L_1})(h * X_{L_2}) + (h * X_{L_1})(h * N) + (h * N)X_{L_2}$ and $E\{UC\} = 0$ because X_{L_1}, X_{L_2} and N are uncorrelated.

For signal X of size K, $E\{X\} = \frac{\sum_j X(j)}{K}$

The above expansion and analysis of the NSNR feature reveal that NSNR takes values closer to 0 when the SNR is very high in both audio streams. On the other hand, NSNR takes values closer to 1 if SNR is very low. From the definition of the NSNR, we need to know $\mathcal{S}_{L_1}$, $\mathcal{S}_{L_2}$ and h. $\mathcal{S}_{L_1}$ and $\mathcal{S}_{L_2}$ are directly known from the data, since they are simply the time domain samples of each bilingual audio segment. However, the

filter h is unknown and, thus, we need to estimate h from the data for each bilingual segment separately.

4.5.3 Filter estimation

To estimate the filter h, we propose two approaches. The first approach is simplistic and faster and assumes that the filter acts on the signal by scaling and shifting it. The second approach tries to estimate a time-varying Least Mean Squares (LMS) filter using the normalized LMS [42] approach.

4.5.3.1 Scaling and Shifting Filter (SSF)

In this case, the assumption is that the filter h is only shifting and scaling the signal. To estimate this signal, we use regions in which there is only noise. Such regions are returned by the algorithm described in [12]. The segments between consecutive speech regions are expected to be noise only. For example, the i^{th} segment has a left and right noise-only region. As Fig. 4.7 shows the left noisy region is between $S_{e_{i-1}}$ and S_{s_i} and the right noisy region of segment i is between S_{e_i} and $S_{s_{i+1}}$.

Now, to compute h using the SSF estimation, we first compute the correlation coefficient for both left and right noisy regions by varying the shift index M as follows:

$$CCleft(i,M) = \frac{\sum_{j=S_{e_{i-1}}+M}^{S_{s_i}} V_{j-M}^{L_1} V_j^{L_2}}{\sqrt{\sum_{j=S_{e_{i-1}}+M}^{S_{s_i}} \left(V_{j-M}^{L_1}\right)^2 \sum_{j=S_{e_{i-1}}}^{S_{s_i}} \left(V_j^{L_2}\right)^2}}$$

and

$$CCright(i,M) = \frac{\sum_{j=S_{e_i}+M}^{S_{s_{i+1}}} V_{j-M}^{L_1} V_j^{L_2}}{\sqrt{\sum_{j=S_{e_i}+M}^{S_{s_{i+1}}} \left(V_{j-M}^{L_1}\right)^2 \sum_{j=S_{e_i}}^{S_{s_{i+1}}} \left(V_j^{L_2}\right)^2}}$$

where $V_j^{L_k} = \mathcal{S}_{L_k}(j) - \bar{\mathcal{S}}_{L_k}(j)$

We define the maximum correlation coefficient by

$$MCC(i) = \max(\max_M(CCleft(i,M)) \max_M(CCright(i,M)))$$

The optimal shift (delay of h) for the i^{th} segment is the value M that corresponds to the $MCC(i)$ value. To compute the scale, we select the noise region (left or right) which corresponds to $MCC(i)$ value. Then, the scale factor is computed as the ratio of the energy between the noisy region L_1 and the noisy region L_2. Thus, we construct h and compute NSNR.

4.5.3.2 Least Mean Squares Filter (LMSF)

In this case, we relax the assumptions of h and we let h to be any filter. To estimate the filter, we use the normalized LMS algorithm as described in [42]. Note at this point that we include the left and right noise regions of $\mathcal{S}_{L_1}$ and $\mathcal{S}_{L_2}$ in the input and target signals to get better estimates of the filter h. We denote the extended signals as SN_{L_1} and SN_{L_2}. Since at each step of LMS we are minimizing the distance between SN_{L_1} and SN_{L_2}, the filter will be such that SN_{L_1} will track SN_{L_2}. Ideally, the error of the LMS will be $h * X_{L_2}$ and ,thus, the output will be $h * SN_{L_1}$. To define the iterative Normalized LMS, we need to define first the truncated versions of SN_{L_1} and SN_{L_2}. We define $SN_{TL_1}(n)$ the truncated signal starting at sample n. The signal is truncated to

have length equal to h and n is used to shift the truncated signal. The input and target signals to Normalized LMS are SN_{TL_1} and SN_{TL_2} respectively. Next, the iterative Normalized LMS to estimate h in the $m^{th}+1$ iteration is performed in the following manner:

$$h^{m+1} = h^m + \frac{\mu(SN_{TL_2} - h^m \cdot SN_{TL_1}) \cdot SN_{TL_1}}{||SN_{TL_1}||^2}$$

After finding h, we use h, $\mathcal{S}_{L_1}$ and $\mathcal{S}_{L_2}$ to compute NSNR.

4.6 Experimental setup

To identify, align and segment speech and noisy speech regions, we used the algorithm described in [12]. Furthermore, we have used the same parameters values optimized in [12], since we are working on the same data set.

After getting the segments, we extracted the various features described in section 4.5. For the computation of SC, we computed 12 MFCCs (excluding the DC coefficient). For computing the filter h using the SSF method, we have searched M in the range -800:800 and, thus, searching correlations of 1601 values which means we are searching for the best shift within 100ms in a 16kHz audio signal. This is a reasonable assumption given the grounding of the audio channel to the video stream; additionally, this helps in constraining the computational cost.

Moreover, using LMSF to estimate h, we had to optimize the learning rate, μ, and filter size, $|h|$. On a development set, by using grid search we picked the parameters that maximize the average K-Nearest Neighbor K-NN performance for $K = 1 - 20$. We computed the performance for μ = [0.1 0.01 0.001 0.0001] and filter size $|h|$ = [30 80 250 800]. We found that the average K-NN performance was maximized for a filter size of 80 and learning rate $\mu = 0.001$. To get better estimates of the h filters for each bilingual segment, we run two iterations over the same segment. This helps the

Normalized LMS algorithm to converge if it did not converge during the first iteration. Of course, more iterations one runs, the better the estimate and convergence of the filter; however, extra iterations over the same segments increase the computational cost significantly.

In all experiments that K-NN is involved, we used Mahalanobis [32] distance as a distance function. Also, in order to strengthen the results of our work, we run a 5-fold cross-validation in all experiments. The split of train/test is 60%/40% and the results reported are the average of the folds.

4.7 Experiments, results and discussion

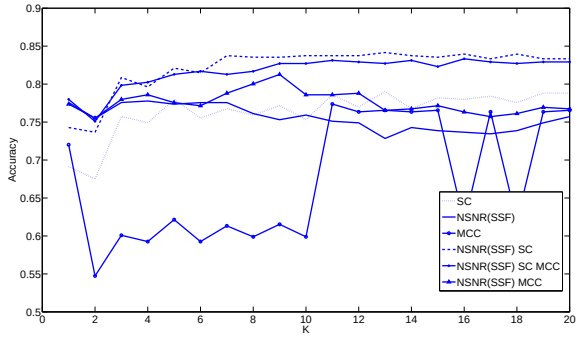

Figure 4.8: This figure shows the K-NN classifier versus the accuracies by varying K and combining various features.

Fig. 4.8 shows the performance of the features considered namely the SC feature, NSNR, and the Maximum Cross Correlation (MCC) value. In addition, we used various combinations of features to test, understand and verify which features contain complementary information. Using the features isolated, NSNR and SC have similar (75%-80%) performance across different values of K. However, it is interesting to observe that by combining the NSNR and SC features, we gain in accuracy. This suggests that these

two features contain complementary information. While the latter contains information about spectral similarity and totally ignores the short-term phase differences and amplitude related information, the former only takes into account phase shifts and scalings into computing the similarity of the two signals. Including all the features in the classifier, the selection accuracy is 83%.

Since the NSNR feature depends on the accurate estimation of h, it is interesting to compare the performance of different approaches in estimating h and, in addition, we consider the performance of different combinations of features using SC, NSNR, NSNR with least means squares filter estimation (LMSF) and (LMSF)-Extended in which we are using more iterations to estimate the filter h. For the (LMSF)-Extended, we are using 10 iterations over each segment for estimating each time the filter h. As shown in Fig. 4.9, the filter estimated with (LMSF)-Extended gives better convergence characteristics with the highest performance among all features. It is interesting to see that by combining NSNR(LMSF)-Extended with NSNR(SSF), SC and MCC, we get the best performance (84%-87%). This fact suggests that those features complement some missing information from NSNR(LMSF)-Extended. Also, due to the computational costs to estimate higher order filters for the least means square filter (i.e, the size of filter $|h|$), we did not experiment with filters higher than 50ms. This might be one reason that longer term information is not captured by NSNR(LMSF) features. The results also suggest that some of that information is captured by the the rest of the features namely NSNR(SSF), SC and MCC.

Overall, our method depends on a set of training data so that the algorithm learns from human annotations. The results indicate that the features provide discrimination upto 87% using the KNN [93] classifier for the data set considered. In addition, the NSNR features which are motivated by Signal-to-Noise ratio ideas have been the best performing.

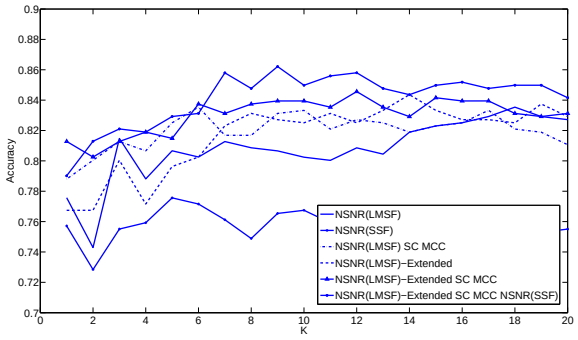

Figure 4.9: This figure shows the K-NN classifier versus the accuracies by varying K and combining various features. In particular, the main focus is to compare different approaches in estimating the filter h which relates the source with the target noise.

4.8 Conclusions

In this work, we focus on identifying clean bilingual speech signals by exploiting the relation between the background noise in two audio streams. We proposed various features to capture this information. The first feature captures the spectral correlation (SC) of the bilingual audio streams and aims to measure their relationship by spectral similarity. The second feature, called the Noise to Speech and Noise Ratio (NSNR) aims to model the relation using a signal plus noise model of two audio streams. NSNR requires an estimation of a filter h and we have proposed two methods to estimate h which vary in speed and performance. Our best performing approach delivers accuracies up to 87% in classifying clean and noisy speech.

Chapter 5:

Paralinguistic perceptual experiments [101]

5.1 A study on the effect of prosodic emphasis transfer on overall speech translation quality

Despite the increasing interest in Speech-to-speech (S2S) translation, research and development has focused almost exclusively on the lexical aspects of translation. The importance of transferring prosodic and other paralinguistic information through S2S devices and evaluating its impact on the translation quality are yet to be well established. The novelty in this work is a large scale human evaluation study to test the hypothesis that cross-lingual prosodic emphasis transfer is directly related to the perceived quality of speech translation. This hypothesis is validated at the 0.53-0.54 correlation level on the data sets considered with results significant at p-value=0.01. The second contribution of this work is an evaluation methodology based on crowd sourcing using English-Spanish language bilingual data from two distinct domains and evaluated

with over 200 bilingual speakers. We also present lessons learned on this type of S2S subjective experiments when using crowd sourcing.

5.1.1 Introduction

The goal of Speech-to-speech (S2S) translation is to allow spoken human interactions across different languages and support communication between people with limited or no knowledge of a certain spoken language. Such need is felt widely in today's increasingly multilingual multicultural world, such as in improving delivery of health-care to patients that do not share the same language [90]. Also due to the rapid expansion of tourism, Internet and smart-phones, S2S translation has attracted researchers attention during the last decade for building and using S2S translation applications that are portable and personal [49, 38, 76].

A typical S2S system has a pipelined architecture [84] in which an automatic speech recognizer (ASR) receives the speech signal and converts it into a sequence of words. Then, the sequence of words is translated with the statistical machine translator (SMT) into the target language. Finally, the words in the target language are synthesized using a Text-to-Speech (TTS) system. This pipelined architecture has its advantages in the sense that each component of this S2S pipeline can be isolated and researched independently. However, it has limitations when additional source language information needs to be exploited and for which the individual components are not designed to model.

5.1.1.1 Relation to prior work

Only limited work has been done in incorporating information that is not supported by the aforementioned components of the typical S2S pipelined architecture. In some works, additional information extracted from the speech signal has been used within

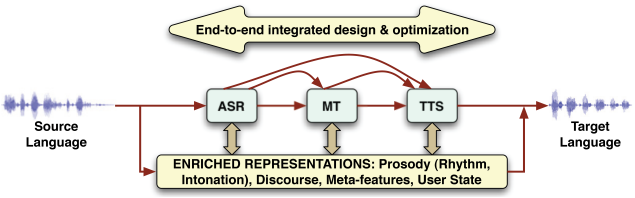

Figure 5.1: A system architecture that can exploit speech information beyond the pipelined architecture used in speech-to-speech systems.

the S2S components individually. For example, Parlikar *et al.* [9] have adapted the TTS output signal using phoneme mappings from the input language and have shown TTS improvement. Aguero *et al.* [71] used an unsupervised method to learn prosodic mappings trained on bilingual read sentences which are then used to enhance the TTS output and have shown benefits in terms of mean opinion score. Rangarajan *et al.* [78] have added dialog acts and prosodic features obtained from the source signal in the SMT component and have shown translation improvements in terms of BLEU score [70].

The importance of paralinguistic information in monolingual human communication has been widely documented [63, 27, 25]. The premise of our work is that such paralinguistic information is important in cross-lingual communication settings, and S2S systems should possess such capability. What is however unclear is what aspects of the multifaceted rich information in the source speech would be beneficial for inclusion in the cross-lingual transfer. Toward that end, in this paper we describe a method to systematically explore and evaluate the role and importance of specific aspects of paralinguistic information in S2S translation. This can be viewed as a design step even before an actual technology system is created.

We perform perceptual evaluation experiments using a crowd sourcing approach widely used in various experimental settings in the past [21, 89, 85, 103]. In particular, we describe a case study aimed at investigating whether the transfer of emphasis of a

word or a phrase in the source language to the target language is related to the perceived translation quality. To carry out our experiments, we use bilingual utterances obtained from dubbed movies and doctor-patient-interpreter interactions, with subjective experiments on Amazon Turk[1] to test our hypothesis. In this paper, we focus on the English-Spanish language pair. We find that the transfer of emphasis significantly correlates with the perceived translation quality for this language pair. In addition, we describe our experience in evaluating this type of S2S experiments using crowd sourcing.

This paper is structured as follows. In subsection 2, we describe the data collected and used in this work. In subsection 3, we elaborate on our hypothesis. Section 4 describes the survey used to conduct the perceptual experiments. Section 5 presents the experimental setup on Amazon Turk. In subsection 6, we discuss the results of this work and, finally, in subsection 7, we summarize the findings of this work and provide some future directions.

5.1.2 Data collection

In this subsection, we describe the data sets collected for testing our hypothesis through human evaluation experiments. We focus on two different data sets.

5.1.2.1 S2S data set

The first data set was collected by SAIL[2] as a part of a medical domain S2S translation project called Speech-links. This data set involves interactions between an English speaking doctor, a Spanish speaking patient and a bilingual interpreter that facilitates this interaction by translating from English to Spanish and vice versa. The doctors are students from USC's Keck School of Medicine and the patients are standardized patients. This method originally was proposed in [17]. The interpreters are professional

[1] http://www.mturk.com
[2] http://sail.usc.edu

English-Spanish interpreters trained to facilitate medical interactions in California's hospitals. The recordings took place in a typical room setting with little background noise coming from air-conditioning etc. Each session lasted up to half an hour. There are a total of six doctors, six interpreters and six patients and at the end of each doctor-patient-interpreter setting, and participants are permuted to ensure variety in the pairs involved in the interaction. The interactions are highly realistic and spontaneous and the interpreters were unconstrained in their task with minimal instructions that they should attempt to minimize overlap.

For the purpose of these experiments, we hired English-Spanish bilingual speakers to randomly pick 50 bilingual utterances from 10 different sessions resulting in a total set of 500 utterance pairs. We also asked the bilingual speakers to manually transcribe and match them in bilingual pairs, for example, two utterances are put together if they are bilingual translations of each other in the interaction. From now on, we will refer to this data set as S2SData set.

5.1.2.2 Movies data set

The second data set we experimented with comprises bilingual utterances that are from dubbed movies. This type of utterance pairs has more or less the same duration as the source utterance due to constraints of the visual channel. Often the translations are made in a way to lip-sync the words spoken in the source language. Dubbed movies are processed off-line and dubbed by professional interpreters, with the possibility of being recorded multiple times and, also, if possible, lip synced to match the video both in timings and visuals. In this sense, dubbed movies data differ from the S2SData set.

To obtain a set of high-quality bilingual utterances, we segmented the data using the approach described in [12]. Then, we processed the bilingual utterances and selected the clean bilingual pairs that do not contain background noise. From 15 dubbed

movies, we randomly selected 781 clean bilingual utterances, ensuring that the pairs were conceptually translations of each other, and transcribed them manually in both languages. From now on, we will refer to this data set as Movies data set.

5.1.3 Hypothesis: Transfer of Prosodic Emphasis

Our goal is to examine the hypothesis whether translation quality is affected by the quality of transfer of paralinguistic cues. In particular, we focus on the transfer of emphasis. In signal processing terms, emphasis/stress is defined as the perceived loudness of a word/phrase. Intuitively, if we want to emphasize a word/phrase, or a concept, we stress specific words/phrases of the utterance. By stressing the word/phrase, we may change the meaning conveyed by the utterance and, thus, such cues have to be taken into consideration in the translation. For example, an emphasized word might be important in the context of the dialog and the annotator might need to pay special attention to that word/phrase. Our main premise is that this paralinguistic cue is important both in terms of production (interpreters transfer this information) and in terms of perception (annotators perceive this information). We perform perceptual evaluation experiments to test this hypothesis for the English-Spanish language pair.

5.1.4 Perceptual evaluation experiments

To perform the perceptual experiments and test our hypothesis, we used the data described in subsection 5.1.2 and created the survey shown in Fig. 5.2. At each instance, we provided one bilingual Spanish-English utterance pair to the annotators and asked them to rate the quality of the translation (on a scale 1-5, question 1) and how well the emphasis of the English audio is preserved in the Spanish audio (on a scale 1-5, question 3). In addition, the annotators were asked to give their confidence in rating

the emphasis preservation (on a scale 1-5, question 4) and, whether, they perceive any words/phrases that are emphasized in the English audio (yes/no answer, question 2).

To examine our hypothesis, we tested the relation between the results on the quality of translation (question 1) with the ratings of emphasis preservation (question 3). Confidence ratings (question 4) were used to examine the hypothesis above for confident annotations. In addition, for quality testing purposes, we asked the annotators to transcribe each utterance (questions 5-6), thus, ensuring they paid attention to each audio signal. The survey given to the annotators is shown in Fig. 5.2[3].

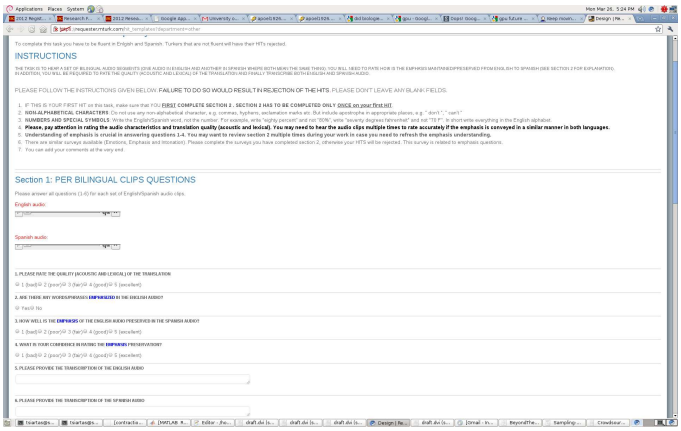

Figure 5.2: The survey used to validate the hypothesis claimed in the paper.

5.1.5 Experimental setup

For the perceptual experiments, we employed crowd sourcing through Amazon Turk. Using the survey described in subsection 5.1.4, we requested that each annotator participating in the survey to be an English-Spanish bilingual speaker. Before filling the survey, each annotator was mandated to go through training. Annotators were presented with samples of speech containing emphasized words/phrases and samples with no emphasized words/phrases so that we were sure it was clear to the annotators what

[3]Note that a one-time demographic survey and training session was given to each annotator

is the definition of emphasis. In addition, they were presented examples in which emphasis was transfered and other cases that emphasis was not transfered. Their attention to training was ensured through monitoring of the accuracy in transcription of each utterance that they had to transcribe. At this point, we accepted annotators that passed the training subsection without transcription errors.

Annotators who cleared the training phase had to answer the four questions explained in subsection 5.1.4 and to transcribe the utterances in both languages (Fig. 5.2, questions 5-6). To ensure the quality of the tagged data, we rejected annotators having Word Error Rate (WER) [61] greater than 25.0%.

Finally, we asked for 8 surveys filled for each individual utterance pair. In total, 202 different annotators participated in the surveys. 32.6% and 58.7% of the annotators reported English and Spanish language as the native language respectively. The rest reported other languages. We collected 5977 samples from the movies data and 3895 samples from the S2SData. If we define emphasis transfer as giving an emphasis transfer rating above 3, then 78.4% of the S2SData samples and 84.7% of the movies data have been rated with transfered emphasis. After the results had been collected, we computed the correlation and mutual information [24] between the emphasis transfer rating and the quality of the translation.

5.1.6 Results and discussion

5.1.6.1 Perceptual prosodic emphasis experiments

Fig. 5.3 shows the normalized counts of translation quality given the rating that emphasis was transfered. By normalized counts, we mean the histogram of the translation

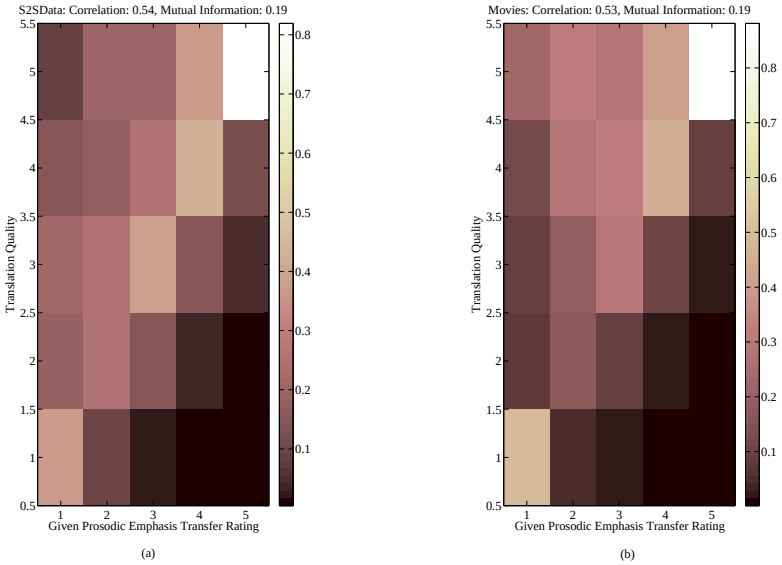

Figure 5.3: shows the normalized counts (normalized histogram) of translation quality given the rating that emphasis was transfered. Thus, each column sums up to 1 and represents the distribution of the translation quality for each emphasis transfer rating for both the S2Sdata and Movies data set.

quality ratings divided by the number of samples. Hence, each column in Fig. 5.3 represents the normalized counts for each emphasis rating. We plot the distribution per column to remove any bias coming from unequal priors of each rating value as reported in subsection 5. In the title of each plot, we provide the correlation and mutual information of emphasis transfer across the various levels of the translation quality variable. The lighter color indicates a high normalized count. Both data sets, show very similar trends. However, the movies data set indicates that some times emphasis transfer is rated as "bad" (i.e. rating 1) but still we get good translation quality rating. That might happen because, in some cases, although conceptually identical, may be paraphrased significantly to make emphasis transfer comparisons difficult (note the temporal synchronicity and potential lip-syncing constraints placed on actors).

	S2SData	Movies
Confidence≥4	0.52	0.49
Confidence≥4, No Emph. present	0.46	0.48
Confidence≥4, Emph. present	0.54	0.50

Table 5.1: Correlation coefficient when the results are conditioned on the confidence of the annotators and on the cases whether there exists emphasis in the English utterance.

Correlation gives a comparison tool to judge whether there is a linear or inversely linear relation between the emphasis transfer and the quality of the speech translation. The lighter color on the diagonals in Fig. 5.3 indicates that the hypothesis that more faithful transfer of prosodic emphasis is correlated with perceived overall translation quality; this is validated at the 0.54 correlation level between the emphasis transfer and the quality of the speech translation for the S2SData set and 0.53 for the Movies data set. All results are significant against the no correlation hypothesis using a t-test at p-value=0.01. For any non-linear relations, mutual information is used which is a measure of the predictive power between the two variables of interest and for both data sets the mutual information between the emphasis transfer and the quality of the speech translation is 0.19. Table 5.1 presents the correlation given that the confidence of the annotators is greater than 3. Similarly, we present the correlation given that the confidence of the annotators is greater than 3 and they perceived an emphasized word/phrase in the English side or they did not perceive an emphasized word/phrase. Results show that the hypothesis is validated at the 0.46-0.54 correlation level even when annotators report confidence greater than 3 with or without the presence of emphasized words for both S2Sdata and Movies data sets. All results are significant at p-value=0.01.

As reported in subsection 5.1.5, there is a bias towards samples that have been rated as prosodic emphasis transfered. To eliminate any effect of this bias, we randomly picked 500 samples from each class (One class contains the points where emphasis

transfer rating is greater than 3 and the rest points are in the other class) for each data set and computed the correlation coefficient. This experiment was repeated 1000 times with replacement. The average correlation of this experiment is 0.54 for S2SData and 0.51 for movies.

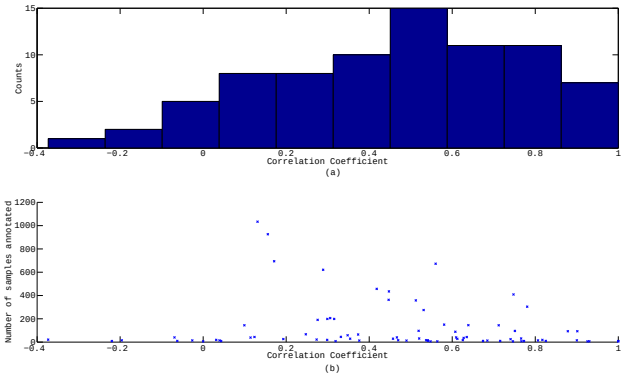

Figure 5.4: Fig. 5.4(a) shows the histogram of correlation between the quality of the translation and prosodic emphasis transfer. Fig. 5.4(b) shows the scatter plot of the number of samples completed by an annotator vs the correlation between the quality of the translation and prosodic emphasis. In both cases, we included annotators with more than 5 samples.

It is also interesting to examine the correlation values across annotators. Fig. 5.4(a) indicates that the main mass of the correlation between the quality of the translation and prosodic emphasis transfer is around 0.5 as expected from the overall correlation figures. In this histogram, the two data sets are reported together. The median point on the histogram is 0.52 which is close to the overall correlation coefficient reported. Fig. 5.4(b) shows a scatter plot of the number of samples annotated by each annotator and the corresponding correlation coefficient. Annotators with negative correlation have annotated very few samples and their effect was minimal on the overall correlation score. Also, a few people that annotated a lot of samples (e.g. above 700 samples) gave average correlation of 0.14 which is well below the overall median. However, the average

completion time per sample of annotators having above 700 samples is 58.5 seconds much lower than the global average of 98.6 seconds which questions those annotators quality.

Overall, we conclude that there is approximately 0.5 correlation between the emphasis transfer and the quality of the speech translation. However, to make a stronger statement with higher correlation we might need to include other prosodic variables, for example, intonation, emotional state, etc.

5.1.6.2 Lessons learned

From our experience with the S2S subjective experiments on Amazon Turk, we learned that it is important to have a training part, mandate annotators to take the training part and have a way to validate that they went through this training procedure. Initially, we did this experiment on a small scale with written instructions but without the training part and many annotators were asking questions about emphasis and what we expect from them. After manually creating clear examples on what we mean by prosodic emphasis the questions on this topic were minimal.

Apart from a well prepared training procedure and explanation, it is important to evaluate the annotators understanding of both languages. Initially, we had only the questions $1-4$ in the survey (Fig. 5.2) and soon realized that we were getting bad annotations (completed extremely fast) from people that we couldn't say if they are fluent in both languages or not. So we added the questions $5-6$ and mandated the annotators to transcribe all utterances in both languages. This helped us to ensure that annotators were actually listening to the samples they were rating and also we filtered a lot of annotators that were not fluent in both languages.

Finally, Amazon Turk provides no procedure to limit the number of samples annotated per person (only ensures that annotators are not presented with the same sample more than once). This created imbalances in number of the samples annotated per

person. To limit such imbalances one has to request annotators to stop after a certain upper bound number of annotations and if they do not comply, then exclude them from the task.

5.1.7 Conclusions and Future directions

In this work, we have presented a perceptual study to establish the hypothesis that there is a relation between the emphasis transfer and the quality of speech translation. The hypothesis is validated at 0.53-0.54 correlation level on the two data sets used. The results are significant at p-value=0.01. We also discussed the lessons we learned in rating perceptually the S2S translation quality in these subjective experiments using Amazon Turk.

Some future directions we want to investigate include expanding this work in carrying out the experiments in other language pairs. In addition, we want to expand the study to the relation of speech translation quality and other paralinguistic cues transfer, for instance, intonation and emotions.

Chapter 6:

Toward transfer of acoustic cues of emphasis across languages [102]

6.1 Toward transfer of acoustic cues of emphasis across languages

Speech-to-speech (S2S) translation has been of increased interest in the last few years with the research focused mainly on lexical aspects. It has however been widely acknowledged that incorporating other rich information such as expressive prosody contained in speech can enhance the cross-lingual communication experience. Motivated by recent empirical findings showing a positive relation between the transfer of emphasis and the quality of the audio translation, we propose a computational method to derive a set of acoustic cues that can be used in transferring emphasis for the English-Spanish language pair. In particular, we present an iterative algorithm that aims to discover the set of acoustic cue pairs in the two languages that maximize the accurate transfer of emphasis.

We find that the relevant acoustic cues can be constructed from a diverse set of features including word/phrase level statistics of spectral, intensity and prosodic cues and can model the acoustic information related to emphasized and neutral words/phrases for the English-Spanish language pair. These features can in turn enable data-driven transformations from source to target language that preserve such rich prosodic information. We demonstrate the efficacy of this approach through experiments on a specially constructed corpus of 1800 English-Spanish words/phrases.

6.1.1 Introduction

Speech-to-speech (S2S) translation's ultimate goal is to allow spoken human communication across different languages, dialects and cultures. S2S is becoming more desirable due to increasingly multicultural societies, people's increased travel, and due to widely available Internet-connected devices such as smart-phones. The need is also evident in improving health-care delivery among patients and doctors that do not speak the same language [90]. This need has attracted research and industry towards the creation of a robust and accurate S2S translation system.

A variety of S2S systems have been proposed in the literature [49, 38, 76]. A typical speech-to-speech (S2S) system is composed of an automatic speech recognizer (ASR) which converts the input into words, the words are translated using a statistical machine translator (SMT) and, finally, a Text-To-Speech (TTS) system is used to compose the target signal. In such pipelined S2S approach, one can isolate and work on subsystems independently. However, S2S translation is beyond this pipelined S2S approach. Recent work [101] has shown that additional paralinguistic cues such as emphasis can be also useful for S2S translation.

There is limited systems-side work in bringing paralinguistic aspects into S2S translation. However, there is early research into exploiting paralinguistic cues in the S2S

framework. Parlikar *et al.* [9] have used phoneme mappings as acoustic units to adapt the TTS output signal from the input language and shown benefits on the TTS side. On the feature side, power and duration have been used in [96] to translate emphasized digits and wherein the prediction of emphasis and root mean squared error rate (RMSE) have been used as an evaluation metric. Aguero *et al.* [71] have used an unsupervised approach in learning prosodic mappings and showed TTS output benefits in terms of mean opinion score. Rangarajan *et al.* [78] used dialog acts and prosodic cues obtained from the input speech signal within the SMT component and have shown translation benefits in terms of BLEU score [70].

While such approaches can offer useful information to various aspects of S2S components, a computational approach to learn paralinguistic representations can be very important for all components and S2S translation in the same way phonemes and words are useful for ASR. In contrast to existing work that focuses on the entire S2S system to show improvement in terms of different aspects of S2S translation, in this paper, we focus on deriving acoustic representations that maximize the direct information transfer across languages. We present a data-driven supervised approach that learns acoustic mappings by discretizing the acoustic space (modeled by diverse speech features such as MFCCs, pitch etc.) with the K-Means algorithm. The code mappings are evaluated using the mutual information between the bilingual discrete representations and the presence of paralinguisticly salient. In addition, the bilingual acoustic representations are evaluated by conditional entropy to measure the uncertainty of the mappings.

Specifically, in this paper, we show the efficacy of the approach by creating a representation for prosodic information transfered and focus on deriving the most informative acoustic representations. The representation is created from acoustic feature vectors discretized and evaluated using mutual information shared between the representation and

the emphasis transfer. The representation is learned from a quadruplets of parallel utterances spoken in neutral (flat tone) English, neutral Spanish, and English, Spanish with appropriate emphasis. In addition, we further attempt to jointly maximize the information transfered and the predictability of the encoding using conditional entropy as a measure.

This paper is structured as follows. In subsection 2, we explain how the 4-way bilingual data have been collected. In subsection 3, we describe the acoustic measures used to create the acoustic representation. In subsection 4, we give a brief description of the word/phrase level features used to model the acoustic space. Section 5 describes the approach used to map the acoustic space to the acoustic representation. In subsection 6, we describe the experimental setup and subsection 7 discusses the results of this work. Finally, in subsection 8, we summarize the findings of this work and provide some future directions.

6.1.2 Data-Driven Learning

To collect data suitable for directly learning emphasis transfer representations for the English-Spanish bilingual pair, we recruited two bilingual actors, one male and one female. We obtained a random utterance set from the IEMOCAP [20] database and translated all English utterances to Spanish.

The utterances were tagged with words to be emphasized. The corresponding word/phrase on the translated Spanish side has been marked as well. The actors spoke the utterance in both languages with emphasis and neutral resulting into a quadruplet. We recorded 450 such quadruplets resulting in 1800 utterances. Next, we extracted the words/phrases that are emphasized with their neutral counterparts in both languages resulting into 1800 words/phrases. The set has been split into half for training and half testing.

6.1.3 Acoustic Representation

In this subsection, we propose a representation for the acoustic cues transfered in cross-lingual spoken translation. To create this representation, we propose a mapping from the continuous acoustic space of speech to a discrete set of acoustic units.

Hence, say we have two words/phrases spoken in two languages L_1 and L_2. Let X_{L_1} and X_{L_2} be signal representations of the words/phrases, we define two mappings independently for the two languages to yield the corresponding discretized (quantized) vectors as follows:

$$X_{L_1} \rightarrow A_{L_1}$$

and

$$X_{L_2} \rightarrow A_{L_2}$$

The signal representation X_{L_i} is composed of a set of features, for example, transformations of MFCC, pitch and other spectral and prosodic features. The mapping defines a discretization of such features which denoted as A_{L_i}. To construct such a mapping, we use K-means clustering [32] to map the continuous space of acoustic cues to a finite discrete set of acoustic units.

6.1.3.1 Transfer of acoustic cues

With the aforementioned representation, we need a way to measure how well cues are transfered by the particular representation. Each feature vector and mapping to acoustic units can create a representation in which some mappings can model the "language" of acoustic cues transfered. Thus, we propose to use an information theoretic approach to evaluate each representation created by different feature vectors and different mappings to acoustic units. Hence, given a perceptual acoustic transfer Y, for example

emphasis transfer, we need to find a representation $A = (A_{L_1}, A_{L_2})$ such that their mutual information is maximized:

$$I(A, Y) = \sum_{a \in \mathcal{A}, y \in \mathcal{Y}} P(a, y) log \frac{P(a, y)}{P(a)P(y)}$$

where P defines the probability measure.

6.1.3.2 Conditional entropy for minimizing code uncertainty

While a specific representation can model the information shared between a language pair for analysis purposes of the acoustic cues transfered, it might have high uncertainty in the translation process. Thus, it is useful to have a measure to model the coding mapping uncertainty. For this reason, we propose to use a soft metric to evaluate how well the acoustic translation representation can be predicted using conditional entropy which can be written as:

$$H(A_{L_1}|A_{L_2}) = \sum_{a_{L_1} \in \mathcal{A}_{L_1}, a_{L_2} \in \mathcal{A}_{L_2}} P(a_{L_1}, a_{L_2}) log \frac{P(a_{L_2})}{P(a_{L_1}, a_{L_2})}$$

Using the conditional entropy metric, we can evaluate the ambiguity of the coding scheme.

6.1.4 Acoustic features

We considered a variety of acoustic feature vectors (X_{L_1},X_{L_2}) to represent different aspects of the speech spectrum and prosodic cues. All features are defined at the word/phrase level.

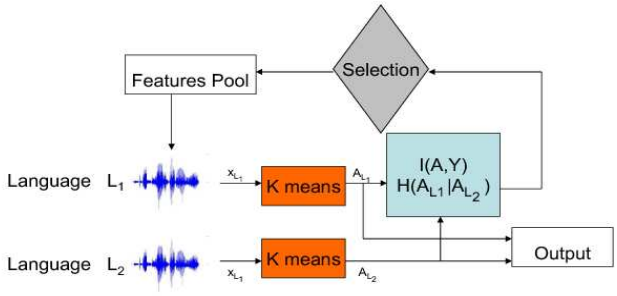

Figure 6.1: The iterative approach used to find the best acoustic representation for the acoustic cues transfered.

6.1.4.1 Mean power

The first feature we used in our representation is mean power to model the transfer of emphasis. Since power has been used widely in a variety of settings for modeling emphasis, we use it to produce a baseline representation.

6.1.4.2 Additional features

In addition to mean power, we have used word/phrase duration and various word/phrase level statistics of features which include MFCC, voicing pitch, etc. Statistics used include quantiles, mean, max, min, etc. In total we have extracted 6126 word/phrase level features for each word/phrase. The feature set has been extracted using OpenSMILE [33] as used in [81].

6.1.5 Acoustic unit estimation approach

In this subsection, we describe the approach used to create the acoustic units. A basic layout of the approach is shown in Fig. 6.1. The approach is iterative and the dimensionality of the feature space is increased progressively and in a greedy manner.

The algorithm is initialized with an empty feature vector. In step one, we add a feature to the feature vector $\mathcal{F}$. In the second step, the K-Means [32] algorithm is run independently for languages $L1$ and $L2$ and the encoding (A_{L_1} and A_{L_2}) is created for each language. Thirdly, the coding is evaluated using the MI and conditional entropy metric as defined in Sec. 6.1.3. If the metric considered improves, the feature replaces the last feature added in $\mathcal{F}$. If all features described in 6.1.4 are exhausted, we increase the dimensionality of $\mathcal{F}$ and go to step one.

6.1.6 Experimental setup

For experimenting with the emphasis transfer problem, we run the algorithm described in Sec. 6.1.5 in different setups. First, we run the algorithm by maximizing the mutual information. Then, we run the algorithm by maximizing the mutual information between the cross-lingual acoustic representations and the emphasis transfer at each step and at the same time minimizing the entropy so that we include as much information about the paralinguistic cue transfer but also find a representation that will minimize the coding prediction error. In addition, as described in Sec. 6.1.4, we used the mean power and duration of the signal to create a representation and form a baseline to evaluate the efficacy of our approach. To perform this optimization, we split the data set into two parts one for training and one for testing with half of the data in each set. The optimization has been run on the training set and we report the results on the testing set. Since the acoustic representations are created on the training set, we assign to the testing feature vector the closest code as defined by its cluster center.

Finally, we repeated the experiments for coding schemes with vocabulary size of 2, 4 and 8 codes in each language. For computational purposes, we run the experiments until at maximum 10 features are added or stopped if no more features can be added to improve the metrics considered.

6.1.7 Results and Discussion

In this subsection, we analyze the results of the computational approach to find appropriate representations for the English-Spanish pair. Fig. 6.2 shows the value of mutual information (MI) between the acoustic representations in English and Spanish and the transfer of emphasis. Results show that the amount of information transfered increases when increasing the number of tokens the acoustic representation is composed. Adding the duration to the power baseline (Power+Dur) can increase the information transfered in the English-Spanish bilingual pair.

Furthermore, when we maximize the mutual information ($I(A, Y)$) the approach can identify acoustic representations that yield up to 0.07 MI measure higher than power and duration together depending on the number of tokens considered which in-turn implies that more information is transfered. Using the Wilcox rank test and breaking the test set into 45 subsets, we find that the results are significant at p-values less than 10^{-8} for both the comparisons.

In applications such as speech translation, it is important not only to ensure that the representations ensure maximal information transfer, but yield as minimal ambiguity as possible to enable correct translation with low uncertainty. For this reason, we repeated our experiments by maximizing the mutual information ($I(A, Y)$) and minimizing the entropy ($H(A_{L_S}|A_{L_E})$ and vice versa for Spanish→English) at each step of the algorithm.

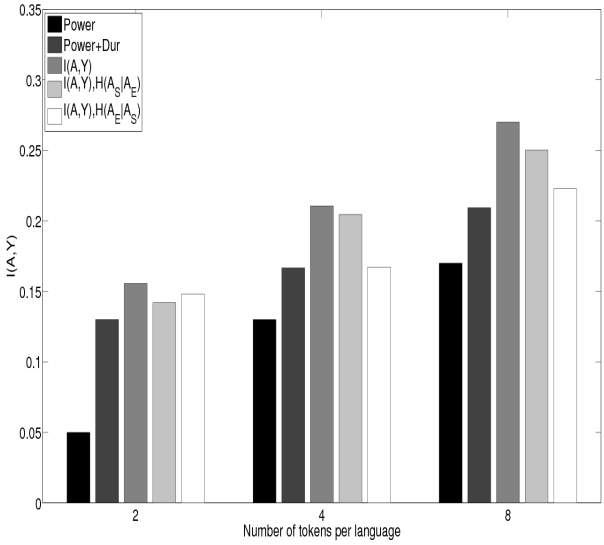

Figure 6.2: This figure shows the mutual information $I(A, Y)$ of the acoustic representations for emphasis transfer for different approaches and different number of tokens.

Results in Fig. 6.2 show that such joint optimization yields less information carried in the bilingual English-Spanish pair than optimizing only on mutual information ($I(A, Y)$) by up to 0.05 depending on the number of tokens considered but still more information than the baselines of up to 0.1 in terms of mutual information.

In addition, while adding duration to the baseline increased the mutual information, in three cases it increased the ambiguity of the coding scheme only when discretizing into two representations. Also, as shown in Fig. 6.3 the computational approach to create acoustic representations yielded codes with much lower ambiguity measured in terms of conditional entropy. In particular, this dual metric can lower the conditional entropy by up to 0.7 points depending on the number of tokens considered. The improvements in conditional entropy are consistent for both sides of the mapping of the acoustic information for all numbers of tokens considered and results are significant at p-values less than 10^{-8} for all cases.

While optimizing only on MI, the conditional entropy remains very close to the baseline but with much more cross-lingual information carried in the representation.

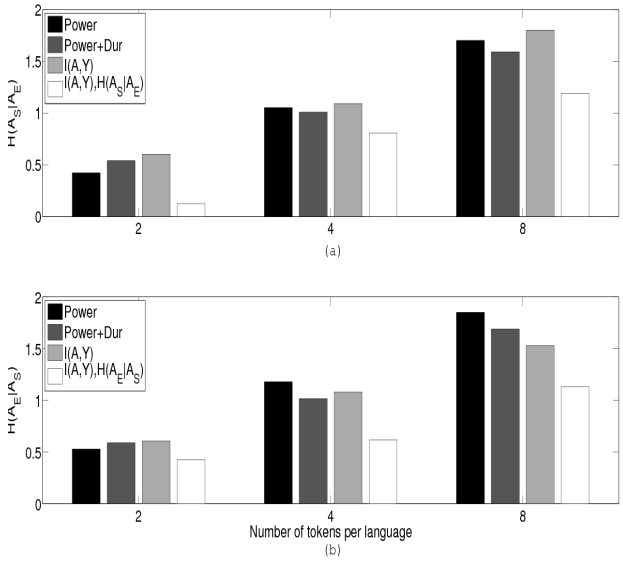

Figure 6.3: This figure shows the conditional entropy for the English to Spanish (a) and Spanish to English (b) translation of the acoustic representations with different approaches for different number of tokens.

6.1.8 Conclusion

In this work, we presented a computational approach to construct a cross-lingua representation for acoustic cues transfer and, in particular for the emphasis transfer. We have presented a mapping from the acoustic feature space to a discrete set of units using an iterative procedure in which at each step the mutual information is maximized. This method can potentially lead to an approach to learn cross-lingual information across speech-to-speech (S2S) components that can be used beyond the pipelined architecture of S2S by exploiting a diverse set of features.

Furthermore, for applications such as speech translation that require the resulting acoustic units to have low uncertainty for the prediction while simultaneously transferring as much information as possible, we added another condition to the algorithm to jointly maximize mutual information and minimize the conditional entropy of the prediction. Our results indicate that for applications in which the information transfered is important we can achieve MI up to three times higher than the baseline considered.

In addition, for applications requiring not only the maximum amount of information transfered but also low ambiguity of the coding scheme, the joint maximization of mutual information and conditional entropy yielded reductions in terms of conditional entropy of up to 3.5 times for the English Spanish bilingual translation.

For future work, we intend to collect and evaluate our approach on more speakers. In addition, we want to explore more features sets that can be used in the approach and also improve the approach with different optimization techniques to yield higher mutual information (MI) and lower conditional entropy as a measure of the coding scheme uncertainty. Also, additional metrics can be useful for extracting different cross-lingual information useful in different S2S components.

Chapter 7:

Conclusions

7.1 Summary and contributions

Speech to speech (S2S) is a system that facilitates translation from one language to another and includes a diverse set of modules. In this thesis, I focused on a subset of components used in a speech to speech (S2S) system. Such components include the front-end of voice activity detection module used to discriminate noisy and speech signals. In addition, techniques to align movie subtitles and bilingual movie audio used to train S2S systems. Furthermore, a perceptual analysis on the importance of paralinguistic cues in S2S translation was presented. Finally, I have shown a method to represent acoustic cues transfer in S2S translation.

Chapter 2 presented a method to identify speech in noisy environments based on the Long-term spectral variability (LTSV) feature. Furthermore, The theoretical and experimental results supporting the robustness of the method in noisy environments and its theoretical independence of Signal to Noise Ratio (SNR) in stationary environments are shown. It is also shown the per frame performance averaged over eleven noises and 5

SNRs to be 92.95% for eleven noisy types and five SNR levels outperforming the state-of-the-art system. In addition, a method for cross-talk voice activity detection (VAD) is proposed for multi-channel audio signals with an accuracy of 92% on the medical speech-to-speech (S2S) data set.

In chapter 3, I have shown an approach for aligning subtitles using information from a dictionary and exploiting the linearity of subtitle time-stamps. In particular, I have used 2000 utterances from the DARPA TRANSTAC data set, translated the English utterances into Spanish and French and evaluated the performance of the approach for the English-Spanish, English-French and vice versa language pairs. We found that the proposed method can facilitate a selection of high quality bilingual utterance pairs that showed performance improvements on average over all language pairs for different corpus sizes of 4.88 BLEU score points compared to past works.

In chapter 4, I have used the observation that parallel audio segments in movies share the same background audio/noise and I have proposed the Long-term spectral distance (LTSD) feature and a method to identify bilingual parallel speech segments from dubbed movies. I have shown that I can identify these segments with accuracy of 89% and identify the corresponding subtitles with 91% accuracy. Going one step further, I have used the least mean squares adaptive filter to estimate a SNR-based feature which was very effective in estimating whether the bilingual audio pair is noisy speech or clean speech. Combining the SNR-based feature with time-domain and spectral correlation features I achieved performance of 87% in detecting clean speech signals.

In chapter 5, I presented and analyzed the data I obtained from two data sets. The focus of this chapter has been on effect of paralinguistic cues in the quality of S2S translation. I have used amazon turk to check the premise whether emphasis, emotions and intonation transfer is correlated with the quality of translation. Based on the ratings of bilingual speakers and using two different data sets, I have found that the

above-mentioned paralinguistic cues correlate with the quality of the translation in the range of 0.41-0.6.

Chapter 6 presented a method for the representation of acoustic cues transfer with the application in emphasis transfer. The method maps features from the continuous acoustic feature space to a discrete set of units using an iterative greedy procedure with objective function the mutual information maximization. In addition, application such as speech translation require the resulting acoustic units to be easily translatable and predictable while maintaining as much information transfered as possible. To reach this goal, I have proposed to jointly maximize mutual information and minimize the conditional entropy of the prediction. The results indicate that predictability of the acoustic cues is important for applications such as speech translation. In the data set tested, I can achieve MI up to three times higher than the baseline considered while maintaining much lower uncertainty.

Appendix

.1 Proof of $\log R = \xi_k^N(m) \geq \xi_k^{S+N}(m) \geq \xi_k^S(m) \geq 0$

From eqn. (2.4), we rewrite the following:

$$\left.\begin{aligned}
&\xi_k^N(m) = -\sum_{n=m-R+1}^{m} \frac{\sigma_k}{R\sigma_k} \log\left(\frac{\sigma_k}{R\sigma_k}\right) = \log R \quad \text{[by setting } SS = 0\text{]} \\
&\xi_k^{S+N}(m) = -\sum_{n=m-R+1}^{m} \frac{SS+\sigma_k}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k} \log\left(\frac{SS+\sigma_k}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k}\right) \\
&\xi_k^S(m) = -\sum_{n=m-R+1}^{m} \frac{SS}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)} \log\left(\frac{SS}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)}\right) \quad \text{[by setting } \sigma_k = 0\text{]}
\end{aligned}\right\}$$

We know that entropy is bounded by two values [4]

$$0 \leq \xi_k^{S+N}(m) \leq \log R = \xi_k^N(m) \quad \text{and} \quad 0 \leq \xi_k^S(m) \leq \log R = \xi_k^N(m) \tag{1}$$

We need to show

$$\xi_k^S(m) \leq \xi_k^{S+N}(m) \tag{2}$$

Consider eqn. (2.2). Let us denote $\frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m} S_x(l,\omega_k)} = p_n,\ n = m-R+1, ..., m.$ Then

$$\begin{aligned}\xi_k^x(m) &= -\sum_{n=m-R+1}^{m} \frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m} S_x(l,\omega_k)} \log\left(\frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m} S_x(l,\omega_k)}\right) = -\sum_{n=m-R+1}^{m} p_n \log p_n \\ &= H(p_{m-R+1},...,p_m) \\ &= H\left(\frac{S_x(m-R+1,\omega_k)}{\sum_{l=m-R+1}^{m} S_x(l,\omega_k)},\ ...,\ \frac{S_x(m,\omega_k)}{\sum_{l=m-R+1}^{m} S_x(l,\omega_k)}\right) \qquad (3)\end{aligned}$$

H is a function with R-dimensional argument $\{p_n\}_{n=m-R+1}^{m}$, where $p_n = \frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m} S_x(l,\omega_k)}$. We know that H is a concave function of $\{p_n\}_{n=m-R+1}^{m}$ [4] and it takes maximum value at $p_{m-R+1} = ... = p_m = \frac{1}{R}$. Let us denote this point in R-dimensional space by $\underline{\eta}_N = \left[\frac{1}{R}\ ...\ \frac{1}{R}\right]^T$, where $[.]^T$ is vector transpose operation. Thus, $\xi_k^N(m) = H(\underline{\eta}_N) = \log R$. Similarly, $\xi_k^S(m) = H(\underline{\eta}_S)$ and $\xi_k^{S+N} = H(\underline{\eta}_{S+N})$, where $\underline{\eta}_S = \left[\frac{S_S(m-R+1,\omega_k)}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)}\ \ \cdots\ \ \frac{S_S(m,\omega_k)}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)}\right]^T$ and

$\underline{\eta}_{S+N} = \left[\frac{S_S(m-R+1,\omega_k)+\sigma_k}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k}\ \ \cdots\ \ \frac{S_S(m,\omega_k)+\sigma_k}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k}\right]^T$. From eqn. (2), we need to show $H(\underline{\eta}_S) \leq H(\underline{\eta}_{S+N})$.

Proof:

$$\frac{SS+\sigma_k}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k} = \lambda\left(\frac{SS}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)}\right) + (1-\lambda)\left(\frac{1}{R}\right),\ \forall n$$

where $\lambda = \frac{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)}{\sum_{l=m-R+1}^{m} S_S(l,\omega_k)+R\sigma_k}$. Thus $\underline{\eta}_{S+N}$ can be written as a convex combination of $\underline{\eta}_N$ and $\underline{\eta}_S$, i.e., $\underline{\eta}_{S+N} = \lambda\underline{\eta}_S + (1-\lambda)\underline{\eta}_N$. Now,

$$
\begin{aligned}
H(\underline{\eta}_{S+N}) &= H(\lambda\underline{\eta}_S + (1-\lambda)\underline{\eta}_N) \\
&\geq \lambda H(\underline{\eta}_S) + (1-\lambda)H(\underline{\eta}_N),\ (H \text{ is a concave function}) \\
&\geq \lambda H(\underline{\eta}_S) + (1-\lambda)H(\underline{\eta}_S),\ (\text{From eqn. (1), } H(\underline{\eta}_S) \leq H(\underline{\eta}_N)) \\
&= H(\underline{\eta}_S) \\
\Longrightarrow \xi_k^{S+N}(m) &\geq \xi_k^S(m),\ (\text{As } \xi_k^S(m) = H(\underline{\eta}_S) \text{ and } \xi_k^{S+N} = H(\underline{\eta}_{S+N})) \qquad (4)
\end{aligned}
$$

Thus eqn. (2) is proved. Hence, combining eqn. (1) and (2),

$$\log R = \xi_k^N(m) \geq \xi_k^{S+N}(m) \geq \xi_k^S(m) \geq 0 \qquad \text{(proved)}$$

.2 A better estimate of $\mathcal{L}_N(m)$ and $\mathcal{L}_{S+N}(m)$, [$N(n)$ is a stationary noise]

When $x(n) = N(n)$, $S_x(n, \omega_k) = S_N(n, \omega_k) = \sigma_k$ and hence $\mathcal{L}_N(m) = 0$. However, σ_k is unknown. We need to estimate these from available noise samples. If we use the periodogram (eqn. (2.3)), the estimate of $S_N(n, \omega_k)$ is biased and has a variance γ_N^2 (say). On the other hand, if we use the Bartlett-Welch method of spectral estimate (eqn. (2.6)), the estimate of $S_N(n, \omega_k)$ is asymptotically unbiased and has a variance of $\frac{1}{M}\gamma_N^2$ [36].

The estimate of $\mathcal{L}_N(m)$ is obtained from eqn. (2.1) and (2.2) by replacing $S_N(n, \omega_k)$ in eqn. (2.2) with its estimate $S_N(\hat{n}, \omega_k)$. From eqn. (2.1) and (2.2), we see that $\mathcal{L}_N(m)$ is a continuous function of $\xi_k^N(m)$ and $\xi_k^N(m)$ is a continuous function of $\left\{S_N(\hat{n}, \omega_k)\right\}_{n=m-R+1}^{m}$. When the Bartlett-Welch method is used, $S_N(\hat{n}, \omega_k)$ converges in probability to σ_k as $M \longrightarrow \infty$ (assuming N_w is sufficiently large to satisfy asymptotic unbiased condition) [36]. And hence, $\mathcal{L}_N(m)$, being a continuous function of $\left\{S_N(\hat{n}, \omega_k)\right\}_{n=m-R+1}^{m}$, also converges in probability to 0 as $M \longrightarrow \infty$ [7]. Thus, for large M we get a better estimate of $\mathcal{L}_N(m)$ using the Bartlett-Welch method. If the periodogram method is used instead, we don't gain this asymptotic property.

A similar argument holds for the case when $x(n) = S(n) + N(n)$. The Bartlett-Welch method of spectral estimate $S_x(\hat{n}, \omega_k)$ always yields a better estimate of $\mathcal{L}_x(m)$ compared to that obtained by the periodogram method.

.3 Dynamic time warping algorithm

The optimization problem described in (3.1) can be solved efficiently using DTW. The goal of the DTW algorithm is to find the best mappings $\left\{m_{ij}^{\star}\right\}$ that minimize the global distance: $\sum_{i,j} m_{ij}\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$. We use the notion of the cumulative distance for the best mappings starting from $(S_1^{L_1}, S_1^{L_2})$ and ending at $(S_i^{L_1}, S_j^{L_2})$, denoted by $\mathcal{L}(S_i^{L_1}, S_j^{L_2})$ [11]. We define $\zeta(S_i^{L_1}, S_j^{L_2})$ as the mapping prior to last in the best path terminating at $(S_i^{L_1}, S_j^{L_2})$. Formally, the DTW algorithm is defined in Table 1.

(i)*Initialization:*

$$\mathcal{L}(S_1^{L_1}, S_1^{L_2}) = \mathcal{DM}\left(S_1^{L_1}, S_1^{L_2}\right)$$

$$\begin{aligned}
\mathcal{L}(S_1^{L_1}, S_j^{L_2}) &= \mathcal{L}(S_1^{L_1}, S_{j-1}^{L_2}) + \mathcal{DM}\left(S_1^{L_1}, S_j^{L_2}\right) \\
\zeta(S_1^{L_1}, S_j^{L_2}) &= \left(S_1^{L_1}, S_{j-1}^{L_2}\right), \quad j = 1, ..., N_2
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}(S_i^{L_i}, S_1^{L_2}) &= \mathcal{L}(S_{i-1}^{L_1}, S_1^{L_2}) + \mathcal{DM}\left(S_i^{L_1}, S_1^{L_2}\right) \\
\zeta(S_i^{L_1}, S_1^{L_2}) &= \left(S_{i-1}^{L_1}, S_1^{L_2}\right), \qquad i = 1, ..., N_1
\end{aligned}$$

(ii)*Iteration:*

$$\begin{aligned}
\mathcal{L}(S_i^{L_1}, S_j^{L_2}) &= \mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right) + \min\left\{\mathcal{L}(S_i^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_j^{L_2})\right\} \\
\zeta(S_i^{L_1}, S_j^{L_2}) &= \arg\min\left\{\mathcal{L}(S_i^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_j^{L_2})\right\}, \; i = 2, ..., N_1, \quad j = 2, ..., N_2 \quad (5)
\end{aligned}$$

(iii)*Backtracking:* Let there be K mappings. Then

$$m_{N_1,N_2} = m(S_{N_1}^{L_1}, S_{N_2}^{L_2}) = 1, \quad m_{1,1} = m(S_1^{L_1}, S_1^{L_2}) = 1$$

$$\xi_K = \zeta(S_{N_1}^{L_1}, S_{N_2}^{L_2})$$

$$\begin{cases} m(\xi_k) = 1 \\ \xi_{k-1} = \zeta(\xi_k), \quad k = K, ..., 2. \end{cases}$$

Table 1: Steps in subtitle alignment using DTW approach.

.4 Optimal α, ϵ-linear function parameters

The optimal minimum mean square sense α, ϵ-linear function, $f^*(x) = m^* + b^*$, of a set of points $P = \{\{x_{1i}, y_{1i}\}, \{x_{2i}, y_{2i}\} : 1 \leq i \leq M\}$ of order $|I| = N$, $I \subseteq \{i : 1 \leq i \leq M\}$ with $\frac{1}{\alpha} < \left|\frac{y_{2i} - y_{1i}}{x_{2i} - x_{1i}}\right| < \alpha \; \forall i \in I$ and $\alpha > 1$ is given by the parameters:

(i)

$$m^* = \frac{N(Y_1 + Y_2) - Z_1 - Z_2}{N \sum_{i \in I} (x_{1i} + x_{2i})^2 - \left(\sum_{i \in I} (x_{1i} + x_{2i})\right)^2}$$

where,

$$\begin{aligned} Y_1 &= \sum_{i \in I} y_{1i}(x_{1i} + x_{2i}) \\ Y_2 &= \sum_{i \in I} y_{2i}(x_{1i} + x_{2i}) \end{aligned}$$

and

$$\begin{aligned} Z_1 &= \sum_{i \in I} y_{1i} \sum_{i \in I} (x_{1i} + x_{2i}) \\ Z_2 &= \sum_{i \in I} y_{2i} \sum_{i \in I} (x_{1i} + x_{2i}) \end{aligned}$$

(ii)

$$b^* = \frac{\left(\sum_{i \in I} y_{1i} + \sum_{i \in I} y_{1i}\right) \sum_{i \in I} (x_{1i} + x_{2i})^2 - (Y_1 + Y_2)\left(\sum_{i \in I} x_{1i} + \sum_{i \in I} x_{2i}\right)}{2\left(N \sum_{i \in I} (x_{1i} + x_{2i})^2 - \left(\sum_{i \in I} (x_{1i} + x_{2i})\right)^2\right)}$$

where Y_1 and Y_2 are defined in (i)

.4.1 Proof of the optimal parameters using least squares

The goal is to find m^* and b^* such that:

$$\begin{aligned} MSE &= \sum_{i \in I} \left(\frac{y_{1i}}{2} + \frac{y_{2i}}{2} - f^* \left(\frac{x_{1i}}{2} + \frac{x_{2i}}{2} \right) \right)^2 && \because \quad MSE \text{ definition} \\ &= \tfrac{1}{4} \sum_{i \in I} \left(y_{1i} - m^* x_{1i} - b^* + y_{2i} - m^* x_{2i} - b^* \right)^2 && \because \quad \text{Linearity of } f \end{aligned}$$

(i) Taking the derivatives with respect to m^* and b^* and setting them to zero gives the equations needed to find m^* and b^* that minimize MSE. Thus,

$$\begin{aligned} \frac{d(MSE)}{dm^*} &= \frac{d}{dm^*} \left(\frac{1}{4} \sum_{i \in I} \left(y_{1i} - m^* x_{1i} - b^* + y_{2i} - m^* x_{2i} - b^* \right)^2 \right) \\ &= -\tfrac{1}{2} \sum_{i \in I} \left(x_{1i} + x_{2i} \right) \left(y_{1i} - m^* x_{1i} - b^* + y_{2i} - m^* x_{2i} - b^* \right) \\ &= \tfrac{1}{2} \left(Y_1 - \sum_{i \in I} m^* (x_{1i} + x_{2i})^2 + Y_2 - \sum_{i \in I} 2b^* (x_{1i} + x_{2i}) \right) \\ &= 0 \end{aligned}$$

where Y_1 and Y_2 are defined in theorem .4.

By rearranging terms, we get the following relation,

$$Y_1 + Y_2 = \sum_{i \in I} m^* (x_{1i} + x_{2i})^2 + \sum_{i \in I} 2b^* (x_{1i} + x_{2i}) \tag{6}$$

Next multiply both sides of (6) by N and we get the following equation of two unknowns, m^* and b^*,

$$NY_1 + NY_2 = N \sum_{i \in I} m^*(x_{1i} + x_{2i})^2 + N \sum_{i \in I} 2b^*(x_{1i} + x_{2i}) \tag{7}$$

The next equation with unknowns m and b is obtained by differentiating MSE w.r.t. b

$$\begin{aligned}
\frac{d(MSE)}{db^*} &= \frac{d}{db^*}\left(\frac{1}{4}\sum_{i \in I}(y_{1i} - m^*x_{1i} - b^* + y_{2i} - m^*x_{2i} - b^*)^2\right) \\
&= -\sum_{i \in I}(y_{1i} - m^*x_{1i} + y_{2i} - m^*x_{2i} - 2b^*) \\
&= -\sum_{i \in I} y_{1i} + m^* \sum_{i \in I} x_{1i} - \sum_{i \in I} y_{2i} + m^* \sum_{i \in I} x_{2i} + 2b^* \sum_{i \in I} 1 \\
&= -\sum_{i \in I} y_{1i} + m^* \sum_{i \in I} x_{1i} - \sum_{i \in I} y_{2i} + m^* \sum_{i \in I} x_{2i} + 2Nb^* \\
&= 0
\end{aligned}$$

By rearranging terms, we get the following relation

$$\sum_{i \in I} y_{1i} + \sum_{i \in I} y_{2i} = m^* \sum_{i \in I} x_{1i} + m^* \sum_{i \in I} x_{2i} + 2Nb^* \tag{8}$$

Next multiply both sides of (8) by $\sum_{i \in I}(x_{1i} + x_{2i})$ and we end up with another equation with two unknowns, m^* and b^*,

$$Z_1 + Z_2 = m^* \left(\sum_{i \in I}(x_{1i} + x_{2i})\right)^2 + 2Nb^* \sum_{i \in I}(x_{1i} + x_{2i}) \tag{9}$$

where Z_1 and Z_2 are defined in .4.

Now subtracting (9) from (7)

$$N(Y_1+Y_2)-Z_1-Z_2=m^*\left(N\sum_{i\in I}(x_{1i}+x_{2i})^2-\left(\sum_{i\in I}(x_{1i}+x_{2i})\right)^2\right)$$

Solving w.r.t. m^*

$$m^*=\frac{N(Y_1+Y_2)-Z_1-Z_2}{N\sum_{i\in I}(x_{1i}+x_{2i})^2-\left(\sum_{i\in I}(x_{1i}+x_{2i})\right)^2}$$

as required

(ii) We substitute m^* in (9), rearrange terms and we end up with

$$\begin{aligned} b^* &= \frac{\sum_{i\in I}y_{1i}+\sum_{i\in I}y_{2i}-m^*\sum_{i\in I}x_{1i}+\sum_{i\in I}x_{2i}}{2N} \\ &= \frac{\left(\sum_{i\in I}y_{1i}+\sum_{i\in I}y_{1i}\right)\sum_{i\in I}(x_{1i}+x_{2i})^2-(Y_1+Y_2)\left(\sum_{i\in I}x_{1i}+\sum_{i\in I}x_{2i}\right)}{2\left(N\sum_{i\in I}(x_{1i}+x_{2i})^2-\left(\sum_{i\in I}(x_{1i}+x_{2i})\right)^2\right)} \end{aligned}$$

.5 Additional experiments

Since the TRANSTAC bilingual corpus is not available in English-Spanish and English-French languages, we present here the comparison with Europarl version-5 (EUROP) [54] and the NEWS corpora for readers interested in the performance of subtitles for this domain. As shown in Fig. 1, SMT models trained on the subtitle corpus (TA-2) improve the performance by up to 14.34 BLEU score points on the TRANSTAC test set. Thus, on this data-set, it is beneficial to use subtitles for training the SMT models.

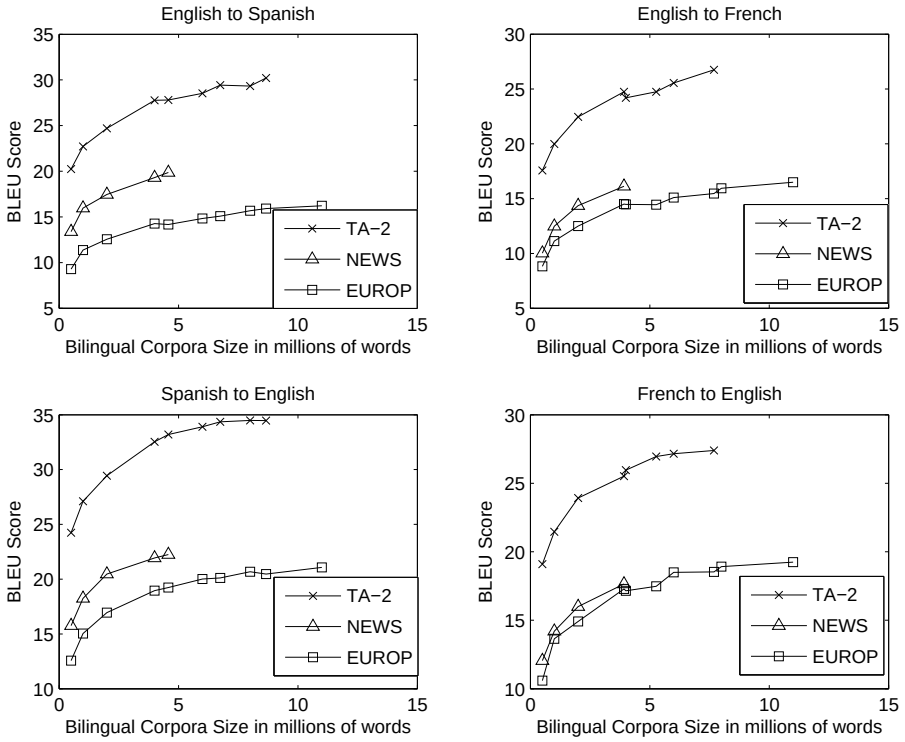

Figure 1: This figure compares the performance of the SMT models trained on time-alignment (TA-2), NEWS and EUROP corpora when the TRANSTAC development and test sets are considered. The experiments were repeated for various bilingual corpora sizes. The comparison is extended for the language pairs between English-Spanish, English-French, and vice versa.

Bibliography

[1] Digital cellular telecommunications system (phase 2+); adaptive multi rate (amr) speech; ansi-c code for amr speech codec. 1998.

[2] Digital cellular telecommunications system (phase 2+); voice activity detector (vad) for adaptive multi rate (amr) speech traffic channel; general description. 1999.

[3] Bies D. A. and Hansen C. H. *Engineering Noise Control: Theory and Practice*. Edition: 3, illustrated, Published by Taylor, 2003.

[4] Cover T. M. Thomas J. A. *Elements of Information Theory*. Wiley-Interscience, August 1991.

[5] Craciun A. and Gabrea M. Correlation coefficient-based voice activity detector algorithm. In *Canadian Conference on Electrical and Computer Engineering*, volume 1, pages 1789–1792, May 2004.

[6] Davis A., Nordholm S., and Togneri R. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. *IEEE Trans. on Audio, Speech and Language Proc.*, 14(2):412–424, March 2006.

[7] Gubner J. A. *Probability and Random Processes for Electrical and Computer Engineers*. Cambridge University Press, June 2006.

[8] Makur A. and Mitra S.K. Warped discrete-Fourier transform: Theory and applications. *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, 48(9):1086–1093, 2001.

[9] Parlikar A., Black A., and Vogel S. Improving speech synthesis of machine translation output. In *INTERSPEECH*, pages 194–197, September 2010.

[10] Sangwan A., Chiranth M.C., Jamadagni H.S., Sah R., Prasad R.V., and Gaurav V. VAD techniques for real-time speech transmission on the Internet. In *IEEE Int. Conf. on High-Speech Networks and Multimedia Comm.*, pages 365–368, 2002.

[11] Tsiartas A., Ghosh P., Georgiou P. G., and Narayanan S. Context-driven automatic bilingual movie subtitle alignment. In *Procedings of Interspeech*, pages 444–447, Brighton, UK, 2009.

[12] Tsiartas A., Ghosh P., Georgiou P. G., and Narayanan S. Bilingual audio-subtitle extraction using automatic segmentation of movie audio. In *the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5624–5627, Prague, Czech Republic, 2010.

[13] Tsiartas A., Ghosh P., Georgiou P. G., and Narayanan S. High-quality bilingual subtitle document alignments with application to spontaneous speech translation. *Computer Speech and Language*, October 2011.

[14] Tsiartas A., Ghosh P., Georgiou P. G., and Narayanan S. Classification of clean and noisy bilingual movie audio for speech-to-speech translation corpora design. In *the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[15] Varga A. and Steeneken H. J. M. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study th effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, July 1993.

[16] Kotnik B., Kacic Z., and Horvat B. A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm. In *Proc. 7th EUROSPEECH*, pages 197–200, Aalborg, Denmark, 2001.

[17] R. Belvin, W. May, S. Narayanan, P. Georgiou, and S. Ganjavi. Creation of a doctor-patient dialogue corpus using standardized patients. *In Proc. LREC, Lisbon, Portugal*, 2004.

[18] P. F. Brown, J. C. Lai, and R. L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176, 1991.

[19] Breithaupt C., Gerkmann T., and Martin R. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In *Proc. ICASSP*, pages 4897–4900, April 2008.

[20] Busso C., Bulut M., Lee C.C., Kazemzadeh A., Mower E., Kim S., Chang J.N., Lee S., and Narayanan S. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, 2008.

[21] Callison-Burch C. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics, 2009.

[22] Lee Y. C. and Ahn S. S. Statistical model-based VAD algorithm with wavelet transform. *IEICE Trans. Fundamentals*, E89-A(6):1594–1600, June 2006.

[23] H. M. Chang. CrossTalk: technical challenge to VAD-like applications in mixed landline and mobile environments. In *Proc. 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, pages 77–80, October 1996.

[24] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (2nd Edition)*. Wiley-Interscience, 2006.

[25] Brazil D. et al. *Discourse Intonation and Language Teaching.* ERIC, 1980.

[26] Cho Y. D. and Kondoz A. Analysis and improvement of a statistical model-based voice activity detector. *IEEE Signal Proc. Letters*, 8(8):276–278, October 2001.

[27] Crystal D. *Prosodic systems and intonation in English*, volume 1. Cambridge University Press, 1976.

[28] Enqing D., Heming Z., and Yongli L. Low bit and variable rate speech coding using local cosine transform. In *Proc. TENCON*, volume 1, pages 423–426, 2002.

[29] Vlaj D., Kotnik B., Horvat B., and Kacic Z. A Computationally Efficient Mel-Filter Bank VAD Algorithm for Distributed Speech Recognition Systems. *EURASIP Journal on Applied Signal Processing*, (4):487–497, 2005.

[30] DARPA-TIMIT. Acoustic-phonetic continuous speech corpus. *NIST Speech Disc 1-1.1*, 1990.

[31] Green D.M. and Swets J.M. *Signal detection theory and psychophysics*. New York: John Wiley and Sons Inc., 1966.

[32] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[33] Florian E., Martin W., and Björn S. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[34] P. Fung and K. Mckeown. Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 81–88, Columbia, Maryland, 1994.

[35] Evangelopoulos G. and Maragos P. Speech event detection using multiband modulation energy. In *Proc. Interspeech*, volume 1, pages 685–688, Lisbon, Portugal, September 2005.

[36] Manolakis D. G., Ingle V. K., and Kogon S. M. *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. Artech House Publishers, April 2005.

[37] W.A. Gale and K.W. Church. A program for aligning sentences in bilingual corpora. *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184, 1991.

[38] Y. Gao, Gu L., Zhou B., Sarikaya R., Afify M., Kuo H.K., Zhu W., Deng Y., Prosser C., Zhang W., et al. IBM MASTOR SYSTEM: Multilingual automatic speech-to-speech translator. In *Proceedings of the Workshop on Medical Speech Translation*, pages 53–56. Association for Computational Linguistics, 2006.

[39] P. Goldberg. RATS evaluation plan. In *SAIC, Tech. Rep.*, 2011.

[40] Chang J. H. and Kim N. S. Voice activity detection based on complex Laplacian model. *IEE Electronics letters*, 39(7):632–634, April 2003.

[41] Krishnan P. S. H., Padmanabhan R., and Murthy H. A. Voice Activity Detection using Group Delay Processing on Buffered Short-term Energy. In *Proc. of 13th National Conference on Communications*, 2007.

[42] Simon Haykin. *Adaptive Filter Theory (4th Edition)*. Prentice Hall, 2001.

[43] E. Itamar and A. Itai. Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC)*, pages 269–272, Marrakech, Morocco, 2008.

[44] ITU. Coding of speech and 8 kbit/s using conjugate structure algebraic code - excited linear prediction. annex b: A silence compression scheme for g.729 optimized for terminals conforming to recommend. v.70. *International Telecommunication Union*, 1996.

[45] ITU. Coding of speech at 8 kbit/s using conjugate structure algebraic code - excited linear prediction. annex i: Reference fixed-point implementation for integrating g.729 cs-acelp speech coding main body with annexes b, d and e. *International Telecommunication Union*, 2000.

[46] Haigh J. and Mason J. S. A voice activity detector based on cepstral analysis. In *Proc. 3rd EUROSPEECH*, pages 1103–1106, Berlin, Germany, September 1993.

[47] Ramirez J., Segura J. C., Benitez C., Torre A., and Rubio A. Efficient voice activity detection algorithms using long-term speech information. In *Speech Communication*, volume 42, pages 271–287, April 2004.

[48] Sohn J., Kim N. S., and Sung W. A statistical model-based voice activity detection. *IEEE Signal Proc. letters*, 6(1):1–3, Jan. 1999.

[49] Zheng J., Mandal A., Lei X., Frandsen M., Ayan NF, Vergyri D., Wang W., Akbacak M., and Precoda K. Implementing SRI's Pashto speech-to-speech translation system on a smart phone. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 133–138. IEEE, 2010.

[50] K. Laskowski Q. Jin and T. Schultz. Cross-correlation-based multispeaker speech activity detection. In *ICSLP*, pages 973–976, Jeju Isand, Korea, October 2004.

[51] Freeman D. K., Southcott C. B., Boyd I., and Cosier G. A voice activity detector for pan-European digital cellular mobile telephone service. In *Proc. IEEE ICASSP*, volume 1, pages 369–372, Glasgow, U.K., 1989.

[52] Itoh K. and Mizushima M. Environmental noise reduction based on speech/non-speech identification for hearing aids. In *Int. Conf. on Acoust. Speech Signal Proc.*, volume 1, pages 419–422, 1997.

[53] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, 1995.

[54] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth machine translation summit*, volume 5, 2005.

[55] C. Lavecchia, K. Smaïli, and D. Langlois. Building Parallel Corpora from Movies. In *The 4th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, Funchal, Madeira, 2007.

[56] Liberman A. M. *Speech: a special code*. MIT Press, 1996.

[57] Pwint M. and Sattar F. A new speech/non-speech classification method using minimal Walsh basis functions. In *IEEE International Symposium on Circuits and Systems*, volume 3, pages 2863–2866, May 2005.

[58] P.C. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55. New Delhi, 1936.

[59] M. Mangeot and E. Giguet. Multilingual aligned corpora from movie subtitles. Technical report, LISTIC, 2005.

[60] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Online edition (c), Cambridge University Press, 2009.

[61] McCowan, A. Iain, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard. On the use of information retrieval measures for speech recognition evaluation. Idiap-RR Idiap-RR-73-2004, IDIAP, Martigny, Switzerland, 0 2004.

[62] Porter M.F. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, July 1980.

[63] Campbell N. On the use of nonverbal speech sounds in human communication. *Verbal and nonverbal communication behaviours*, pages 117–128, 2007.

[64] Meshkati N. Cultural factors inuencing safety need to be addressed in design and operation of technology. *International Civil Aviation Organization (ICAO) Journal*, 51(8):1718, 2728.

[65] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K Vesely, and Matejka. Developing a speech activity detection system for the DARPA RATS program. In *Proceedings of Interspeech*. Portland, OR, USA, 2012.

[66] World Tourism Organization. Tourism 2020 Vision: Global Forecast and Profiles of Market Segments. January 2001.

[67] Ghosh P., Tsiartas A., Georgiou P. G., and Narayanan S. Robust voice activity detection in stereo recording with crosstalk. In *Interspeech*, 2010.

[68] Ghosh P., Tsiartas A., Georgiou P. G., and Narayanan S. Robust Voice Activity Detection Using Long-Term Signal Variability. *IEEE Transactions Audio, Speech, and Language Processing*, 2011.

[69] Renevey P. and Drygajlo A. Entropy based voiced activity detection in very noisy conditions. In *Proc. EUROSPEECH*, pages 1887–1890, Aalborg, Denmark, September 2001.

[70] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics (ACL)*, pages 311–318, 2002.

[71] Aguero P.D., Adell J., and Bonafonte A. Prosody generation for speech-to-speech translation. In *ICASSP*, Toulouse, France, May 2006.

[72] T. Pfau, D. Ellis, and A. Stolcke. Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. In *Proc. ASRU*, pages 107–110, Trento, Italy, December 2001.

[73] T. Pfau and D. P. W. Ellis. Hidden markov model based speech activity detection for the ICSI meeting project. In *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[74] Padmanabhan R., Krishnan P. S. H., and Murthy H. A. A pattern recognition approach to VAD using modified group delay. In *Proc. of 14th National conference on Communications*, pages 432–437, IIT Bombay, February 2008.

[75] Prasad R., Saruwatari H., and Shikano K. Noise estimation using negentropy based voice-activity detector. In *47th Midwest Symposium on Circuits and Systems*, volume 2, pages 149–152, July 2004.

[76] Prasad R., Krstovski K., Choi F., Saleem S., Natarajan P., Decerbo M., and Stallard D. Real-time speech-to-speech translation for Pdas. In *Portable Information Devices, 2007. PORTABLE07. IEEE International Conference on*, pages 1–5. IEEE, 2007.

[77] A. Buzo R. Gra and, A. Gray, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Trans. Acoust., Speech, and Signal Proc.*, 28(4):367–376, August 1980.

[78] V. Rangarajan, S. Bangalore, and S. Narayanan. Enriching machine-mediated speech-to-speech translation using contextual information. *Computer Speech and Language*, 2011.

[79] Beritelli F. Casale S. and Cavallaro A. A robust voice activity detector for wireless communications using soft computing. *IEEE J. Select. Areas Commun.*, 16(9):1818–1829, December 1998.

[80] Beritelli F. Casale S. and Ruggeri G. A physcoacoustic auditory model to evaluate the performance of a voice activity detector. In *Proc. Int. Conf. Signal Proc.*, pages 69–72, Beijing, China, 2000.

[81] Björn S., Stefan S., Anton B., Elmar N., Alessandro V., Felix B., Rob V., Felix W., Florian E., Tobias B.and Mohammadi G., and Weiss B. The interspeech 2012 speaker trait challenge. *Interspeech, Portland, Oregon*, 2012.

[82] Greenberg S., Ainsworth W. A., Popper A. N., and Fay R. R. *Speech Processing in the Auditory System.* Illustrated edition, Springer, 2004.

[83] McClellan S. and Gibson J. D. Variable-rate CELP based on subband flatness. *IEEE Trans. Speech Audio Proc.*, 5(2):120–130, 1987.

[84] Narayanan S., Ananthakrishnan S., Belvin R., Ettaile E., Ganjavi S., Georgiou PG, Hein CM, Kadambe S., Knight K., Marcu D., et al. Transonics: A speech to speech system for english-persian interactions. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 670–675. IEEE, 2003.

[85] Kunath S.A. and Weinberger S.H. The wisdom of the crowd's ear: speech accent rating and annotation with amazon mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 168–171. Association for Computational Linguistics, 2010.

[86] Soleimani S.A. and Ahadi S.M. Voice Activity Detection based on Combination of Multiple Features using Linear/Kernel Discriminant Analyses. In *International Conference on Information and Communication Technologies: From Theory to Applications*, pages 1–5, April 2008.

[87] R. Sarikaya, S. R. Maskey, R. Zhang, E. Jan, D. Wang, B. Ramabhadran, and S. Roukos. Iterative Sentence–Pair Extraction from Quasi–Parallel Corpora for Machine Translation. In *Procedings of Interspeech*, pages 432–435, Brighton, UK, 2009.

[88] C. Schlenoff, BA Weiss, M.P. Steves, G. Sanders, F. Proctor, and A. Virts. Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies. In *Proc. of the Performance Metrics for Intelligent Systems (PerMIS) Workshop*, 2009.

[89] Mohammad S.M. and Turney P.D. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, 2010.

[90] B. D. Smedley, Stith, A., and A. Nelson. Institute of medicine committee on understanding and eliminating racial and ethnic disparities in health care. 2003.

[91] Stevens S.S., Volkmann J., and Newman EB. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[92] A. Stolcke. SRILM – an Extensible Language Modeling Toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3, pages 901–904, 2002.

[93] Cover T. and Hart P. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[94] Kinnunen T., Chernenko E., Tuononen M., Fränti P., and Li H. Voice activity detection using MFCC features and support vector machine. In *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, volume 2, pages 556–561, 2007.

[95] Takezawa T., Sumita E., Sugaya F., Yamamoto H., and Yamamoto S. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, volume 1, pages 147–152, 2002.

[96] K. Takatomo, S. Sakriani, T. Shinnosuke, N. Graham, T. Tomoki, and N. Satoshi. A method for translation of paralinguistic information. *Proceedings IWSLT 2012*, 2012.

[97] J. Tiedemann. Building a multilingual parallel subtitle corpus. In *Proceedings of CLIN*, volume 17, Leuven, Belgium, 2007.

[98] J. Tiedemann. Improved sentence alignment for movie subtitles. In *Proceedings of RANLP*, pages 582–588, Borovets, Bulgaria, 2007.

[99] J. Tiedemann. Synchronizing Translated Movie Subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 1902–1906, Marrakech, Morocco, 2008.

[100] A. Tsiartas, T. Chaspari, N. Katsamanis, P. Ghosh, M. Lee, , M. Van Segbroeck, A. Potamianos, and S. Narayanan. Multi-band long-term signal variability features for robust voice activity detection. In *Proc. Interspeech, Lyon*, 2013.

[101] A. Tsiartas, P. Georgiou, and S. Narayanan. A study on the effect of prosodic emphasis transfer on overall speech translation quality. In *Proc. IEEE ICASSP*. IEEE, 2013.

[102] A. Tsiartas, P. G. Georgiou, and S. Narayanan. Toward transfer of acoustic cues of emphasis accross languages. In *Proc. Interspeech, Lyon*, 2013.

[103] Ambati V., Vogel S., and Carbonell J. Active learning and crowd-sourcing for machine translation. *Language Resources and Evaluation (LREC)*, 7:2169–2174, 2010.

[104] K. Walker and S. Strassel. The RATS Radio Traffic Collection System. In *Odyssey 2012-The Speaker and Language Recognition Workshop*. Singapore, 2012.

[105] S. N. Wrigle, G. J. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multi-channel audio. *IEEE Trans. Speech and Audio Proc*, 3(1):84–91, January 2005.

[106] H. Xiao and X. Wang. Constructing Parallel Corpus from Movie Subtitles. *Proceedings of International Conference on Computer Processing of Oriental Languages*, pages 329–336, 2009.

[107] M. Xu, L. Y. Duan, J. Cai, L. T. Chia, C. Xu, and Q. Tian. HMM-based audio keyword generation. *Advances in Multimedia Information Processing - PCM 2004: 5th Pacific Rim Conference on Multimedia.*, 2004.