FUSING SHORT TERM AND LONG TERM FEATURES FOR IMPROVED SPEAKER DIARIZATION

A. Gerald Friedland, B. Oriol Vinyals, C. Yan Huang

Intern'l Computer Science Institute
1947 Center Street Suite 600,
Berkeley, CA, 94704

*D. Christian Müller**

German Research Center for AI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany

ABSTRACT

The following article shows how a state-of-the-art speaker diarization system can be improved by combining traditional short-term features (MFCCs) with prosodic and other long-term features. First, we present a framework to study the speaker discriminability of 70 different long-term features. Then, we show how the top-ranked long-term features can be combined with short-term features to increase the accuracy of speaker diarization. The results were measured on standardized data sets (NIST RT) and show a consistent improvement of about 30 % relative in diarization error rate compared to the best system presented at the NIST evaluation in 2007. This result was also verified on a wide set of meetings, which we call CombDev, that contains 21 meetings from previous evaluations. Since the prosodic and long-term features were selected using a diarization-independent speaker-discriminability study, we are confident that the same features are able to improve other systems that perform similar tasks

Index Terms— Speaker Diarization, Prosody, Long-Term Features

1. INTRODUCTION

Traditionally, in speech research a small set of standard features, such as MFCC or PLP are used for almost any speech-related task even when problems seem to be orthogonal, such as speech and speaker recognition. The field of speaker diarization is no exception: Current systems usually rely on the combination of Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [1]. In the related field of speaker recognition, however, task-specific features have been successfully applied in combination with MFCCs. These features are often obtained on portions of speech longer than one frame and are therefore referred to as long-term features.

Short-term cepstral features are generally referred to as *low-level* features reflecting the voice parameters of the speaker as opposed to *higher-level* features that capture phonetic, prosodic, and lexical information. In [2], the author summarizes approaches using higher-level information for speaker recognition, a field that is closely related to diarization, and describes them in terms of their type, temporal span, and relevance to the task. It was shown that systems using a combination of cepstral and higher-level features outperformed standard systems, especially when the amount of available training data was increased. This confirms the assumption that short-term cepstral systems generally perform well because they reflect information about the speaker's physiology and do not rely on the phonetic content (which makes them inherently text-independent). However, long-range information that also resides in the signal is only exploited in the combined systems. In addition, as pointed out by [2], higher-level features also have the potential of increased robustness to channel variation, since lexical usage or temporal patterns do not change with the change of acoustic conditions.

Clearly, lexical idiosyncrasies are not investigated here at all. Therefore, rather than using the broader term *higher-level*, we refer to non-cepstral features as *prosodic* and *long-term* features. Following the definition proposed by [2], long-term information refers to features that are extracted over regions longer than a frame. Prosodic features capture variations in intonation, timing, and loudness that are specific to the speaker. Because such features are supra-segmental i.e., extend beyond one segment, they can be considered a subset of *long-term* features. Here, mainly pitch and energy dynamics are investigated. However, [2] itemizes further types of prosodic features such as (explicit) syllable-based prosody sequences, inter-pause/conversation level statistics, and durational features.

2. BASELINE SPEAKER DIARIZATION SYSTEM

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]. While in speaker recognition, models are trained for a specific set of

*The fourth author performed the work while at ICSI

target speakers which are applied to an unknown test speaker for acceptance (target and test speaker match) or rejection (mismatch), in speaker diarization no prior knowledge about the identity or number of the speakers in the recording is given.

The ICSI Speaker Diarization engine extracts MFCC features from a given audio track, discriminates between speech and non-speech regions (speech activity detection), and uses an agglomerative clustering approach to perform both segmentation of the audio track into speaker-homogeneous time segments and the grouping of these segments into speaker-homogeneous clusters in one step.

The audio track is processed as 19th-order MFCC features using a frame size of 30 ms, with a step size of 10 ms. Speech activity regions are determined using a state-of-the-art speech/non-speech detector [3]. The detector performs iterative training and re-segmentation of the audio into three classes: speech, silence, and audible non-speech. To bootstrap the process, an initial segmentation is created with an Hidden Markov Model (HMM) trained on broadcast news data. The non-speech regions are then excluded from the agglomerative clustering, which is explained in the following paragraph.

The algorithm is initialized using k clusters, where k is larger than the (unknown) number of speakers that are assumed to appear in the recording. An initial segmentation is generated by uniformly partitioning the audio into k segments of the same length. On the basis of this segmentation, k Gaussian Mixture Models are trained. As classifications based on 30 ms frames are very noisy, a minimum duration of 2.5 seconds is assumed for each speech segment. The algorithm then performs the following loop:

(1) Re-Segmentation: Run Viterbi Alignment to find the optimal path of frames and models, using a minimum duration constraint of 2.5 s. (2) Re-Training: Given the new segmentation of the audio track, compute new Gaussian Mixture Models for each of the segments. (3) Cluster Merging: Given the new GMMs, try to find the two clusters that most likely represent the same speaker. This is done by computing the BIC score (Bayesian Information Criterion) of each of the clusters and the BIC score of a new GMM trained on the merged segments for two clusters. If the BIC score of the merged GMM is smaller than or equal to the sum of the individual BIC scores, the two models are merged and the algorithm loops at the re-segmentation using the merged GMM. If no pair is found, the algorithm stops.

The Diarization Error Rate (DER) can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker errors (mapped reference is not the same as hypothesized speaker). The ICSI Speaker Diarization System has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art systems. The current official score is 21.74 % DER for the single-microphone case (RT07 evaluation set). This error can be decomposed in 6.8 % speech/non-speech error and 14.9 % speaker clustering error. The total speaker error includes all wrongly classified segments, including overlapped speech.

3. FEATURE SELECTION

The list of initial candidate features can be assigned to five different categories: pitch, energy, formants, harmonics-to-noise ratio, and long-term average spectrum.

The default **pitch** as well as pitch range available to the speaker is influenced by the length and mass of the vocal folds in the larynx [4]. Individual speakers vary in the range of frequencies they are capable of producing as well as the range of frequencies they actually use in everyday speech. Hence, pitch can be regarded as a capable speaker discriminant feature which has been confirmed in numerous speaker recognition studies. From the actual pitch track, various long-range statistics were calculated: mean (the average value); median (the value of the 50th percentile which is generally less sensitive to outliers than the mean); min, max (5th and 95th percentiles); diff (the difference between max and min as a measure of the local range); stddev (the standard deviation as a measure of the variance); swoj (the slope of the pitch curve ignoring octave jumps).

Compared to pitch, changes in **loudness (or energy)** are much less directly induced by anatomical characteristics. Rather than that, they are predominantly relevant to the marking of stress and to express emotions (which is also the case for pitch but to a smaller proportion). Still, energy features are considered as potentially speaker discriminant and therefore used in the candidate list of features for this study. The following statistics have been calculated on the basis of this contour: min, max, diff (the difference between the former two), mean, and stddev.

Formants are concentrations of acoustic energy around particular frequencies at roughly 1000 Hz intervals. Formants occur around frequencies that correspond to the speaker-specific resonances of the vocal tract and are therefore suitable measures to help recognize the speaker. The variation related to the phonetic content happens for the most part in the first two formants while the higher ones are generally assumed to capture mainly speaker-specific information. The following formant-related statistics were used as candidate features (calculated for formants 1—5): mean, median, min (5th percentile), max (95th percentile), and standard deviation. Additionally, a formant dispersion measure was used. It was calculated as the sum of the differences (of min, max, and mean) between consecutive formants. The sampling rate of the NIST meeting data is 16 kHz so we consider frequencies up to 8 kHz.

The **harmonics-to-noise-ratio (HNR)** quantifies the relative amount of additive noise in the voice signal. Spectral noise can be caused by aperiodic vocal fold vibration and tur-

bulent airflow generated by inadequate closure of the vocal folds during phonation [5]. HNR is therefore considered one of the parameters that can be used to quantify a perceptual impression of a rough voice. A fine-grained analysis requires the phone identity to be known. Also, a noisy channel compromises the usefulness of the feature. We nevertheless calculated mean, min, max, diff, and stddev.

In order to obtain the **long-term average spectrum** (LTAS), the spectral energy in 100 Hz-wide frequency bands is measured over a relatively large portion of speech. The standard deviation (stddev) was used as a measure of the variance. In addition, the slope of the curve (slope), the frequency associated with the lowest energy (fmin) and highest energy (fmax), and the peak heights (lph) were calculated.

In total, the list of initial candidate features had 52 elements. To obtain a smaller set of features, we estimated their general speaker discriminability in a pre-experiment using the TIMIT database [6]. We are aware of the limitations of this database in terms of the lack of intersession and interchannel variability. However, these limitations are not crucial for the task of sub-selecting speaker-discriminant features [7]. TIMIT incorporates a large number of speakers (462) which are divided into roughly two-thirds male and one-third female, and each speaker has 10 utterances. Also, the database is reasonably small, which reduced the complexity of the pre-experiment.

For each utterance, one value per feature was obtained (i.e. the range is the entire utterance). The relative speaker discriminability was estimated on the basis of the ratio of the within-speaker variability (w) and the between-speaker variability (b), where $max(\frac{b}{w})$ indicates the best feature in the test. This method is both well-known in machine learning as Fisher discriminant analysis [8] and its validity is also reflected in the final diarization experiments (see Section 4).

The ten features with the highest rank were selected: pitch (mean and median), 4th formant (stddev), pitch (min), 4th formant (min), HNR (mean), 4th formant (mean), 5th formant (mean), ltas (stddev), 5th formants (stddev). Generally, the results appeal to our intuition: The median and average fundamental frequency are the best features, followed by high formants (F4, F5). Also, the mean harmonics-to-noise ratio and the variance of the long-term average spectrum achieved a high rank. Although pitch_median and pitch_mean are likely to be highly correlated, we decided to keep them both since their ranks are outstanding.

4. INTEGRATION INTO ICSI SPEAKER DIARIZATION

We performed a second set of experiments using the ICSI Speaker Diarization system. In these experiments, the speech/non-speech detection remains the same – only the feature input to the agglomerative clustering step is modified.

data set	length	#speakers	#turns	avg l.
CombDev	693 / 3946	2 / 6	8646	3.68s
Eval07	1352 / 2826	2 / 6	5424	3.04s

Table 1. Statistics of the NIST RT meeting evaluation data. Min/Max value for the meeting length (in seconds) and the number of speakers, the number of speaker turns, and the average length per speaker turn (in seconds).

All experiments were performed on the NIST RT07 evaluation data for single distant microphone condition. It contains eight meetings recorded in several geographic locations with differing numbers of people (referred to as NIST Eval07 hereafter). Even though the diarization task is unsupervised, some parameters can be adjusted and optimized, such as the initial number of Gaussians or the weights for the various feature streams. Another set of 21 meetings, based on all NIST RT evaluation and development meeting data of the previous years (excluding the evaluation data Eval07), is used for parameter selection (hereafter referred to as CombDev). Basic statistics of the data sets are shown in Table 1.

In order to combine the new set of features with "traditional" MFCCs, one feature value per frame must be extracted while maintaining a minimum length of the actual extraction region. This is obtained by using a Hamming window of 500 ms and a step size of 10 ms. Another issue is how to deal with missing values. Pitch features, for example, are naturally undefined on unvoiced regions of speech. We applied the most straightforward solution of replacing the undefined values by the mean value of the respective feature calculated over the entire meeting.

The approach we propose for combining several features is similar to the one in [9]. In particular, the function performed by the diarization engine is to maximize the likelihood of the observed data given the model (in our case, the model is an ergodic HMM). We can then define the combined likelihood for the emission probabilities as:
$p(x_{MFCC}, x_{PROS}|\theta_i) = p(x_{MFCC}|\theta_{i1})^{1-\alpha} p(x_{PROS}|\theta_{i2})^{\alpha}$
where x_{MFCC} and x_{PROS} represent the feature vectors (the MFCC vector being 19-dimensional and the prosodic vector being 10-dimensional), θ_{i1} represent the parameters of cluster i using the MFCC features extracted from the observed data and θ_{i2} are the parameters using the prosodic features. The model we use for the emission probabilities are GMMs where the number of components varies for each feature stream. Note that there is an assumption of independence between the two sets of features. Finally, as we observed that MFCC features alone tend to perform better than prosodic features, we used the α parameter to weight the confidence given to each feature stream. If α is set such that $\alpha < 1$, the likelihoods of the prosodic features given each class are flattened (the extreme case where $\alpha = 0$ map all the likelihoods to 1). Hence, the effect of this parameter is to give a different confidence value to each feature stream.

The CombDev set is used to find the optimal value for α. The initial number of Gaussians of the prosodic features is set to 2 and we use the top 10 performing prosodic features. The rest of the parameters are the same as the ones used in the RT07 evaluation (16 initial clusters and 5 Gaussians per cluster for the MFCC feature vector).

First, results on the development were obtained with the optimal value of $\alpha = 0.1$. The use of the top-ten prosodic features resulted in a $24.36\,\%$ relative improvement of the DER (from 17.57 % to 13.29 % DER absolute). Table 2 shows the results using the top-ten features on the Eval07 set, compared to the system that performed best in the NIST Evaluation for the SDM condition (baseline system). The relative improvement is $25.36\,\%$, which is consistent with what was observed on the CombDev data.

We also analyzed the DER evolution per algorithm stage of the baseline system vs. our combined approach. It was to be seen that the top-ten prosodic features contribute especially in the last stages of the agglomerative clustering approach. Since the α value found using the development set was low, the effect of the prosodic features on the first iterations is unnoticed by the algorithm: the MFCCs alone are able to refine the segments and merge clusters that belong to the same speaker. As the clusters are merged, the average length increases and thus the long-term dependencies that the prosodic features extract are more robust. Moreover, in the last stages of the algorithm the clusters are more pure (each cluster contains speech from only one person), and, as a consequence, the discriminative power that the prosodic features have is amplified by the fact that the clusters represent speech from mostly one person.

We also conducted diarization experiments using the top 11–20 ranked prosodic features on CombDev. The resulting DER is 17.29 % absolute, which is almost the same as the baseline system. Compared to the 24.36 % relative improvement generated from using the top 10 ranked features, the use of the top 11–20 ranked features does not give significant improvement over the baseline system, which further verifies our feature selection approach discussed in Section 3.

Meeting ID	Sp/nsp	SpkrSeg	Total DER
CMU_20061115-1030	13.9 %	9.1 %	22.98 %
CMU_20061115-1530	6.7 %	8.6 %	15.25 %
EDI_20061113-1500	10 %	16.5 %	26.43 %
EDI_20061114-1500	6.2 %	14.6 %	20.75 %
NIST_20051104-1515	3.8 %	1.5 %	5.29 %
NIST_20060216-1347	3.3 %	4.1 %	7.43 %
VT_20050408-1500	5 %	2 %	7.05 %
VT_20050425-1000	5.7 %	21.6 %	27.36 %
ALL	6.80 %	9.50 %	16.28 %
ALL (baseline)	6.80 %	15 %	21.81 %

Table 2. DER breakdown for the NIST Eval07 data by using MFCC+prosodic features (baseline is MFCC only). Sp/nsp (speech/non-speech) is the error due to Speech Activity Detection (same system as in baseline) while SpkrSeg is the error due to the speaker segmentation algorithm.

5. CONCLUSION

We presented a systematic investigation of the speaker discriminability of 70 long-term features. We provided additional evidence that despite the dominance of short-term cepstral features in speaker recognition, a number of long-term features can provide significant information for speaker discrimination. Using a combination of the top-ten ranked prosodic and long-term features combined with regular MFCCs we obtained a 30 % relative improvement in terms of the diarization error rate (DER). The results were measured on the NIST RT test and evaluation data sets and were compared to the top-performing system of the NIST RT evaluation in 2007.

6. REFERENCES

[1] DA Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," *Proceedings of ICASSP'05*, vol. 5, pp. 953–956, March 2005.

[2] E. Shriberg, "Higher-Level Features in Speaker Recognition," in *Speaker Classification I*, Christian Müller, Ed., vol. 4343 of *LNAI*. Springer, Heidelberg, 2007.

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

[4] Volker Dellwo, Mark Huckvale, and Michael Ashby, "How is individuality expressed in voice? An introduction to speech production & description for speaker classification," in *Speaker Classification*, Christian Müller, Ed., vol. 4343 of *LNAI*. Springer, Heidelberg - Berlin - New York, 2007.

[5] C. Ferrand, "Harmonics-to-noise ratio: An index of vocal aging," *Journal of Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[6] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallet, and Nancy L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," Tech. Rep. NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, USA, 1993.

[7] Christian Mueller, *Sprecherklassifikation nach Alter und Geschlecht*, Akademische Verlagsgesellschaft Aka GmbH, 2006.

[8] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wileys and Sons, 2nd edition, 2001.

[9] A. Gallardo-Antolin, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, September 2007.