SINGLE-CHANNEL SPEAKER-PAIR IDENTIFICATION: A NEW APPROACH BASED ON AUTOMATIC FRAME SELECTION

Ramji Srinivasan, Ji Ming, Danny Crookes

Institute of Electronics, Communications and Information Technology
Queen's University Belfast, Belfast BT3 9DT, United Kingdom

ABSTRACT

Given single-channel recordings of simultaneous speakers, we may need to identify the individual speakers for separating their voices. In this paper, we consider the problem of identifying two simultaneous speakers based on single-channel data, i.e., speaker-pair identification. We model the problem as identifying speakers using noisy speech with partial temporal corruption, which corresponds to the heavily mixed speech frames. Inclusion of these noisy frames will damage the accuracy of both speakers identification. In this paper, we propose a new approach to automatically and optimally select the single-speaker dominated speech frames for identification. The new algorithm has been evaluated using two databases: 1) the GRID speech separation database and 2) the Wall Street Journal (WSJ0) database. The new approach has shown better performance than other approaches. On the Grid database, for example, the new approach outperformed the state of the art IBM approach in 5 out of 6 test conditions.

Index Terms— speaker-pair identification, partial temporal corruption, speaker recognition, speech separation

1. INTRODUCTION

Accurate identification of the constituent speakers from their speech mixtures is a challenging problem. It has received considerable attention in recent years due to its importance in speech separation applications. The need for this study was well emphasized in the Pascal Speech Separation Challenge [1], for separating two simultaneous speakers given their single-channel data. In the challenge task, the speaker-pair identification formed an integral part of many separation systems ([3]–[5]), and was also shown to be one of the deciding factors on the overall performance. In this paper, we propose a new technique for speaker-pair identification from single-channel data of two simultaneous speakers. We aim to improve the identification accuracy over existing approaches.

The above problem has been addressed by using computational auditory scene analysis (CASA), statistical models, and combination. In the CASA based techniques (e.g., [2], [3]), human perceptual cues such as pitch, temporal continuity and harmonic structures are used to identify the constituent speakers in the speech mixture. In the model based approaches (e.g., [4]–[7]), a statistical model such as Gaussian mixture model (GMM) or hidden Markov model (HMM) is used to represent each constituent speaker; given a speech mixture, likelihoods are calculated for the compositions of the constituent speaker models via log-max, Algonquin, Max-VQ, lifted max, non-negative matrix factorization (NMF) or parallel model combination (PMC). For a review of these composition methods, see [6]. The model-based approach in [4] further incorporated a frame selection scheme based on a manually decided likelihood threshold, aiming to ignore the heavily mixed frames for improving the identification accuracy. Studies on this problem are still in active progress. More recently, an attempt was made to improve the performance of speaker-pair identification using double talk detection [7].

In this paper, we propose a new model-based approach for single-channel based identification of two simultaneous speakers, i.e., speaker-pair identification. We consider the problem as one to identify the speakers using noisy speech with partial temporal corruption, which corresponds to the heavily mixed speech frames not matching any single speaker's feature. Inclusion of the heavily corrupted speech frames in calculation will damage the accuracy of both speakers identification. Hershey et al. [4] studied the problem along similar lines and used a manually decided likelihood threshold to select the single-speaker dominated frames for identification. In this paper, we present a new algorithm aiming to automatically and optimally select the frames. We hope that the new automatic algorithm can be more flexible and robust than the manual-threshold based algorithm, for dealing with variable speech features, models and variable acoustic conditions. Experiments conducted on two databases have demonstrated the merit of the new algorithm.

2. SPEAKER AND GAIN MODELING

In this paper, we calculate the log power spectrum for each speech frame, and use a GMM to model the probability distribution of the frame features for each speaker. Denote by G_λ the GMM for speaker λ, which can be expressed as

$$G_\lambda = \{\mathcal{N}(x; \mu_m^\lambda, \Sigma_m^\lambda), w_m^\lambda : m = 1, 2, ..., M^\lambda\} \quad (1)$$

where $\mathcal{N}(x; \mu_m^\lambda, \Sigma_m^\lambda)$ is the m'th Gaussian component, with μ_m^λ and Σ_m^λ being the training-data based mean vector and covariance matrix, and w_m^λ is the corresponding weight, for speaker λ. Given a test speech frame y, the likelihood of y associated with speaker λ is given by:

$$p(y|\lambda) = \sum_{m=1}^{M^\lambda} w_m^\lambda \mathcal{N}(y; \mu_m^\lambda, \Sigma_m^\lambda) \quad (2)$$

In identification, we need to model for each constituent speaker the gain difference between the training and test data. Rewrite the component Gaussian $\mathcal{N}(x; \mu_m^\lambda, \Sigma_m^\lambda)$ in a gain-updated form $\mathcal{N}(x; \mu_m^\lambda + a^\lambda, \Sigma_m^\lambda)$, where a^λ is a gain update value (in dB). Thus,

This work was supported by the UK EPSRC under grant number EP/G001960/1.

the likelihood of a test frame y associated with speaker λ with gain update value a^λ can be expressed as

$$p(y|\lambda, a^\lambda) = \sum_{m=1}^{M^\lambda} w_m^\lambda \mathcal{N}(y; \mu_m^\lambda + a^\lambda, \Sigma_m^\lambda) \quad (3)$$

For each test utterance, we will consider a range of gain update values for each constituent speaker for identification. Specifically, we assume that $a^\lambda \in \mathcal{G}^\lambda$, where $\mathcal{G}^\lambda$ is a predefined gain-update value set, for each speaker λ.

3. AUTOMATIC FRAME SELECTION FOR SPEAKER-PAIR IDENTIFICATION

Let $\mathbf{y}_{1:T} = \{y_t : t = 1, 2, ..., T\}$ be a test utterance, containing T frames with y_t being the frame at time t, composed of two simultaneous speech utterances spoken by two speakers. Assume that we can divide the frames in $\mathbf{y}_{1:T}$ into two classes: those which are dominated by one of the single speakers, and those which are heavily mixed and hence not matching any single speaker's feature. Our objective is to extract the single-speaker dominated frames to perform both speakers identification. The problem can be viewed as identifying the speakers using a noisy speech signal, with unknown, partial temporal corruption. The temporal corruption corresponds to those heavily mixed frames. In the following, we describe an algorithm to automatically extract the single-speaker dominated frames for the identification.

Given the probability distribution (i.e., the GMM) of each constituent speech frame, and given the assumption that the test signal is an additive mixture of two constituent speech signals, there can be several methods, for example, log-max, Algonquin, lifted max, NMF or PMC [6], that can be used to derive the likelihood of a test frame associated with two constituent speakers. However, many of these statistical inference methods are computationally expensive [4]. In this paper, we use a new and simpler method. Specifically, we use the following GMM-based expression to calculate the likelihood of test frame y_t associated with two constituent speakers λ and γ with respective gains a^λ and a^γ

$$\begin{aligned} p(y_t|\lambda, \gamma, a^\lambda, a^\gamma) &\simeq \frac{1}{2} p(y_t|\lambda, a^\lambda) + \frac{1}{2} p(y_t|\gamma, a^\gamma) \\ &= \frac{1}{2} \sum_{m=1}^{M^\lambda} w_m^\lambda \mathcal{N}(y_t; \mu_m^\lambda + a^\lambda, \Sigma_m^\lambda) \\ &\quad + \frac{1}{2} \sum_{m=1}^{M^\gamma} w_m^\gamma \mathcal{N}(y_t; \mu_m^\gamma + a^\gamma, \Sigma_m^\gamma) \end{aligned} \quad (4)$$

When the test frame y_t is dominated by a single speaker, λ or γ, and has the correspondingly correct gain, $p(y_t|\lambda, \gamma, a^\lambda, a^\gamma)$ should be large. Therefore, we can identify the speaker pair by using the frames y_t producing large likelihoods, which are likely to correspond to the single-speaker dominated frames of the two speakers. Denote by $p(y_t^*|\lambda, \gamma, a^\lambda, a^\gamma)$ the test-frame likelihoods sorted in descending order, with y_t^*, $t = 1, 2, ..., T$, corresponding to the test frames from the highest likelihood to the lowest likelihood associated with speaker pair λ, γ and gains a^λ, a^γ. If we know the optimal number of single-speaker dominated frames $\mathcal{T}$ in y_t^*, we can combine these frames to form a likelihood for speaker pair λ, γ. Taking into consideration all possible gain update values, this likelihood of speaker pair based on single-speaker dominated frames can be expressed as

$$p(\mathbf{y}_{1:\mathcal{T}}^*|\lambda, \gamma) = \sum_{a^\lambda \in \mathcal{G}^\lambda, a^\gamma \in \mathcal{G}^\gamma} \prod_{t=1}^{\mathcal{T}} p(y_t^*|\lambda, \gamma, a^\lambda, a^\gamma) P(a^\lambda, a^\gamma) \quad (5)$$

where $\mathbf{y}_{1:\mathcal{T}}^* = \{y_t^* : t = 1, 2, ..., \mathcal{T}\}$ stands for the frame sequence containing the $\mathcal{T}$ highest-likelihood test frames, and $P(a^\lambda, a^\gamma)$ is the prior probability of the gain updates which we assume to be equal for all the value combinations. Without knowledge of the optimal $\mathcal{T}$, we formulate a posterior probability problem to jointly estimate the speaker pair and the optimal $\mathcal{T}$. Let $P(\lambda, \gamma|\mathbf{y}_{1:\mathcal{T}}^*)$ represent the posterior probability of speaker pair λ, γ, as a function of number of test frames with the highest likelihoods. This posterior probability can be expressed as

$$P(\lambda, \gamma|\mathbf{y}_{1:\mathcal{T}}^*) = \frac{p(\mathbf{y}_{1:\mathcal{T}}^*|\lambda, \gamma) P(\lambda, \gamma)}{\sum_{\lambda', \gamma'} p(\mathbf{y}_{1:\mathcal{T}}^{*'}|\lambda', \gamma') P(\lambda', \gamma') + \epsilon} \quad \mathcal{T} = 1, 2, ..., T \quad (6)$$

where $p(\mathbf{y}_{1:\mathcal{T}}^*|\lambda, \gamma)$ is the likelihood function defined in (5), $P(\lambda, \gamma)$ is the prior probability of speaker pair λ, γ (which we assume to be equal for all speaker pairs), and ϵ is a small positive number modeling the likelihood of the test frame sets $\mathbf{y}_{1:\mathcal{T}}^*$ in which there are no matching single-speaker dominated frames and hence the denominator would approach zero. Based on (6), the most-likely speaker pair can be obtained by jointly maximizing $P(\lambda, \gamma|\mathbf{y}_{1:\mathcal{T}}^*)$ over all speaker pairs and all possible numbers of the highest-likelihood frames $\mathcal{T}$, i.e.,

$$\hat{\lambda}, \hat{\gamma}, \hat{\mathcal{T}} = \arg \max_{\lambda, \gamma, \mathcal{T}} P(\lambda, \gamma|\mathbf{y}_{1:\mathcal{T}}^*) \quad (7)$$

where $\hat{\lambda}$, $\hat{\gamma}$ represent the most-likely speaker pair and $\mathbf{y}_{1:\hat{\mathcal{T}}}^*$ is the optimal frame set found for $\hat{\lambda}$, $\hat{\gamma}$. Equation (7) is an automatic approach for speaker-pair identification from unknown partial temporal corruption, which corresponds to those heavily mixed frames without prior knowledge. Similar problems have been discussed previously in *speech* recognition (e.g., [8]). Assume that single-speaker dominated frames produce higher likelihood for the correct speaker pair as against incorrect speaker pairs. We can show that the expression is capable of extracting *all* the single-speaker dominated frames, in terms of the maximum likelihood ratio, for each speaker pair for identification, and hence providing greater discrimination.

We have found that it is helpful to impose a constraint on the minimum value of the optimal frame number $\mathcal{T}$. The constraint reflects a balance between retaining sufficient frames for identification and ignoring noisy frames for robustness.

4. EXPERIMENTAL STUDIES

4.1. Evaluation on the GRID database

First, we used the GRID database [1] to evaluate the new algorithm. The GRID corpus consists of 34 speakers (18 male, 16 female) with 500 training speech utterances for each speaker. Each utterance in GRID has an average length of about 2 s. The test set is formed by mixing two different speech utterances between the 34 speakers at 6 different target-to-masker ratios (TMRs): 6, 3, 0, -3, -6 and -9 dB. The test set is divided into three groups: different gender (DG), same gender(SG) and same talker (ST), with roughly equal numbers of utterances in each group. In our experiments, we used the test utterances from the DG and SG groups for evaluation (the ST

Table 1. Speaker identification accuracy (%) on the GRID database, for different-gender (DG) and same-gender (SG) mixed utterances at variable TMR levels, for the new algorithm compared to the IBM system algorithm and a No-frame-selection algorithm.

TMR	No frame selection			IBM			New		
dB	DG	SG	Avg	DG	SG	Avg	DG	SG	Avg
6	87.5	90.5	88.9	99.0	97.0	98.0	94.0	97.2	95.5
3	91.2	96.1	93.5	99.0	98.0	98.5	98.3	99.2	98.7
0	92.0	96.1	93.9	98.0	98.0	98.0	99.3	99.2	99.2
-3	94.0	95.8	94.8	98.0	97.0	97.5	98.8	98.9	98.8
-6	96.8	96.4	96.5	97.0	97.0	97.0	98.3	98.0	98.2
-9	95.8	94.7	95.2	96.0	96.0	96.0	97.0	96.9	96.9
Avg	92.9	94.9	**93.8**	97.8	97.2	**97.5**	97.6	98.2	**97.8**

Table 2. Speaker identification accuracy (%) on the WSJ0 database, for the new algorithm under constraint $\mathcal{T} \geq T/2$ and $\mathcal{T} \geq T/3$ respectively, compared to the No-frame-selection algorithm.

TMR	No frame selection			New			New		
dB				$\mathcal{T} \geq T/2$			$\mathcal{T} \geq T/3$		
	DG	SG	Avg	DG	SG	Avg	DG	SG	Avg
10	86.9	90.0	88.3	95.3	96.5	95.9	95.0	95.7	95.3
5	93.1	95.7	94.3	99.8	99.7	99.7	99.4	99.3	99.3
0	97.1	97.6	97.4	100	99.7	99.9	99.5	99.7	99.6
-5	97.2	98.8	98.0	99.9	99.6	99.7	99.9	99.4	99.7
-10	96.1	96.2	96.1	97.0	97.2	97.1	96.6	97.4	96.9
Avg	94.0	95.6	**94.8**	98.4	98.5	**98.4**	98.0	98.3	**98.1**

group is neglected as it always yields high speaker identification accuracy rates [6]). As in other approaches, we used a gain update set $\mathcal{G}^{\lambda} = [-9, -6, -3, 0, 3, 6]$, all values in dB, for each speaker λ, to model the variable TMR/gain changes from the training data to test data. The GRID database was used in the PASCAL Speech Separation Challenge [1], and is still being used for the evaluation of single-channel speech separation and speaker-pair identification systems (e.g. [7]). The speech utterances in the database were formed on a small vocabulary (51 words), each utterance obeying a fixed command-sentence grammar.

We divided the speech utterances into frames of 20 ms with a frame period of 10 ms. We then represented each frame in the form of Mel-frequency log filterbank power spectrum. We tested filterbanks of variable numbers of channels, from 26 as typically used in speech analysis for speech recognition, to some higher resolutions up to 128. In general, a higher-resolution power spectrum representation gave improved identification accuracy, but also resulted in higher computational load. For the experiments in this paper, we used a 50-channel filterbank representation, which appeared to provide a good balance. After extracting the frame features for the training utterances, we trained a GMM [i.e., (1)] containing 512 Gaussian components with diagonal covariance matrices, to represent each speaker.

For the experiments on the GRID database, we forced the new algorithm to select at least half of the test frames from each test utterance to perform the speaker-pair identification. That is, $\mathcal{T} \geq T/2$ in the expression (7), where $\mathcal{T}$ is the optimal number and T is the total number of frames in the given test utterance. We compared our new algorithm with the IBM system algorithm [4], which used a manually decided likelihood threshold to choose the optimal frames for identification and produced the best separation results in the speech separation challenge. We also conducted a comparative study with a variant of our new algorithm which uses all the test frames, i.e., $\mathcal{T} \equiv T$. We call this variant algorithm *No frame selection*. Following convention, we present the results in terms of the accuracy for identifying all the speakers in the given test utterances, with each test utterance containing two speakers (target and masker), as a function of the TMR.

Table 1 shows the identification results for the new algorithm compared to the those produced by the IBM system algorithm and the No-frame-selection algorithm. It can be seen that both the new algorithm and the IBM system algorithm outperformed the No-frame-selection algorithm for all the gender groups and TMR conditions. Averaged over all the gender groups and TMR conditions, the IBM system algorithm improved accuracy by absolute 3.7%, and the new algorithm further improved the IBM algorithm accuracy by over absolute 0.3%. The new algorithm, hence, achieved the highest average accuracy of the three algorithms. These results demonstrated the importance of selecting single-speaker dominated frames for speaker pairs identification. Compared to the new algorithm, the IBM system algorithm showed higher DG identification accuracy in the 6 dB and 3 dB TMR conditions. Otherwise, the new algorithm performed consistently better than the IBM algorithm in terms of higher average identification accuracy, through all the remaining TMR conditions. The improvement is more significant for the lower TMR conditions. The new algorithm achieved about 1% absolute improvement in average accuracy over the IBM system algorithm for TMR = 0 dB and lower.

4.2. Evaluation on the WSJ0 database

Then, we used a second database, WSJ0 [9], to further validate the new algorithm. In contrast to the GRID database, the WSJ0 database contains free-text speech utterances formed on a much larger vocabulary (5k words). WSJ0 consists of 101 speakers providing short-term data for speaker-independence training (SI-TR-S). From these speakers, we selected 20 speakers (10 male, 10 female) to construct our experimental data set. Each speaker has about 140 speech utterances, with an average utterance duration of about 7 s. For each speaker, we chose two C-type (Common read no verbal punctuation) utterances to be used to form test utterances, and used the remaining (about 138) utterances for training; the training utterances and test utterances had no sentence texts in common. The two test utterances of each speaker (target) were mixed with the two test utterances of each of the other 19 speakers (masker), first utterance to first utterance, and second utterance to second utterance, at five different TMRs: 10, 5, 0, -5, and -10 dB. Therefore, at each TMR level, there were $20 \times 19 \times 2 = 760$ mixed speech utterances, of which 400 were different-gender (DG) mixtures and 360 were same-gender (SG) mixtures. For this database, we used a gain-update set $\mathcal{G}^{\lambda} = [-12, -9, -6, -3, 0, 3, 6, 9, 12]$ for each speaker λ, to model the variable TMR/gain changes from the training data to test data. As for the GRID database, we calculated the 50-channel, Mel-scale log filterbank power spectrum for each frame and used a 512-component GMM with diagonal covariance matrices to model each speaker.

We conducted two sets of experiments for the new algorithm by using different constraints on the minimum number of frames to be

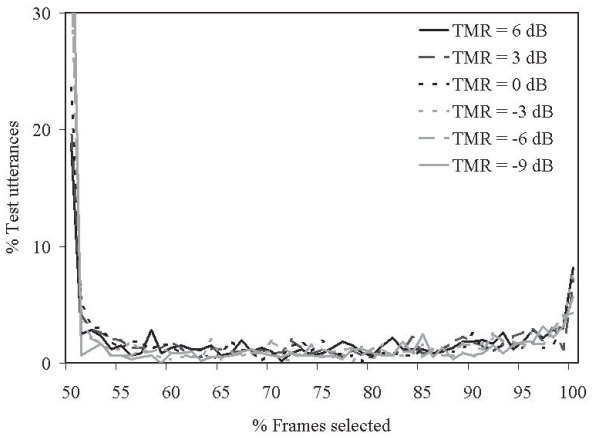

Fig. 1. Histogram of % of frames selected by the new algorithm for identifying speaker pairs on the GRID database, as a function of TMR.

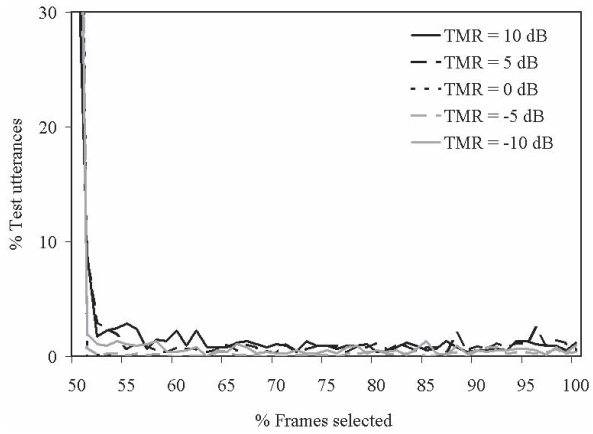

Fig. 2. Histogram of % of frames selected by the new algorithm for identifying speaker pairs on the WSJ0 database, as a function of TMR.

used for identification. First, we assumed that the optimal frame number $\mathcal{T} \geq T/2$, i.e., the algorithm must choose at least half of the frames from each test utterance to identify the speakers. Second, we relaxed the constraint by assuming that $\mathcal{T} \geq T/3$, i.e., we allowed the algorithm to ignore more frames if necessary to seek the maximization of the posterior probability (7). Table 2 shows the results for the new algorithm for each of the constraints in operation, compared to the results produced by the No-frame-selection algorithm with $\mathcal{T} \equiv T$. We see that the new algorithm outperformed the No-frame-selection algorithm under both constraints, for all the gender groups and TMR conditions. Greater than 3% absolute improvement in average accuracy was achieved by the new algorithm through optimal frame selection. The new algorithm performed similarly under the two different constraints, and obtained slightly higher accuracy with the tighter constraint $\mathcal{T} \geq T/2$. Comparing Table 1 and Table 2, we can see that the new algorithm achieved similar average accuracy for identifying speaker pairs on the two different databases. These demonstrate the robustness of the new algorithm for automatically selecting the optimal frames for speaker-pair identification.

Finally, for each test utterance, we counted the number of frames chosen by the new algorithm, expressed as a percentage of the total number of frames. The histograms for the two databases, with the constraint that the optimal frame number $\mathcal{T} \geq T/2$, are shown in Fig. 1 and Fig. 2, as a function of the TMR level. We have noticed some similar frame selection patterns for the new algorithm on the two databases. For heavily mixed utterances, e.g., of a TMR = 0 dB, many frames became unusable, and hence the algorithm tended to select fewer frames from each utterance for identification. The numbers of selected frames were found to be increasing with increasing/decreasing TMR levels. For a significant number of test utterances, the algorithm selected more than half of the frames for identification.

5. CONCLUSION

In this paper, a new algorithm was described for identifying the speaker pairs from single-channel mixed speech with two simultaneous speakers. We aim to find an automatic and optimal algorithm to extract single-speaker dominated frames for the identification. Given a mixed utterance, the new algorithm seeks the most-likely speaker pair by jointly maximizing the posterior probability over all speakers and all possible optimal frames. The new algorithm offers greater flexibility and robustness over existing algorithms which use manually chosen thresholds for frame selection. Two databases, GRID and WSJ0, were used for evaluation. The experimental results have demonstrated improved performance for the new algorithm.

6. REFERENCES

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, "The speech separation and recognition challenge," Computer Speech and Language, vol. 24, pp. 1-15, 2010.

[2] K. Hu, and D. L. Wang, "An approach to sequenctial grouping in co-channel speech," ICASSP'2011, pp. 4636-4640.

[3] J. Barker, N. Ma, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," Computer Speech and Language, vol. 24, pp. 94-111, 2010.

[4] J. R. Hershey, S. J. Rennie,, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: a graphical modeling approach," Computer Speech and Language, vol. 24, pp. 45-66, 2010.

[5] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," Computer Speech and Language, vol. 24, pp. 67-76, 2010.

[6] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," IEEE Signal Processing Magazine, pp. 66-80, November 2010.

[7] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. G. Christensen, S. H. Jensen, and P. Frnti, "Improving monaural speaker identification by double-talk detection," Interspeech'2010, pp.1069-1072.

[8] J. Ming and F. J. Smith, "Speech recognition with unknown partial feature corruption - a review of the union model," Computer Speech and Language, vol. 17, pp. 287-305, 2003.

[9] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," Fifth DARPA Speech and Natural Language Workshop, 1992, pp. 357-362.