BILINGUAL AUDIO-SUBTITLE EXTRACTION USING AUTOMATIC SEGMENTATION OF MOVIE AUDIO

Andreas Tsiartas, Prasanta Ghosh, Panayiotis G. Georgiou, Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory,
Department of Electrical Engineering,
University of Southern California,
Los Angeles, CA 90089

tsiartas@usc.edu, prasantg@usc.edu, georgiou@sipi.usc.edu, shri@sipi.usc.edu

ABSTRACT

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%.

***Index Terms*—** bilingual movie audio, movie subtitle, audio segmentation

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 [1] 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*. [2] showed BLUE score [3] improvements on a large-scale S2S system. In our past work [4], we focused in linking the speech transcriptions of movies as shown in Fig. 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 [5] and Basic Travel Expression Corpus (BTEC) [6]. However, very little work has been done to automatically acquire and align bilingual speech utterances. In this work, we focus on segmenting parallel audio from movies and aligning the segments with the corresponding subtitles as shown in Fig. 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.

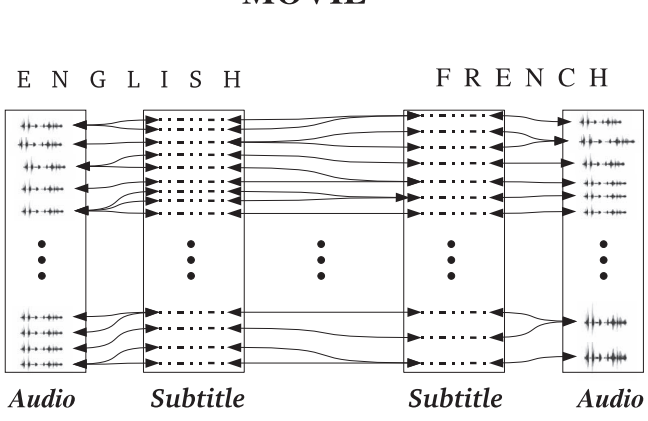

Fig. 1. An illustration of bilingual audio streams and subtitles alignment and segmentation between English and French.

This paper 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.

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

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

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. 2.

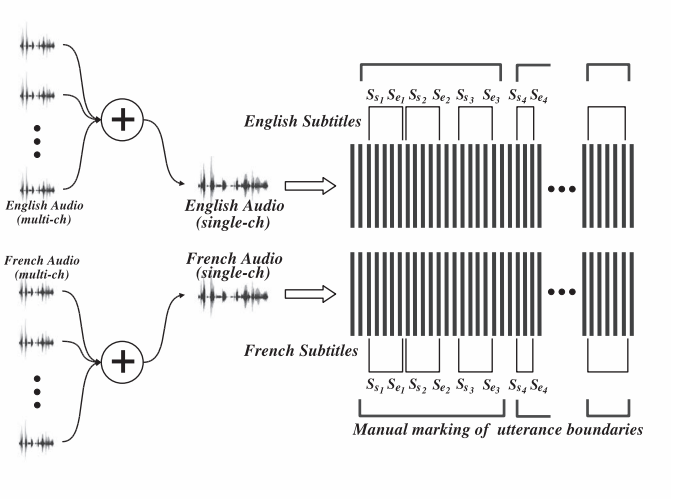

Fig. 2. An illustration of the manually tagged bilingual audio streams for English and French.

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.

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:

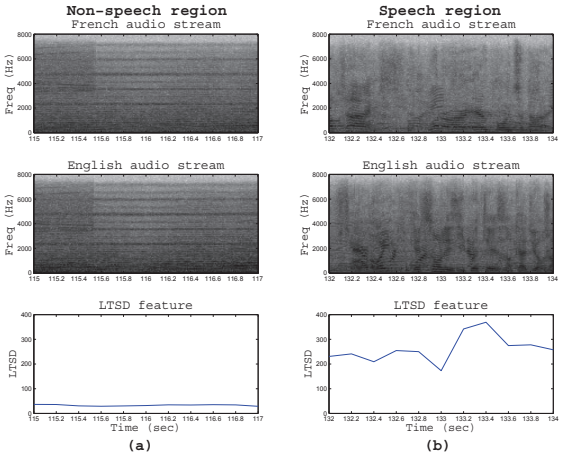

Fig. 3. Fig. 3(a) shows the spectrogram of French and English non-speech audio regions along with the value of LTSD. Fig. 3(b) shows the spectrogram of French and English speech audio regions along with the value of LTSD.

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

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. 3 indicates that the LTSD values differ significantly for parallel speech and non-speech regions.

3.2. Spectral distance

To approximate the acoustical and perceptual proximity of the two audio streams, we use the mel-frequency cepstral coefficients (MFCC) [7], 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

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

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 suggests that, 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.

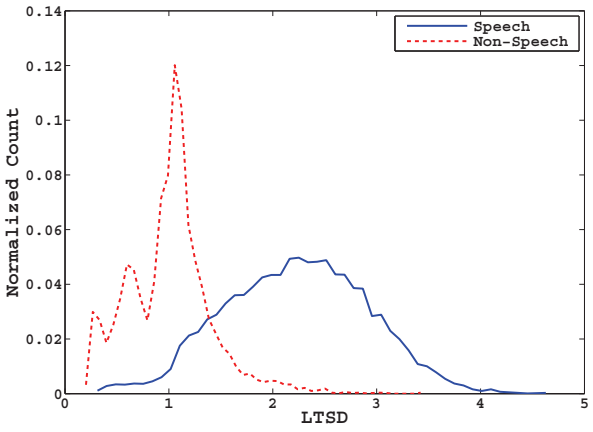

Fig. 4. Distribution of LTSD for speech and non-speech frames

	Accuracy
LTSD	41.39%
Subtitles	37.88%

Table 1. Table shows the percentage of segments that gave perfect segmentations using subtitles and the LTSD feature.

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 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 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.

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. 2, 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. 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. Experimental evaluation setup

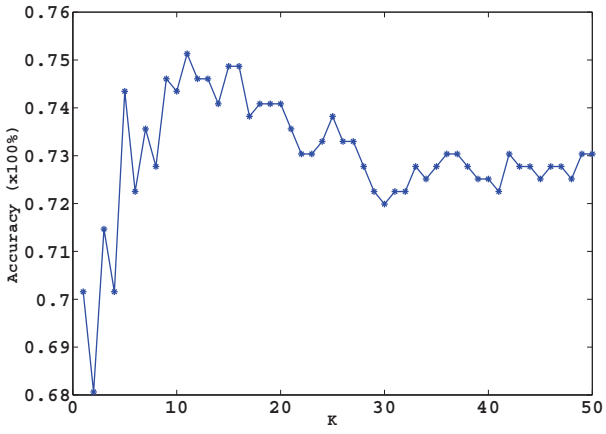

Fig. 5. K-NN accuracy on the development set for various K.

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. 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

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

Table 2. Table shows the percentage of the audio segments and subtitle alignments rated as "Full", "Partial" and "None".

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".

5. RESULTS AND DISCUSSION

Table 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 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. 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.

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.

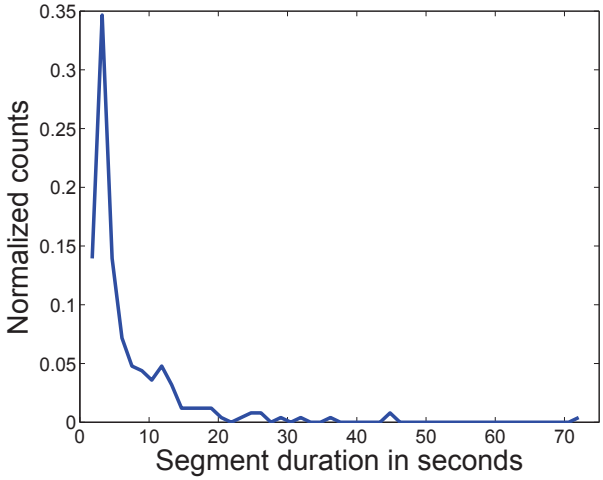

Fig. 6. Distribution of the duration of the resulting segments.

7. REFERENCES

[1] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the tenth machine translation summit*, 2005, vol. 5.

[2] 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*, Brighton, UK, 2009, pp. 432–435.

[3] 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)*, 2002, pp. 311–318.

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

[5] 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.

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

[7] J. Cai L. T. Chia C. Xu M. Xu, L. Y. Duan and Q. Tian, "HMM-based audio keyword generation," *Kiyoharu Aizawa, Yuichi Nakamura, Shin'ichi Satoh. Advances in Multimedia Information Processing - PCM 2004: 5th Pacific Rim Conference on Multimedia.*, 2004.