ADVANCED APPROACHES TO SPEAKER DIARIZATION OF AUDIO DOCUMENTS

Konstantin Markov[*]

Human Interface Lab, School of Computer Science and Engineering,
The University of Aizu, Fukushima, Japan

ABSTRACT

Speaker diarization is the process of annotating an audio document with information about the speaker identity of speech segments along with their start and end time. Assuming that audio input consists of speech only or that non-speech segments have been already identified by another method, the task of speaker diarization is to find "who spoke when". Since there is no prior information about the number of speakers, the main approach is to apply segment clustering. According to the clustering algorithm used, speaker diarization systems can be divided into two groups: 1) based on agglomerative clustering, and 2) based on on-line clustering. Agglomerative clustering is an off-line approach and is used in most of the current systems because it gives accurate results and can be fine tuned by performing several processing passes over the data. This, however, comes at the cost of high computational load which increases exponentially with the number of segments and the requirement of having the whole audio document available in advance. In contrast, on-line clustering based systems have almost constant computational load, work on-line in real time with small latency, but are generally less accurate than off-line systems. As we show in this paper, when using advanced on-line learning methods and original design, on-line systems can make less errors than off-line systems and can even work faster than real time with very low latency.

Index Terms— Speaker diarization, Speaker segmentation, On-line GMM learning.

1. INTRODUCTION

The task of efficient and effective automatic indexing and searching of the growing volumes of recorded spoken documents, such as broadcasts, voice mails, meetings and others, requires human language technologies that can not only transcribe speech, but can also extract different kinds of nonlinguistic information. This information, often called metadata, includes speaker turns, channel changes, and others. Identifying and labeling the sound sources within a spoken document is the task of audio diarization. A main part of the audio diarization process is the speaker diarization or speaker segmentation and clustering. In other words, it is the task to find out "who spoke when".

Speaker diarization is currently the focus of the most efforts in the audio diarization research. Broadcast news audio, meetings recordings or telephone conversations are one of the main domains for speaker diarization research and development. In some cases, prior information about the task can be available. This may be an example speech from speakers of a meeting or from the main anchors of a broadcast. However, from a system portability point of view, it is better to use less or no prior knowledge at all.

Most of the current speaker diarization systems perform several key sub-tasks which are: Speech detection, Speaker change detection, Gender classification and Speaker clustering [1]. To improve the performance, in some cases, cluster recombination and re-segmentation are also used [2]. The speech detection is aimed to find those regions of the audio which consist of speech only. The most popular technique to perform this task is the maximum-likelihood classification with Gaussian mixture models (GMM). They are usually trained in advance from some labeled data and, in the simplest case, there are only two models for speech and non-speech data [3]. Some systems use several models depending on the speaker gender and the channel type [4, 5]. Another approach that has been found useful is to perform a single or multi-pass Viterbi segmentation of the audio stream [6, 7]. After speech segments are identified, speaker change detection is used to find out any possible speaker change within every segment. If such is detected, the segment is further split into smaller segments each of which belongs to a single speaker. There are two main techniques for change detection. The first one finds potential change point in a window by determining whether it is better modeled by two rather than one distribution using the Bayesian information criterion (BIC) [6]. The second one is based on measuring the distance, Gaussian divergence [8] or generalized likelihood ratio [9], between two fixed length windows represented most often by a single Gaussian. A distance peak that is above certain threshold is then considered as a change point. The gender classification is used to split the segments into two groups (male and female) which reduces the load of the next clustering task as well as to give more information about the speakers. Typi-

[*] This work was done while the author was with the ATR Spoken Language Communication Research Lab, Japan.

cally, two GMMs, one for each gender, are trained in advance and maximum-likelihood is used as decision criterion. The last sub-task, the speaker clustering, is to assign each segment with its correct speaker label. This is done by clustering segments into sets corresponding to speakers. The most widely used approach is hierarchical, agglomerative clustering with BIC stopping criterion [7, 10]. Each cluster is usually represented by a single Gaussian and the generalized likelihood ratio (GLR) [11] has been commonly used as between clusters distance measure. Variations of this method have also been proposed [5, 12], but they are still based on the same bottom-up clustering technique. Although, quite successful, agglomerative clustering approach has several drawbacks that limit the potential use of the speaker diarization systems in the real-world, real-time applications. First, it requires all the speech segments to be available before the clustering starts and, therefore, makes on-line processing impossible. Second, the computational load increases almost exponentially with the number of segments [13]. Finally, the performance is greatly affected by the stopping criterion which is considered a critical part of the algorithm [1].

There are situations in which the task of speaker diarization must be performed on-line as the data steams in and the clustering has to be done sequentially. A generic method known as the leader-follower clustering [14] is the basis of most of the on-line systems. One such system has been proposed recently [13], where, as in the agglomerative clustering method, the speech segments are modeled by a single Gaussian distribution and the GLR is used as a distance metric. This reduces the clustering accuracy for short segments and delays the decision until the whole segment is received. In consequence, the system latency becomes dependent on the segment's length which can be up to 30 sec. or even longer. Another sequential technique where speakers are represented by subspaces has also been studied [15]. However, it requires at least 5 sec. long speech segments and has high miss and false alarm rates. The speaker diarization system we have developed [16] is also based on the leader-follower idea, but speakers are represented by Gaussian Mixture models (GMMs) rather than clusters of speech segments. In our system, when assigning speaker label to a given segment, first, it is decided whether it belongs to one of the known speakers or to a new speaker. Then, in the former case, speaker identification is performed and the winning speaker label is assigned to the segment. In the latter case, new speaker is registered to the system and his/her model is created. This is similar to the classical open-set speaker identification task. Each GMM is learned on-line every time it has been a winner. New speaker's GMM is created by spawning a speaker independent male or female GMM trained in advance. In addition, each speaker GMM has a time counter which is set to zero whenever it wins the identification. Otherwise, the counter is incremented by the current segment length. Models whose counter reaches some threshold T, are deleted from the system. This way, the system can operate indefinitely, adapting itself to the environment changes.

2. AGGLOMERATIVE CLUSTERING BASED SYSTEM

2.1. Overview

As we mentioned in Section 1, most of the speaker diarization systems perform voice activity detection, speaker change detection, gender identification and speech segment clustering and the block diagram of such system based on agglomerative clustering is shown in Fig.1.

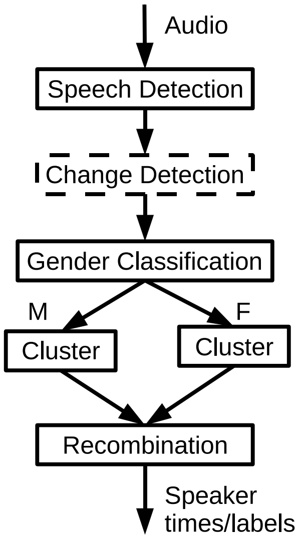

Fig. 1. Block diagram of agglomerative clustering speaker diarization system.

Unsegmented audio data is fed to the voice activity detection module which outputs speech segments start and end time points. In most cases, detected speech segments are homogeneous, i.e. they come from one speaker, but if there are likely multi-speaker segments, optionally, speaker change detection is performed. In our system, we dont use speaker change detection. Next, for each speech segment, speaker gender is determined by the gender identification module and segments from male and female speakers are pooled into two separate sets. Then segment clustering is performed on each set and after simple time re-ordering, output speaker labels and times are obtained.

Next subsections briefly describe each module of the system.

2.2. Voice activity detection

For the voice activity detection (VAD), we use the standard model based approach. Non-speech events (pauses in this case, but other event can also be modeled) are represented by a single GMM and the speech is modeled by the two gender dependent GMMs. For each frame, the non-speech and speech (the better one from the two GMMS) likelihoods are passed through two separate median filters and the frame's label (speech / non-speech) is assigned by comparing the filters output. Then, a simple logic decides segments start and end points taking into account such parameters as minimum segment length, maximum pause in segment and maximum speech in pause.

2.3. Gender identification

The gender identification module uses the same gender dependent GMMs as the VAD module. Frame likelihoods calculated already during the voice activity detection are accumulated from the segment's start point. Then, the speaker gender is determined by a simple maximum-likelihood classification.

2.4. Segment clustering

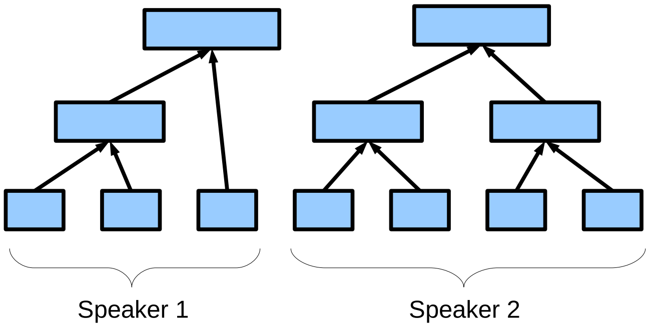

Fig. 2. Agglomerative clustering. At the end of the algorithm, segments from two speakers (bottom row) are clustered into two clusters (top row).

The agglomerative clustering procedure is schematically shown in Fig. 2, where the lowest row shows seven speech segments from two speakers. Next row shows how they are clustered into four clusters. The top row represents the clustering result where segments from the two speakers are clustered into two clusters. Each cluster is modeled by a single Gaussian function with full covariance matrix and generalized likelihood ratio (GLR) was used as inter-cluster distance measure. At each iteration of the clustering procedure, two most closest clusters are merged. Merging is stopped when the change in the Bayesian information criterion statistic (ΔBIC) turns positive. The GRL and ΔBIC are defined as follows:

$$GLR_{x,y} = \frac{|\Sigma_{x\cup y}|^{N_{x\cup y}/2}}{|\Sigma_x|^{N_x/2}|\Sigma_y|^{N_y/2}} \qquad (1)$$

$$\Delta BIC = \log GLR_{x,y} - \alpha \left(\frac{d(d+3)}{4}\right) \log N_{x\cup y}$$

where x and y are the two clusters to be merged, N is the number of frames in the cluster, d is the feature vectors dimension, and α is a free parameter tuned on the development data.

3. ON-LINE CLUSTERING BASED SYSTEM

3.1. Overview

The on-line system uses the same VAD and gender identification modules as the agglomerative clustering system. The main difference is the way speech segments are clustered.

The system operation is schematically shown in Fig. 3. The speech segments and their reference speaker labels are at the top of the figure. The bottom part shows the speaker models and how they change in time. For each speech segment, there is a winning model indicated by a thick border line. At the beginning, there are only three GMMs: one for pause (not shown for clarity) and two for each speaker gender. They are trained in advance from some labeled data. For the first segment, the speaker gender is identified (male in the figure) and a new GMM is created from the male GMM. It is learned online with the segment's data, and from this point it becomes the GMM for Speaker 1 (SP1 in the figure). The next segment is from the same speaker, so the SP1 GMM will be the winner. It is again learned on-line with the second segment's data. The third segment comes from a female speaker and the same procedure is repeated resulting in a set of two speaker GMMs. This way, the system generates a set of speaker models on the fly. If some GMM (SP1 in the figure) has not been a winner for a long time, it is deleted from the system (indicated by an "X" on the figure). Such operating mode allows the system to work indefinitely.

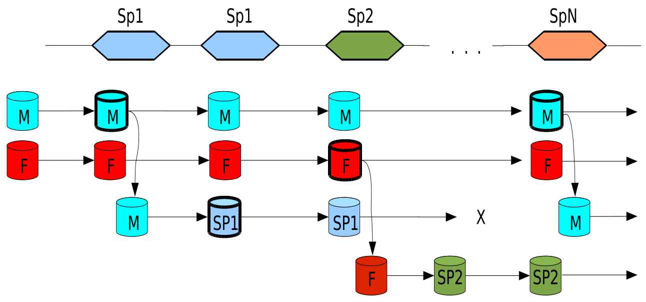

Fig. 3. System operation. For each speech segment, the winning GMM is denoted by bold border lines. The pause GMM is not shown for clarity.

3.2. Novelty detection

The purpose of novelty detection is to decide whether the current segment comes from one of the registered speakers or from a new speaker. This is a typical hypothesis testing problem, where the standard solution is the likelihood ratio test. It is formulated as follows:

$$X \in \left\{ \begin{array}{ll} \omega_0, & \text{if } \ \mathrm{L(X)} > \theta \\ \omega_1, & \text{if } \ \mathrm{L(X)} < \theta \end{array} \right. \tag{2}$$

where $X = \{x_i\}, i = 1, \ldots, DL$ is a decision length speech segment, ω_0 is a class corresponding to the hypothesis H_0, i.e. old speaker. Respectively, ω_1 corresponds to H_1, i.e. new speaker. The likelihood ratio is:

$$L(X) = \frac{p(X|\omega_0)}{p(X|\omega_1)} \tag{3}$$

There are various ways to define $p(X|\omega_i)$. Considering the available set of GMMs, a straightforward approach is to define them as:

$$\begin{array}{lcl} p(X|\omega_0) & = & P_{sp} \ = \max\limits_{\lambda_j \in \Lambda} p(X|\lambda_j) \\ p(X|\omega_1) & = & P_{gen} = \max(p(X|\lambda_{male}), p(X|\lambda_{female})) \end{array} \tag{4}$$

where $\Lambda = \{\lambda_j\}$ is the current set of speaker GMMs. Another approach, often used in speaker verification is to define $p(X|\omega_1)$ as:

$$p(X|\omega_1) = P_{ave} = \frac{1}{n-1}(\sum_j p(X|\lambda_j) - P_{sp}) \tag{5}$$

i.e. the average of all model likelihoods except for the winning model. Here $n = |\Lambda|$ is the size of the speaker set. Experimentally we verified that combining the two approaches works better than either of them. In this case the likelihood ratio is:

$$L(X) = \frac{P_{sp}^2}{P_{gen}P_{ave}} \tag{6}$$

The threshold θ is usually estimated using a development data set.

Although separated in a different module, the speaker identification is implicitly performed during the novelty detection task since the best speaker likelihood is required for the likelihood ratio calculation. The same holds for the gender identification. If the winning hypothesis is H_0, then the best speaker is identified from P_{sp}. Otherwise, the winning gender is found from P_{gen}.

3.3. On-line GMM learning

This technique is the one that allows the whole system to operate on-line and makes it different from all other systems. The main algorithm for off-line GMM parameter estimation is the Expectation-Maximization (EM) algorithm. Not long ago, incremental versions of it were proposed [17, 18], which facilitated the development of on-line variants [19, 20]. In the on-line EM, statistics and parameters are updated after each observation x using the following equations:

$$\begin{array}{l} \ll f(x,y) \gg_i (t) = \ll f(x,y) \gg_i (t-1) + \\ \qquad \eta(t)[f(x(t),y(t))P_i(t) - \ll f(x,y) \gg_i (t-1)] \end{array} \tag{7}$$

where $\ll f(x,y) \gg_i (t)$ is the statistic function of the complete data (x,y). The posterior probability of the Gaussian component i given the previous parameter set Θ_{t-1} is defined as $P_i(t) \doteq P(i|x(t), y(t), \Theta_{t-1})$. The learning rate $\eta(t)$ satisfies the constraints:

$$1 \geq \eta(t) \geq 1/t \tag{8}$$

The new parameters Θ_t are obtained from:

$$\begin{array}{lcl} c_i(t) & = & \ll 1 \gg_i (t) \\ \mu_i(t) & = & \ll x \gg_i (t) / \ll 1 \gg_i (t) \\ \sigma_i^2(t) & = & \ll x^2 \gg_i (t) / \ll 1 \gg_i (t) - \mu_i^2(t) \end{array} \tag{9}$$

The on-line EM converges faster than the standard EM, but even few iterations could increase too much the computational load for a real-time system. On the other hand, given an infinite number of data drawn from the same distribution, the on-line EM can be considered as a stochastic approximation [21]. In practice, this means that as long as there is enough data, model parameters can be approximated in one pass. In this case, the learning rate $\eta(t)$ should satisfy the conditions:

$$\eta(t) \stackrel{t \to \infty}{\longrightarrow} 0, \ \sum_{t=1}^{\infty} \eta(t) = \infty, \ \sum_{t=1}^{\infty} \eta^2(t) < \infty \tag{10}$$

Commonly used function that satisfies these conditions as well as Eq.(8) is:

$$\eta(t) = \frac{1}{at+b} \quad 1 > a > 0 \tag{11}$$

where a and b are parameters which control the learning process. The past samples forgetting speed depends on a, while b sets the learning speed of the new samples.

This algorithm allows fast and inexpensive on-line learning of the system GMMs. As in the batch EM case, the initial parameter values play important role in the learning speed and the precision of the final estimates. Therefore, it is desirable for the initial values to be as close as possible to the true ones. In our system, the gender dependent GMM parameters are the best available initial values for every speaker model and that is why they are used for the new GMM generation.

4. EXPERIMENTS

4.1. Database and pre-processing

For the system evaluation, we used the data released for the TC-STAR 2007 evaluation campaign [22]. The data consists of recordings of the European Parliament plenary speeches. From the training part of the database, we selected about 20 min of silence data for building the pause model. For the gender dependent models, about 2 min. of speech from each of 20 male and 15 female speakers was used. The official development set was used as development data (“dev”), and the evaluation set from the TC-STAR 2006 campaign was used for the final system evaluation (“eval”).

All audio data were transformed into 26 dimensional feature vectors consisting of 12 MFCC coefficients, power and their first derivatives. The frame length and rate were 20 and 10 ms. respectively.

4.2. Voice activity detector performance

We first evaluated the performance of the voice activity detector. The evaluation metric was the speaker diarization error rate (DER) given that all speech segments have correct speaker label. The DER is a time wighted sum of miss errors, false alarms and speaker errors. Since there will be no speaker errors in this setup, the DER will show the VAD performance and it is shown in Table 1 for both the development “dev” and evaluation “eval” data. The minimum segment length for detection was set to 1 or 2 seconds. Bigger values did not improve the results. Typically, a forgiveness collar of 0.25 sec around the reference segment boundaries is set when the DER is calculated. Results with no collar are also presented in the table.

Table 1. VAD performance in terms of DER (%).

Min. segment	Collar = 0.0		Collar = 0.25	
length	dev	eval	dev	eval
1 sec.	4.3	4.5	1.9	2.5
2 sec.	4.5	4.6	2.3	2.5

4.3. Agglomerative clustering system performance

Table 2 shows the speaker diarization error rate (DER) for the agglomerative clustering system when the forgiveness collar around the reference segments boundaries is set to 0.0 or 0.25 sec. The free parameter α is tuned on the “dev” set.

Table 2. DER (%) for the baseline system with α tuned on the “dev” data set.

Collar = 0.0		Collar = 0.25	
dev	eval	dev	eval
10.9	9.5	8.4	7.6

4.4. On-line clustering system performance

For the on-line speaker diarization system, the DER results for both the development and evaluation data are summarized in Table 3. Each row corresponds to the case when the system latency was fixed to 1 to 5 seconds.

Table 3. The full system performance in terms of DER (%).

System	Collar = 0.0		Collar = 0.25	
latency	dev	eval	dev	eval
1 sec.	14.1	21.2	11.5	19.4
2 sec.	9.4	18.8	6.7	16.8
3 sec.	7.2	13.8	4.6	11.9
4 sec.	6.6	13.1	4.0	11.3
5 sec.	6.6	12.1	3.9	10.2

As can be seen, the performance improves rapidly when the system latency, is increased to $3 \sim 4$ sec. and then stays almost the same. The error rates for the evaluation data are about two times higher than the development data, which suggests that the DER is sensitive to the irrecoverable errors inherent in the on-line, one-pass systems. Nevertheless, the overall performance is less than 10%, which is in the range of the best off-line multi-pass speaker diarization systems. As for the processing speed, the system showed real time factor of less than 0.1xRT.

5. CONCLUSIONS

We described two systems for the speaker diarization task: one based on agglomerative clustering and another using on-line clustering approaches. The former has more popular design and is quite accurate, but has several drawbacks, such as off-line operation and high computational cost. The latter works on-line, operates in real-time and thanks to the usage of advanced on-line learning techniques and original design in some cases it performs even better that the off-line system.

6. REFERENCES

[1] S. Tranter and D. Reynolds, “An Overview of Automatic Speaker Diarization Systems,” *IEEE Trans. ASLP*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[2] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, “Improving Speaker Diarization,” in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Nov. 2004.

[3] C. Wooters, J. Fung, B. Peskin, and X. Anguera, “Toward Robust Speaker Segmentation: The ICSI-SRI Fall 2004 Diarization System,” in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Nov. 2004.

[4] P. Nguyen, L. Rigazio, Y. Moh, and J.-C. Junqua, “Rich transcription 2002 site report: Panasonic speech technology laboratory (PSTL),” in *Proc. Rich Transcription Workshop (RT-02)*, 2002.

[5] D. Liu and F. Kubala, “Fast speaker change detection for broadcast news transcription and indexing,” in *Proc. Eurospeech*, Sept. 1999, pp. 1031–1034.

[6] D. Reynolds and P. Torres-Carrasquillo, “The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations,” in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Nov. 2004.

[7] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu, “The LIMSI 2006 TC-STAR transcription system,” in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 123–128.

[8] M. Siegler, U. Jain, B. Raj, and R. Stern, “Automatic segmentation, classification and clustering of broadcast news,” in *Proc. DARPA Speech Recognition Workshop*, Feb. 1997, pp. 97–99.

[9] A. Rosenberg, A. Gorin, Z. Liu, and S. Parthasarathy, “Unsupervised speaker segmentation of telephone conversations,” in *Proc. ICSLP*, Sept. 2002, pp. 565–568.

[10] F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, D. Pineda, D. Seppi, and G. Stemmer, “The ITC-irst transcription systems for the TC-STAR-06 evaluation campaign,” in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 117–122.

[11] S. Stuker, C. Fugen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y.-C. Tam, and M. Wofel, “The ISL TC-STAR Spring 2006 ASR evaluation systems,” in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 139–144.

[12] M. Ben, M. Betser, F. Bimbot, and G. Gravier, “Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs,” in *Proc. ICSLP*, Oct. 2004, pp. 2329–2332.

[13] D. Liu and F. Kubala, “Online Speaker Clustering,” in *Proc. ICASSP*, May 2004, pp. 333–336.

[14] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, Inc., Second edition, 2001.

[15] M. Nishida and Y. Ariki, “Real time speaker indexing based on subspace method - Application to TV news articles and debate,” in *Proc. ICSLP*, Dec. 1998, vol. 4, pp. 1347–1350.

[16] K. Markov and S. Nakamura, “Never-Ending Learning System for On-line Speaker Diarization,” in *Proc. IEEE ASRU Workshop*, Dec. 2007, pp. 699–704.

[17] R. Neal and G. Hinton, “A view of the EM algorithm that justifies incremental, sparse and other variants,” in *Learning in Graphical Models*, M. Jordan, Ed., pp. 355–368. The MIT Press, 1999.

[18] S. Nowlan, *Soft competitive adaptation: Neural Network learning algorithms based on fitting statistical mixtures*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1991.

[19] M. Sato and S. Ishii, “On-line EM algorithm for the Normalized Gaussian Network,” *Neural Computation*, vol. 12, pp. 407–432, 2000.

[20] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne, “On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms,” *Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 275 – 300, May 2004.

[21] H. Kushner and G. Yin, *Stochastic approximation algorithms and applications*, Springer-Verlag, New York, 1997.

[22] TC-STAR, “Technology and Corpora for Speech to Speech Translation,” Online: http://www.tc-star.org/.