INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

Speeding Up Speaker Diarization by Using Prosodic Features

Yan Huang, Gerald Friedland, Christian Müller,

and Nikki Mirghafori

TR-07-004

November 2007

Abstract

In this article we present a method to speed up agglomerative clustering used in speaker diarization by using long-term prosodic features. A set of these features is used to decide which clusters should be merged. This strategy reduces the number of decisions that have to be performed using the more calculation-intensive method based on the Bayesian Information Criterion (BIC). We show a speedup of 30 % to a state-of-the-art diarization system.

This work was partly funded by DTO VACE contract number NBCHC060157. Gerald Friedland and Christian Müller were supported by a fellowship within the postdoc program of the German Academic Exchange Service (DAAD).

Speeding Up Speaker Diarization by Using Prosodic Features

Yan Huang[12],*Gerald Friedland*[1],*Christian Müller*[1], *Nikki Mirghafori*[1]

[1]International Computer Science Institute, Berkeley
[2]Department of Computer Science, University of California, Berkeley
{yan,fractor,cmueller,nikki}@icsi.berkeley.edu

Abstract

In this article we present a method to speed up agglomerative clustering used in speaker diarization by using long-term prosodic features. A set of these features is used to decide which clusters should be merged. This strategy reduces the number of decisions that have to be performed using the more calculation-intensive method based on the Bayesian Information Criterion (BIC). We show a speedup of 30 % to a state-of-the-art diarization system.

Index Terms: speaker diarization, prosodic features

1. Introduction

The goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question “who spoke when?” [1]. State-of-the-art systems use a combination of agglomerative clustering with Bayesian Information Criterion (BIC) [2] and Gaussian Mixture Models (GMMs) of frame-based cepstrum features (MFCCs) [1][3]. While these systems obtain a satisfactory accuracy in terms of the speaker diarization error, this approach exhibits inherent complexity due to the sophisticated online modeling of spectral features leading to very slow processing. However, for most of the applications of speaker diarization, either as a preprocessing step for automatic speech recognition (ASR), for large volume audio retrieving, or for multi-modal meeting understanding, high-speed processing is required.

In this paper, we propose a novel approach to speeding up an existing state-of-the art system, the ICSI speaker diarization system [4][3], by introducing a component exploiting speaker discriminant long-term prosodic features modeled in a comparatively simple and light-weight fashion. Together with other high-level features such as lexical features, long-term prosodic features have attracted increasing interest in the automatic speaker recognition community. Despite the dominance of short-term cepstral features in speaker recognition, a range of high-level features can provide significant information for speaker discrimination, as summarized in [5]. By looking at patterns derived from a larger segment of speech, they reveal individual characteristics of the speakers’ voices as well as their speaking behavior, which can not be captured by frame-based short-term cepstral analysis.

Particularly, we propose a new scoring scheme to compare two clusters using simple single Gaussian modeling of long-term prosodic features and KL-divergence measurement. When profiling the ICSI speaker diarization system, we find that BIC-based merge score calculation takes more than half of the total running time. BIC-score calculation is computationally expensive because each iteration involves BIC merge score computation for all pair-wise merge hypotheses and each of them involves a new round of GMM training. Instead of performing full pair-wise BIC score calculation, we use our new merge score based on long-term prosodic features as a pre-processing step to filter out many highly unlikely merge hypotheses. This strategy is also known as fast match [6].

The rest of this article is organized as following: Section 2 introduces the initial set of candidate long-term prosodic features we explored; Section 3 analyses the runtime bottleneck of most state-of-the-art speaker diarization systems; Section 4 explains the experimental setup; Section 5 presents the results; Section 6 finally summarizes this article and points out future work.

2. Initial Set of Prosodic Features

Our initial set of candidate long-term prosodic features consists of a total of 39 measures that are expected to be speaker discriminative. Each of the variables detailed below is calculated on the basis of the entire speech segment and can therefore be considered a long-term feature. All of them are extracted using the system PRAAT [7] and rendered by simple statistics derivation. Further research in exploring the parameters of extraction and representing the features to better capture the complex patterns is certainly warranted.

Pitch. The speaking fundamental frequency (pitch) is obtained by performing an acoustic periodicity detection using the *accurate autocorrelation method* as described in [8]. On the basis of the resulting pitch track, the following statistical derivatives are calculated: the mean, the median (or 50 % quantile), the 5 % quantile, the 90 % quantile, the difference between the latter two, as well as the standard deviation. The 5 % and 90 % quantiles are used rather than the minimum and the maximum values, respectively, to eliminate the effect of outliers caused by artifacts. The difference value is intended to capture the pitch range of the speaker. However, we are aware that speakers are unlikely to use the entire range of their voices within a single utterance. The standard deviation is used as a simplified measure of the variance of the speaking fundamental frequency.

Formants. The formants F1 to F5 are approximated performing a short-term spectral analysis. Thereby, the waveform is re-sampled to 11 kHz which corresponds to twice the value of maximum formant. Subsequently, a pre-emphasis of +6 dB per octave for frequencies above 50 Hz is performed to flatten the spectrum. The LPC coefficients are finally computed based on a Gaussian-like window applied to each frame using Burg’s algorithm[9]. Similar to pitch, a set of statistics is calculated for each formant: the mean, the 5 %, 50 %, and 90 % quantiles, the standard deviation, and the formant dispersion which is defined as the sum of the differences of subsequent formants. One dispersion value is obtained for each of the respective quantiles, leading to a total of 28 formant features.

Long-term Average Spectrum. The long-term average spectrum (ltas), representing the power spectral density as a function of frequency, is calculated using a bin-bandwidth of 100 Hz. This measure is intended to represent the individual articulatory behavior of the speaker. The resulting vector of db-values for each frequency bin is analyzed according to the following statistics: frequency of the minimum value, frequency of the maximum value, the slope, the local peak height, and the standard deviation.

3. Runtime Analysis of the ICSI Speaker Diarization System

The ICSI speaker diarization system [4][3], similar to other state-of-the-art systems, uses the agglomerative clustering approach followed by Viterbi re-alignment of speaker segments based on frame-based MFCCs. At each iteration, a merge score based on Bayesian Information Criterion (BIC) is calculated between each merge candidate. The score is then used to determine which clusters should be merged at this stage and whether the merge should continue or terminate. Subsequently, a new GMM is trained for the merged cluster and the Viterbi alignment is repeated to re-align the data. The computational load can be decomposed into three components: 1) finding the best merge pair and merge, 2) re-training and re-alignment and 3) everything else. The time break-up for the whole system depicted in Table 1 shows that the BIC score calculation takes 62% of the total running time.

Table 1: *Runtime distribution of the ICSI Speaker Diarization System.*

Component	Runtime
Finding Best Merge Pair and Merge	62 %
Re-training/Re-alignment	28 %
Everything else	10 %
Total	100 %

Analyzing how the best merge hypothesis is found, the reason for the high cost of the score calculation can be identified. Let D_a and D_b represent the data belonging to cluster a and cluster b, which are modeled by θ_a and θ_b, respectively. D represents the data after merge a and b, i.e. $D = D_a \cup D_b$, which is parameterized by θ. Thus, in the case of GMM, θ can be written as $\{(w_i, \mu_i, \Sigma_i)\}$ and for single Gaussian modeling, θ is (μ, Σ). The merge score (MS) is calculated as:

$$MS(\theta_a, \theta_b) = \log p(D|\theta) - (\log p(D_a|\theta_a) + \log p(D_b|\theta_b)) \tag{1}$$

For each merge hypothesis a and b, a new GMM θ needs to be trained and this needs to be performed for all merge hypotheses. When the system is configured to use more initial clusters, which is preferable for better initial cluster purity, the computation load becomes prohibitive.

4. Experimental Setup

As shown in Figure 1, we propose a new merge scoring based on long-term prosodic features (described in Section 2) together with an inexpensive distance measure. These are used as a preprocessing step to filter out many highly unlikely merge hypotheses. Only the promising merge pairs are kept and passed to the BIC merge score calculation. This way, the overall system speed can be substantially increased. This section describes the experimental setup. A Universal Background Model (UBM) is trained in order to obtain a more robust stastical base. We then introduce our basic distance measure, KL-divergence, along with our feature selection process.

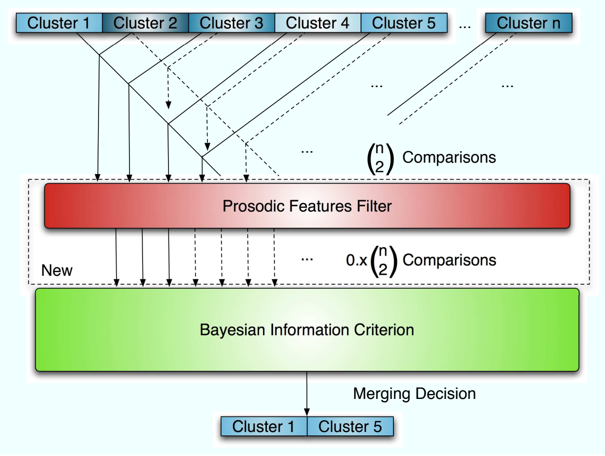

Figure 1: *Diagram of the proposed improved agglomerative clustering step. A prosodic features filter speeds up the entire system by filtering out unlikely merge decisions.*

4.1. Universal Background Model training

To obtain more reliable variances for feature combination and normalization we use a Universal Background Model (UBM) that is trained as a single Gaussian Model to correspond to our online distance measure. Since Meeting Databases typically contain a small number of different speakers, we use TIMIT [10] as our training database. It comprises a total number of 630 speakers from 8 major dialect regions of the United States, each contributing ten sentences. We use the TIMIT training set of 462 speakers, including 136 female speakers and 326 male speakers. For each sentence (treated as a natural speech segment), prosodic features are extracted and used as a training sample for a single Gaussian UBM. One of our objectives of future work is to train the UBM based on a larger dataset of conversational speech. However, the results obtained here indicate that using the TIMIT database for background modeling can be an admissible simplification.

4.2. Feature selection

Another advantage of using a UBM is that it facilitates feature selection. We performed a speaker discriminability study using TIMIT data by measuring the ratio of inter-speaker variance and intra-speaker variance:

$$rank = \frac{\Sigma_{i=1}\Sigma_{j=1}(\mu_i - \mu_j)(\mu_i - \mu_j)^T}{\Sigma_i \Sigma_{j:y_j=i}(x_j - \mu_i)^2}. \tag{2}$$

where y_j is the speaker index for th jth sample.

We assume that features perform better when the ranking of the ratio between inter- and intra-speaker variance is higher. Table 2 shows the results of that ranking. Alternatively, we also use the RankSearch attribute subset selection technique as described in [11]. Interestingly, the RankSearch algorithm, which is based on correlation, comes out with exactly the same subset.

Finally, we take a closer look at the selected feature set, we simply remove the pitch mean because it is highly correlated with the pitch median.

Table 2: *Ranking of the different Prosodic Features according to the ranking discussed in Section 4.2*

Prosodic Feature	Intra-Spk Var	Inter-Spk Var	$\frac{Inter}{Intra}$
Pitch Median	17.0	971.2	57.0
Pitch Mean	89.6	1721.9	19.2
F4 Stddev	7.9	56.8	7.2
F4 Min	12.8	80.4	6.2
Pitch Min	28.3	164.6	5.8
LTAS Stddev	90.6	516.4	5.7
F4 Mean	114.9	649.2	5.6
F5 Mean	180.7	929.0	5.1
F5 Stddev	66.7	327.4	4.9
F5 Min	218.3	1032.5	4.7

This analysis shows that the average pitch is a measure that exhibits a very small intra-speaker variance together with a high inter-speaker variance which is desirable for the task at hand. This finding corresponds to the results of related studies which found pitch to be the single best long-term feature for speaker recognition (see, [12]. The superiority of the median to the mean can be explained by the fact that the median is less sensitive to outliers: artifacts in pitch tracking can lead to a significantly different mean value, but leave the median unaffected. However, the mean and the median pitch are obviously highly correlated, which has to be taken into account when combing the measures. The priority given to F4 and F5 are explained to the lower formants is presumably due to the fact that higher formants capture more speaker-discriminative information whereas lower formants discriminate mainly between different voiced phonemes. The results indicate furthermore that, given the relatively short extraction segments, the minimum pitch (here calculated more reliably by using the 5 % quantile) is a more meaningful measure than the actual difference between the maximum and the minimum.

4.3. Calculating merge score based on KL-divergence

Each cluster is represented as an adapted single Gaussian of long-term prosodic features. KL-divergence is a natural distance measurement for two distributions and in case of single Gaussian it has close-form solution as:

$$KL(\theta_a, \theta_b) = tr(\Sigma_a \Sigma_b^{-1}) - d + tr((\Sigma_a^{-1})(\mu_a - \mu_b)(\mu_a - \mu_b)^T) \quad (3)$$

The symmetric version of KL-divergence is,

$$\begin{aligned} \hat{KL}(\theta_a, \theta_b) = &\quad KL(\theta_a, \theta_b) + KL(\theta_b, \theta_a) \\ = &\quad tr(\Sigma_a \Sigma_b^{-1}) + tr(\Sigma_b \Sigma_a^{-1}) - 2d + \\ &\quad tr((\Sigma_a^{-1} + \Sigma_b^{-1})(\mu_a - \mu_b)(\mu_a - \mu_b)^T) \end{aligned} \quad (4)$$

Since we are using adapted single Gaussian modeling, all Gaussians share the same covariance trained from the UBM described before, i.e. $\Sigma_a = \Sigma_b = \Sigma$ Eq. 4 can be simplified to:

$$\hat{KL}(\theta_a, \theta_b) = 2tr(\Sigma^{-1}(\mu_a - \mu_b)(\mu_a - \mu_b)^T)) \quad (5)$$

Since Σ is trained offline, Σ^{-1} can be calculated offline as well. The new merge score calculation is a simple matrix multiplication.

5. Experiments and Results

Our experiments are performed on the RT06 Meeting Evaluation speaker diarization development set that contains 12 meetings and a total of 4 hours of audio data. Our experiments aimed to answer two questions:

- How much speed up are we able to gain without making the speaker diarization error rate (DER) worse?
- How well do prosodic features discriminate speakers compared to BIC measurements in this particular setup?

5.1. Speedup gained without influencing DER

In order to guarantee that the Prosodic Features Filter will not affect the end-result, we have to make sure that the pair that will be merged according to BIC measurement is not filtered-out. In order to test this we compared the merge decisions of the actual ICSI speaker diarization engine (as described in Section 3) to our filtering decisions for each meeting and each iteration. We found that we can safely filter out 50 % of the comparisons without affecting the system. We rank the possible merge pairs by discarding all the pairs that rank poorly according the KL-divergence merge score. As shown in Table 3, we can expect a speedup of the entire system of about 33 %.

5.2. BIC vs prosodic features

Finally, we compare the rankings of the prosodic filter with the ranking of the BIC measurement. Again, we ranked the possible merge pairs and discarded all those pairs that rank poorly according the KL-divergence merge score. The amount of eliminated possible pairs defines the speed-up. This time, however, we compared against the optimal merge decision according to the ground truth. Additionally, we also ranked the scores generated by the BIC scoring. Accuracy is defined as the average number of iterations (over all meetings) where the ground truth merge pair is not filtered out. This benchmark enables us to get an idea of the quality of the ranking observed by different features and BIC. The results are illustrated in Figure 2. The combination of the top nine features gives better results than BIC ranking. We interpret this as an encouraging sign that in the future we will be able to dispense with the BIC measurement entirely.

Table 3: *Example speed up result for a meeting with 4 speakers. The Section 4.1 for details on the experiment.*

No of Comparisons	Engine Runtime	$x\times$Real time
100 %	100 %	4.2
75 %	83 %	3.4
66 %	78 %	3.2
50 %	67 %	2.8

6. Summary and Future Work

We present a new cluster merge score evaluation using KL-divergence measurement and a set of long-term prosodic features. These include derivatives of $f0$, formants and long-term average spectrum (*ltas*), and apply it to the ICSI speaker diarization system as a pre-processing step to filter out unlikely merge hypotheses. This application largely reduces the amount of heavy-weight BIC merge score calculation and speeds up the system without changing the BIC merge decision.

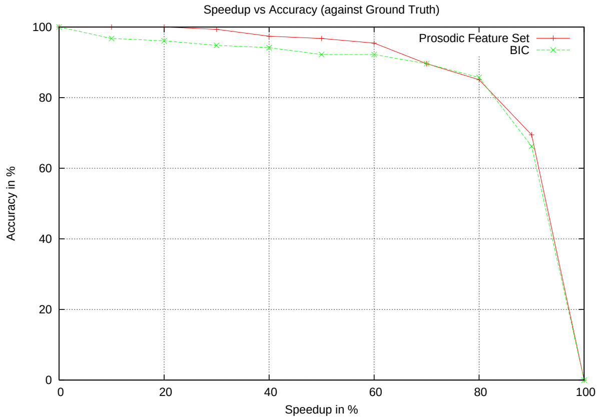

Figure 2: *Speed up vs accuracy for different prosodic features. See Section 4.2 for the details of the experiment.*

This work points into a new direction of combining short-term spectral features and long-term prosodic features for speaker diarization. Short-term frame-based cepstrum features with GMMs modeling are good for speaker change detection (segment re-alignment), where accurate modeling and high time resolution features are needed; long-term prosodic features, which carry rich speaker discriminant information, however, are suitable for cluster modeling and cluster merge decisions. The new merge score approach is very fast compared to the BIC merge score approach based on short-term spectral features. We believe short-term spectral features are still very important in accurately locating speaker change points, but long-term prosodic features can be combined with short-term features for faster speaker clustering.

We are planning on a further exploration of other long-term prosodic features and parameterization. Instead of using our new scoring approach as a pre-processing step of BIC scoring, we would like to try to combine, or ideally replace the BIC decisions.

7. Acknowledgments

The authors would like to thank Nelson Morgan and Chuck Wooters for helpful discussion in this work. We would like to especially address our acknowledgement to Marijn Huijbregts for providing us with the experimental setup. Special thanks go to Mike Ellsworth for proofreading.
This work was (partly) funded by DTO VACE contract number NBCHC060157. Gerald Friedland and Christian Müller were supported by a fellowship within the postdoc program of the German Academic Exchange Service (DAAD).

8. References

[1] D. A. Reynolds and P. Torres-Carrasquillo, “Approaches and applications of audio diarization,” in *Proc. of International Conference on Audio and Speech Signal Processing*, 2005.

[2] S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in *Proc. DARPA speech recognition workshop, 1998.*, 1998.

[3] B. P. X. Anguera, C. Wooters and M. Aguilo, “Robust speaker segmentation for meetings: The icsi-sri spring 2005 diarization system,” in *Proc. of NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.

[4] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *Proc. IEEE Automatic Speech Recognition Underdtanding Workshop*, 2003.

[5] E. Shriberg, “Higher-Level Features in Speaker Recognition,” in *Speaker Classification I*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Heidelberg / Berlin / New York: Springer, 2007, vol. 4343, to appear.

[6] L. Bahl, D. Gopalakrishnan, P.S.and Kanevsky, and D. Nahamoo, “Matrix fast match: a fast method for identifying a short list of candidate words for decoding,” in *Acoustics, Speech, and Signal Processing*, 1989.

[7] P. Boersma, “PRAAT, a system for doing phonetics by computer,” *Glot International*, vol. 9, no. 5, pp. 341–345, 2001.

[8] ——, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proceedings of the Dutch Institute of Phonetic Sciences (IFA)*, Amsterdam, Netherlands, 1993, pp. 97–110.

[9] W. H. Press, S. Teukolsky, and B. Vetterling, W.T. ans Flannery, *Numerical Recipes in C: the art of scientific computing*, 2nd ed. Cambridge University Press, 1992.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren, “The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM,” National Institute of Standards and Technology, Gaithersburg, MD, USA, Tech. Rep., 1993.

[11] I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufman, 2000.

[12] E. Shriberg, S. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, “Modeling Prosodic Feature Sequences for Speaker Recognition,” *Speech Communication*, vol. 46, pp. 455–472, 2005.