INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

Using Acoustic Diarization for Duplicate Detection

Mary Knox, Gerald Friedland, and R. Paul Smith

TR-12-005

March 2012

Abstract

The following article describes the use of an acoustic diarization engine for duplicate detection on broadcast news. Diarization is typically used to partition audio into speaker homogeneous regions, or in other words, to determine “who spoke when.” In this setting, however, we use diarization to segment the recordings and group the segments into homogeneous clusters. Diarization is performed both on the full length broadcast news recordings as well as the short clips (which we are classifying as either a duplicate or not). We then compare the similarity of models trained on the clusters to determine whether the time allocated to the cluster from the short clip is from the original broadcast news recording, or a duplicate. We tested our system under a variety of audio conditions: unmodified, with reverberation, resampled, and lowpass filtered. On our test set, the areas under the receiver operating characteristic curve for the audio conditions were 0.91, 0.89, 0.61, and 0.64 respectively.

1. INTRODUCTION

The topic of duplicate detection is relevant to a variety of areas, including data deduplication, copyright infringement, and social networking. In particular, the problem of data deduplication has been receiving increasing attention in recent years due to the increase in the amount of data taken, stored, and shared. A key aspect of data deduplication is comparing chunks of data to previously stored data and identifying whether there is an appropriate match, or "duplicate". Copyright infringement is another area in which it is important to be able to identify a duplication of copyrighted material. From the social networking perspective there is growing awareness that finding others who have done mashups or have performed simple multimedia modifications on the same data could be highly useful tools for connecting individuals together or identifying piracy.

In this paper, we investigate a novel method of determining whether a short clip is a "duplicate" from a broadcast news recording based on acoustic diarization. In this setting a duplicate is a recording that has the same content as another recording, though the two files do not necessarily have identical binary encodings (due to editing or filtering). We tested our algorithm under four audio conditions: unmodified, with reverberation, resampled, and lowpass filtered. Our algorithm performed well under both the unmodified and reverberation conditions achieving areas under the receiver operating characteristic curve (ROC AUC) values of 0.9. Performance degraded in the resampled and lowpass filtered conditions, achieving ROC AUC values of 0.6.

This paper is outlined as follows: in Section 2 we describe related work, in Section 3 we describe our duplicate detection system, in Section 4 we provide and discuss the experiments and results, and in Section 5 we give our conclusions as well as areas of future work.

2. BACKGROUND

Searching and identifying similar content is a long standing problem in areas of multimedia research. Similarity detection has been used for recommendation systems (e.g., suggest songs to listen to), searching, and copyright infringement. These tasks have different goals but all measure similarities between items.

We briefly outline some of the previous work in these varied areas with an emphasis on searching for perceptually similar content [1, 2]. A review of audio fingerprinting is given in [3]. There exist many techniques in the computer vision community on video copy detection [4] and TRECVID [5] has a copy detection evaluation track. In [6], the authors map each test frame to the nearest query frame and achieve robust audio copy detection using normalized detection cost rates. A query-by-example audio retrieval framework by indexing audio clips in a generic database as points in a latent perceptual space is presented in [7].

There has also been similarity work done in the music community. In [8], the authors describe the algorithm behind Shazam, a popular commercial application used on many mobile devices to recognize music. In [9], the authors aim to identify remixed audio tracks using audio shingles with locality sensitive hashing. Their method identifies remixes based on whether the shingles are similar, thus the remix does not need to have similar spectral content for the entire song. In [10], the authors also investigate duplicate detection for the music setting.

In this work, we investigate the use of a diarization engine for duplicate detection. The goal of speaker diarization is to partition an input stream into speaker homogeneous speech regions, as shown in Figure 1, where the number of speakers as well as the speaker identities are not known a priori. The most common approach to speaker diarization is agglomerative hierarchical clustering [11]. In this method, the system is initialized with many clusters, much greater than the number of speakers, and iteratively merges clusters until a stopping criterion is met. Though the goal of speaker diarization differs from that of duplicate detection, we found that it is a useful method to segment the data and cluster similar data together. We can then measure the "distance" between clusters to determine if one of the clusters is a duplicate of the other.

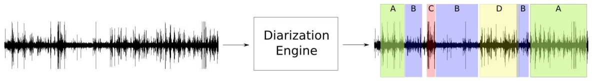

Fig. 1: Overview of speaker diarization. From an input audio signal (with no prior knowledge of the number of speakers, speaker identities, or speech segmentation), segment the signal into nonspeech and speech segments, the latter labeled by speaker (e.g., A, B, C, D).

3. SYSTEM DESCRIPTION

First, diarization is performed on each of the original broadcast news recordings. We use the diarization engine to segment each recording and group similar segments together into clusters, where a GMM is trained on each of the clusters. We then run diarization of each of the short clips. The short clips are then evaluated to determine if they are in fact a duplicate. In order to determine if a short clip is a duplicate, we compute the symmetric KL divergence between each cluster from the short clips and all of the clusters from all of the original broadcast news recordings. A small symmetric KL divergence value means the two clusters are very similar and the time associated with the cluster from the short clip is likely a duplicate.

3.1. Features

We extract Mel-Frequency Cepstral Coefficients (MFCCs) to describe the recordings. We compute the first 19 MFCCs, which are computed over a 30 ms window with a 10 ms forward shift. The MFCC features are extracted using the Hidden Markov Model Toolkit (HTK) [12]. MFCCs are standard features in the speaker diarization community.

3.2. Diarization Engine

The diarization engine used in this study is based on the state-of-the-art ICSI speaker diarization system, which is described in more detail in [13, 14]. The system performs both segmentation and clustering, which are performed iteratively using an agglomerative clustering approach. Segmentation entails identifying the boundaries where audio changes occur (e.g. speaker changes). Clustering is grouping segments which contain similar audio together. Usually, the speaker diarization engine first separates the speech and non-speech regions and then subsequently deals only with the speech regions. However, since the goal of the work is to detect duplication we suspected that omitting nonspeech time (which includes silence as well as other not speech sounds) could be detrimental so we use all portions of the recording.

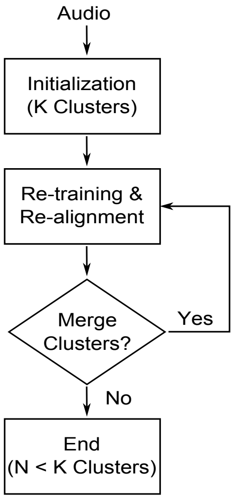

Fig. 2: Flowchart of the diarization engine.

An overview of the diarization engine is shown in Figure 2. Diarization is performed using a Hidden Markov Model (HMM) where each state (or cluster) is modeled as a GMM with a minimum duration constraint of m seconds. The minimum duration m lower bounds the shortest segment duration and prevents the system from having many "speaker" changes within a short amount of time. We initially define K clusters, where the number of clusters is equal to the number of HMM states. K is chosen to be much greater than the number of speakers in the recording. The GMM parameters are initialized after segmenting the data into K uniform regions. Re-segmentation is performed using Viterbi decoding and the GMMs are retrained based on the new segmentation. The clusters are merged based on the delta Bayesian Information Criterion (ΔBIC), shown in Equation (1). More specifically, the cluster pair with the greatest ΔBIC score greater than zero is merged. In this system, when two clusters are merged the number of parameters for the new cluster is equal to the sum of the parameters in the clusters that are merged which results in the simplified BIC equation shown below.

$$\Delta\text{BIC} = \log p(D_{1,2}|\theta_{1,2}) - (\log p(D_1|\theta_1) + \log p(D_2|\theta_2)) \tag{1}$$

where D_1 and D_2 are the data from clusters 1 and 2, $D_{1,2}$ is the data from $D_1 \cup D_2$, and θ represents the parameters for the respective models [15]. After two clusters are merged, we repeat the process of retraining the GMMs, re-segmenting the data, and determining which clusters to merge (assuming Equation (1) is greater than zero, as shown in Figure 2. Diarization concludes once no cluster pair has a ΔBIC value greater than zero.

3.3. Symmetric Kullback-Liebler Divergence

We use the symmetric Kullback-Liebler (KL) divergence to quantify the difference between the probability distributions of two clusters (one cluster from the short clip and one cluster from the original broadcast news recording). The KL divergence is defined as

$$D_{KL}(f(x)\|g(x)) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx \tag{2}$$

where $f(x)$ is the probability distribution of the first cluster and $g(x)$ is the probability distribution of the second cluster. Similarly, we define the symmetric KL divergence as

$$D_{KL,sym}(f(x), g(x)) = D_{KL}(f(x)\|g(x)) + D_{KL}(g(x)\|f(x)) \tag{3}$$

We use the unscented transform based approximation of the KL divergence [16]. This approximation, used specifically for the case of GMM probability distributions, was shown to work well for speaker recognition [16] as well as speaker diarization [17]. The unscented transform based approximation is deterministic and subsequently very efficient to compute [16].

Let X be a D-dimensional GMM with distribution $f(x) = \sum_{i=1}^{M} w_i N(\mu_i, \Sigma_i)$, where M is the number of mixture components, w_i is the mixture weights, μ_i is the mean vector of the ith component, and Σ_i is the covariance matrix of the ith component. Then the unscented transform can be used to approximate the expectation of $\log g(x)$ by evaluating Equation (4) at a number of sigma points $x_{i,k}$.

$$\begin{aligned}\mathbb{E}[\log g(x)] &= \int_{-\infty}^{\infty} f(x) \log g(x) dx \\ &\approx \frac{1}{2D} \sum_{i=1}^{M} w_i \sum_{k=1}^{2D} \log g(x_{i,k}) \qquad (4)\end{aligned}$$

where

$$\begin{aligned} x_{i,k} &= \mu_i + (\sqrt{D\Sigma_i})_k \quad k = 1, ..., D \\ x_{i,D+k} &= \mu_i + (\sqrt{D\Sigma_i})_k \quad k = 1, ..., D \qquad (5)\end{aligned}$$

and $(\sqrt{\Sigma})_k$ is the kth column of the matrix square root of Σ. In our work, we used a diagonal covariance matrix so Equation 5 is further simplified to

$$\begin{aligned} x_{i,k} &= \mu_i + \sqrt{D}\sigma_{i,k} \mathbb{1}_{index=k} \quad k = 1, ..., D \\ x_{i,D+k} &= \mu_i + \sqrt{D}\sigma_{i,k} \mathbb{1}_{index=k} \quad k = 1, ..., D \quad (6)\end{aligned}$$

where $\mathbb{1}_{index=k}$ is a D dimensional vector where the kth index is one and all other values are zero. We use Equations 4-6 to approximate the symmetric KL divergence between GMMs trained on clusters from the short clips and GMMs trained on clusters from the original broadcast news recordings. The symmetric KL divergence was then used to determine whether the time associated with the cluster from the short clip was from one of the original broadcast news recordings.

4. EXPERIMENTS AND RESULTS

In this section, we describe the dataset used to evaluate our duplicate detection system, the method of scoring, and the results obtained on both the development and test sets.

4.1. Dataset

We evaluated our results on approximately 6.5 hours of broadcast news video recordings, consisting of thirteen 30 minute recordings (which included commercials in addition to the news program). Though both video and audio were available, in this work we focus only on the audio.

In order to explore how the system works for a variety of audio clips, we evaluated the system using clips of variable duration and under different audio conditions. More specifically, we extracted 15, 30, and 60 second clips at regular intervals. The clip midpoints were every 100 seconds with the first midpoint at 100 seconds and the last midpoint at 1600 seconds. We also investigated performance when the audio was unmodified, lowpass filtered with a 1750 Hz cutoff, downsampled from 44.1 kHz to 8kHz, and included reverberation. We use `sox` [18] to modify the audio recordings. More specifically, for the reverberation setting we use a 75% gain and a 75 ms delay.

We split the broadcast news recordings into a development set and test set. The development set consists of eight recordings and the test set consists of five records, resulting in a total of 1536 and 960 clips respectively. In Table 1, we show the breakdown of the development and test sets which were randomly chosen. The names given to each recording include the year, month, day, start time, end time, and network the program aired on.

Table 1: Development and test set broadcast news recordings.

Development	Test
19980513-1130-1200-CNN	19980515-1130-1200-CNN
19980513-1830-1900-ABC	19980518-1130-1200-CNN
19980518-1830-1900-ABC	19980519-1830-1900-ABC
19980519-1130-1200-CNN	19980520-1830-1900-ABC
19980520-1130-1200-CNN	19980522-1830-1900-ABC
19980523-1130-1200-CNN	
19980523-1830-1900-ABC	
19980524-1130-1200-CNN	

4.2. Scoring

In order to evaluate our results we use the Receiver Operating Characteristic (ROC) Area Under Curve (AUC). The ROC is a plot of the true positive rate versus the false positive rate. In order to compute the true positive and false positive rates, we thresholded the symmetric KL divergence between the GMMs trained on clusters from the clips and GMMs trained on clusters from the original broadcast news recordings. If the symmetric KL divergence for a given cluster pair, where one cluster is from the clip and the other is from the original broadcast news recording, is less than the threshold then the cluster from the clip is classified as a duplicate of the original broadcast news recording. Otherwise, the cluster from the clip is not a duplicate. Cluster pairs are labeled in the reference as a match if time annotated to the cluster from clip overlapped with any time annotated to the cluster from the original recording. Note that the ROC plots were computed such that each cluster pair had equal weight.

4.3. Results

In this section we describe results on both our development and test sets.

4.3.1. Development Set Results

We first needed to determine parameter settings for the diarization engine, specifically the number of initial clusters K

and minimum duration m. These were the only parameters we tuned for the duplicate detection system.

We first focused on the number of initial clusters K used to run the diarization engine on the original broadcast news recordings. We ran experiments using 16, 32, 64, 128, and 256 initial clusters for the original 30-minute broadcast news recordings. Empirically, we found that 128 clusters performed best since it resulted in a number of clusters most similar to the number of speakers in the recording.

Next, we investigated performance for a number of minimum duration values. We also varied the number of initial clusters for the short clips. We ran the diarization engine with 128 initial clusters and a number of minimum duration values (1, 1.5, 2, and 2.5 seconds) on the original broadcast news recordings. For each of the clips we used $K = 1, 2, ..., 8$ initial clusters and $m = 1, 1.5, 2, 2.5$ seconds minimum durations. We only computed the symmetric KL divergence between the GMMs from the clips and original broadcast news recordings that had the same minimum duration. We evaluated the results on the four audio settings (unmodified, lowpass filtered with a 1750Hz cutoff, downsampled from 44.1 kHz to 8kHz, and reverberation) and for the various duration clips (15, 30, and 60 seconds).

We found that a minimum duration $m = 2.5$ seconds worked best for the unmodified and reverberation setting while $m = 1.5$ seconds worked best for the resampled and lowpass filtered settings. Though the variances of the ROC AUC values were small for all of the settings, the results for the resampled and lowpass filtered settings had less variance so we set the minimum duration to 2.5 seconds. Based upon the development set results for the short clips as well as the previous conclusion to use 128 initial clusters for the original 30-minute broadcast news recordings, we set the number of initial clusters to $K = \text{round}(\text{clip duration in seconds}/14.0625 + 1)$, where 14.0625 was chosen since it is equal to $128/(30 \cdot 60)$. Thus, for the 15, 30, and 60 second clips we started with 2, 3, and 5 clusters respectively. Though again, we found that the variance of the ROC AUC values was very small when varying the number of initial clusters. Having small variance in the ROC AUC values when using a number of initial clusters and minimum durations is promising since the results are not too different based on the parameter selection. Figure 3 shows the results on the development set for all of the audio conditions using the diarization parameters settled upon in this section. The numbers included in the legend are the ROC AUC values for the respective settings.

4.3.2. *Test Set Results*

Using the parameters determined from the development set results, namely a 2.5 second minimum duration and 2,3,and 5 initial clusters for the 15, 30, and 60 second clips respectively, we evaluated the test set clips. We compared the GMMs trained on each cluster from the short test set clips to the GMMs trained on all of the clusters from the five original 30-minute broadcast news recordings which make up the test set. The ROC plots for the unmodified, lowpass filtered, downsampled, and reverberation settings are shown in Figure 4. Each plot shows the results for the 15, 30, and 60 second clips as well as the result when the all of the clips are included. We also included all of the audio conditions into a single ROC plot shown in Figure 5.

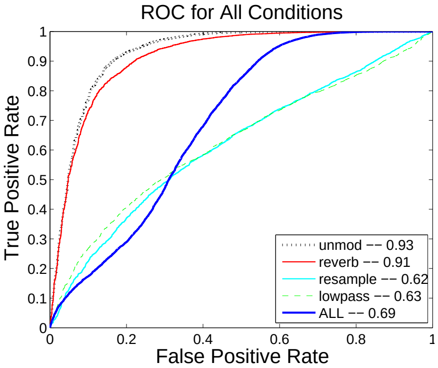

Fig. 3: ROC plot for all audio conditions on the development set.

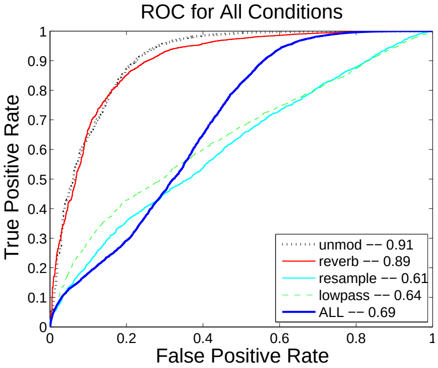

Fig. 5: ROC plot for all audio conditions on the test set.

5. CONCLUSIONS AND FUTURE WORK

We introduced a novel method utilizing diarization for identifying duplicate clips. The diarization engine was used to split both the original broadcast news recordings as well as the short clips into homogeneous clusters. Then the symmetric KL divergence was used to determine whether the the time annotated to the cluster from the short clip was a duplicate of the original broadcast news recording.

There were two tunable parameters in the diarization system, the number of initial guassians K and the minimum

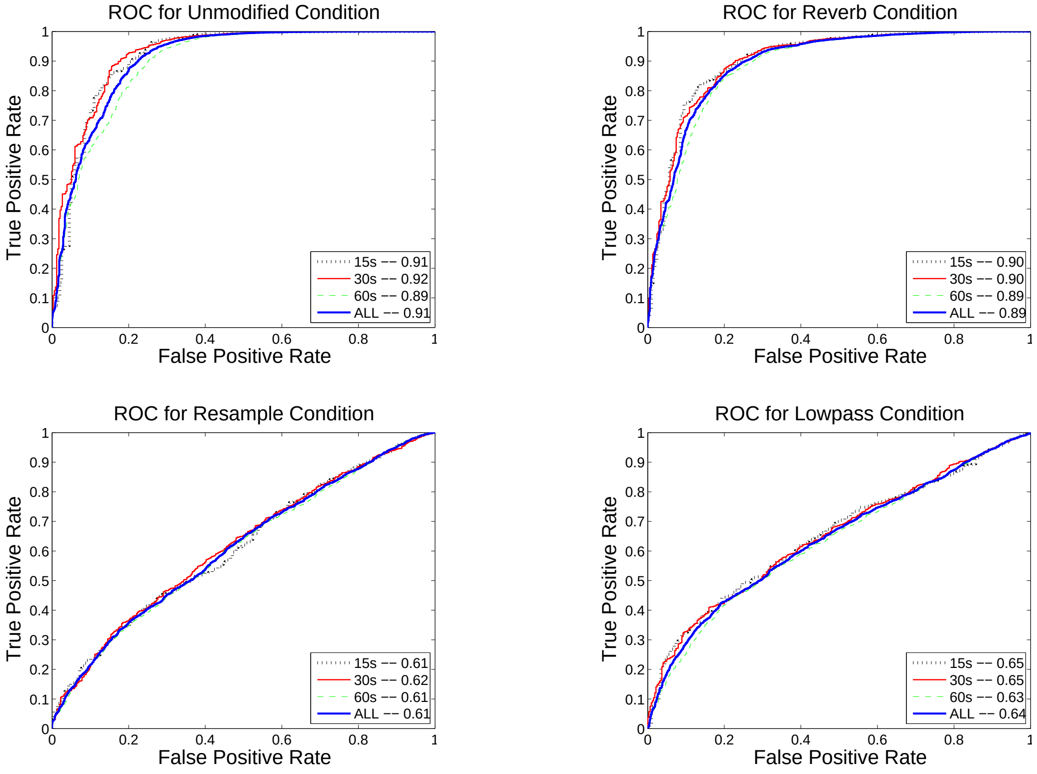

Fig. 4: ROC plots when the test set clip audio is unmodified, contains reverberation, resampled, and lowpass filtered.

duration m. The results obtained on the development set were not very sensitive to the parameter settings. However, we settled on using a minimum duration of 2.5 seconds and $K = \text{round}(\text{clip duration in seconds}/14.0625 + 1)$.

We tested our method on a variety of clips. Our test set included 15, 30, and 60 second clips. We also evaluated our results on four audio conditions: unmodified, with reverberation, resampled, and lowpass filtered. We found that performance was best under the unmodified and reverberation conditions, achieving ROC AUC values of 0.9. Performance degraded under the resampled and lowpass filtered condition, however, we were still able to achieve ROC AUC values of 0.6. For all of the clips, we achieved an ROC AUC value of 0.7.

In the future, we plan to investigate our method on other datasets, particularly those used in the TRECVID evaluations [5]. We also plan to incorporate video features into our system.

6. REFERENCES

[1] Y. Jiao, B. Yang, M. Li, and X. Niu, "MDCT-based perceptual hashing for compressed audio content identification," in *IEEE 9th Workshop on Multimedia Signal Processing*, 2007, pp. 381–384.

[2] Q. Li, J. Wu, and X. He, "Content-based audio retrieval using perceptual hash," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2008.

[3] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," in *Journal of VLSI Signal Processing*, 2005, vol. 41, pp. 271–284.

[4] M. Douze, H. Jegou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," in *IEEE Transactions on Multimedia*, 2010, vol. 12.

[5] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[6] V. Gupta, G. Boulianne, and P. Cardinal, "Content-based audio copy detection using nearest-neighbor mapping," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, 2010, pp. 261–264.

[7] S. Sundaram and S. Narayanan, "Audio retrieval by latent perceptual indexing," in *International Conference*

on Acoustics Speech and Signal Processing (ICASSP), Las Vegas, Nevada, 2008, pp. 49–52.

[8] A. Wang, "An industrial strength audio search algorithm," in *Int. Conf. Music Info. Retrieval*, Baltimore, Maryland, 2003.

[9] M. Casey and M. Slaney, "Fast recognition of remixed music audio," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, 2007.

[10] C.J.C. Burges, D. Plastina, J.C. Platt, E. Renshaw, and H.S. Malvar, "Using audio fingerprinting for duplicate detection and thumbnail generation," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2005, pp. 9–12.

[11] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization : A review of recent research," in *Accepted for publication in IEEE Transactions on Audio, Speech and Language Processing (TASLP), special issue on New Frontiers in Rich Transcription*, 2011.

[12] "Hidden markov model toolkit (HTK)," http://htk.eng.cam.ac.uk/.

[13] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proceedings of the RT07 Meeting Recognition Evaluation Workshop*, 2007.

[14] G. Friedland, A. Janin, D. Imseg, X. Anguera, L Gottlieb, M. Huijbregts, M. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," in *IEEE Transactions on Audio, Speech and Language Processing*, to appear 2012.

[15] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of IEEE Speech Recognition and Understanding Workshop*, St. Thomas, US Virgin Islands, 2003.

[16] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unscented transform and its application to speaker recognition," in *Proceedings of Interspeech*, 2005, pp. 1985–1988.

[17] Y. Huang, O. Vinyals, G. Friedland, C. Muller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *IEEE ASRU*, Kyoto, Japan, 2007, pp. 693–698.

[18] "Sound eXchange," http://sox.sourceforge.net.