IDIAP RESEARCH REPORT

SPEAKER CHANGE DETECTION WITH PRIVACY-PRESERVING AUDIO CUES

Sree Hari Krishnan Parthasarathi
Mathew Magimai.-Doss Daniel Gatica-Perez
Hervé Bourlard

Idiap-RR-23-2009

AUGUST 2009

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Speaker Change Detection with Privacy-Preserving Audio Cues

Sree Hari Krishnan Parthasarathi[1,2], Mathew Magimai.-Doss[1], Daniel Gatica-Perez[1,2], Hervé Bourlard[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(sparta, mathew, gatica, herve.bourlard)@idiap.ch

ABSTRACT

In this paper we investigate a set of privacy-sensitive audio features for speaker change detection (SCD) in multiparty conversations. These features are based on three different principles: characterizing the excitation source information using linear prediction residual, characterizing subband spectral information shown to contain speaker information, and characterizing the general shape of the spectrum. Experiments show that the performance of the privacy-sensitive features is comparable or better than that of the state-of-the-art full-band spectral-based features, namely, mel frequency cepstral coefficients, which suggests that socially acceptable ways of recording conversations in real-life is feasible.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Human Factors

Keywords

Modeling social interactions, Multiparty conversations, Speaker change detection, Privacy-sensitive features

1. INTRODUCTION

Modeling real-life social interactions using multi-modal sensor data is the central goal of our work. Capturing spontaneous, multiparty conversations, also referred to as personal audio logs, is a step towards this. However, recording and storing raw audio could breach the privacy of people whose consent has not been explicitly obtained. On the other hand, features can be stored instead of raw audio, such that neither intelligible speech nor lexical content can be reconstructed [1]. It is clear that to make progress in ubiquitous analysis of conversations, privacy in addition to computational complexity and performance, needs to be factored in the design equation.

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

The analysis of multiparty conversations can typically involve speaker turn taking patterns. For example, speaker change detection (SCD) can be applied to deduce the type of the conversation (e.g., monologues vs discussions), to estimate the amount of floor control in role recognition, and to detect whether a conversation is competitive or cooperative in nature [2].

State-of-the-art SCD systems [3] use short-term spectral-based features such as Mel Frequency Cepstral Coefficients (MFCC) or Linear Predictive Cepstral Coefficients (LPCC). These features tend to model the peaks in the smoothed spectral envelope (formants). However, speech recognition and synthesis studies show that information about the first two formants may be sufficient to reconstruct lexical content or synthesize intelligible speech [4].

In this paper, we present a study on privacy-sensitive features for speaker change detection, an area that has been relatively unexplored in the area of conversation analysis. We focus on extracting privacy-sensitive features by (a) adaptively filtering the short-term spectrum and characterizing the resulting signal as excitation source information [5], (b) characterizing the subband spectrum shown to contain speaker information [8], and (c) characterizing the general shape of the spectrum [7]. Experiments on the standard HUB-4 1997 broadcast evaluation set show that the performance of the three privacy-sensitive features is comparable or better than that of the baseline MFCC features. In addition, we also show that SCD performance is sensitive to linear prediction (LP) order. We emphasize that our intention is not to design the best SCD system, but to investigate and benchmark privacy-sensitive features.

The rest of the paper is organized as follows. The motivation for the selected features is provided in Section 2. The experimental setup comprising of the dataset, SCD system, baseline and privacy-sensitive features, and evaluation measures is described in Section 3. Finally, the results, discussion and conclusions are provided in Sections 4, 5, and 6, respectively.

2. MOTIVATION

State-of-the-art SCD systems [3] use short-term spectral-based features. These features tend to model the peaks in the spectral envelope, which also carry linguistic information. Speech synthesis studies [4] have shown that information about the first two formants are more important to synthesize intelligible speech. Our approach to selecting features that preserve privacy is based on: (a) representations of the excitation source information by adaptively filtering

out the spectral peaks, (b) selectively characterizing regions in spectrum shown to contain speaker information, and (c) representing the general shape of the spectrum.

2.1 Adaptive filtering

Synthesis of an intelligible speech or reconstruction of lexical content can be rendered difficult by adaptively filtering out information about formants. This approach is motivated by the speech production model. The source-filter model of speech production assumes that the excitation source can be considered to be independent of the vocal tract response. Performing a short-term LP analysis of the speech signal estimates these components [9] (a) an all-pole model, representing the vocal tract system (models smooth envelope of short-term spectrum) (b) a residual, representing the excitation source (c) a gain, correlating with the energy of the signal. The vocal tract response estimated by the LP coefficients contains formant information and therefore does not preserve privacy [4]. On the other hand, the excitation source is estimated by inverse filtering the speech signal with the estimated autoregressive (LP) model. Thus the LP residual (LPR) can be considered to be privacy preserving.

Previous works have shown that the LP residual carries speaker information [5], [6]. A key challenge with utilizing LP residual as a feature is to find a suitable representation. One way to represent the LP residual is to estimate its real cepstrum [5]. Other representations of the residual have been explored; [6] uses the residual without any transformation for a dyadic SCD task, and a group delay representation of the residual was explored in [10].

In this study, as done in [5], we use a real cepstral representation of the LP residual. In contrast to [5], where LP residual cepstrum with a fixed LP order of 14 was used as a complimentary feature to standard short-term spectral-based features, we investigate residual cepstrum independently. Furthermore, this study also explores the effect of varying the LP order. This allows us to control the amount of information that is filtered out. A lower LP order fits fewer number of peaks to match the spectrum.

2.2 Characterizing subband information

Previous studies have shown that the spectral subband from 2500 Hz to 3500 Hz, carries speaker specific information [8]. In this study, we also investigate the two neighboring non-overlapping subbands, namely, 1500 Hz - 2500 Hz and 3500 Hz - 4500 Hz, to assess the importance of the subband 2500 Hz to 3500 Hz. The information in these subbands needs to be suitably represented. We investigate two different representations of the subband information: (a) Computing three MFCC coefficients from the subband. (b) Computing the log-energy from a single filter (centroid) on a subband.

The advantage of the MFCC representation over simple subband filterbank energies is that it decorrelates the filterbank energies and makes these suitable for a Gaussian Mixture Model (GMM) with diagonal covariances. Computing the log-energy of a subband yields a simple representation of the subband information that is suitable for modeling with a Gaussian random variable.

2.3 Characterizing spectral shape

Speakers differ from each other in the distribution of spectral energies within their speech [7]. Further, it is known that male and female speakers exhibit different spectral energy distribution. In general, the spectrum of female speakers show a steeper slope than male speakers. Spectral slope (SS) is thus a way to characterize the shape of the spectrum. In our study first cepstral coefficient (c_1) obtained from LP analysis was used as a measure of the spectral slope.

3. EXPERIMENTAL SETUP

The experimental setup was designed to compare the proposed privacy-sensitive features with the baseline MFCC features. In this section, we describe the dataset, SCD system, baseline, and proposed features used to evaluate the features.

3.1 Dataset

The HUB-4 1997 evaluation set was used to test the performance of the proposed features. The HUB-4 database consists of nearly 3 hours of broadcast news data in different acoustic conditions. This data contains a total of 515 speaker changes from a large variety of speakers.

3.2 SCD system

In our experiments we compare the baseline features with the proposed features using a state-of-the-art SCD system proposed in [3]. A brief summary of this SCD system is provided below.

Speaker change at a time t in an analysis window is hypothesized by modeling each of the two test subsegments by using a single Gaussian density with the same number of parameters, and by modeling the entire segment with a single GMM. The GMM is modeled with diagonal covariance.

Two neighboring windows are compared using a dissimilarity function based on simplified Bayesian Information Criterion (BIC). This function is computed as the difference between the sum of the log likelihood values obtained from subsegment models and the log likelihood value from the single GMM. A peak value of the distance metric in regions greater than 0, is hypothesized as a speaker change point. Furthermore, it was shown in [3] that using the simplified BIC criterion avoids the selection of the threshold used in BIC. It is to be emphasized that this system is kept constant while experimenting with baseline and proposed features.

3.3 Baseline features

The baseline features from [3] are 12-dimensional MFCC feature vectors extracted every 10 ms, using a hamming window of size 30 ms. Similar to the previous work [3], delta and acceleration features are not used. These baseline features are used with the SCD system described in Section 3.2.

3.4 Proposed features

The speech signal is first pre-emphasized, and then analyzed with a hamming window of length and shift 30 ms and 10 ms, respectively. The effect of the LP order was investigated by varying the LP order from 4 to 14. A 16th order real cepstrum of the LP residual was estimated. The choice of the cepstral order was based on previous work [5]. The first cepstral coefficient (c_1) obtained from a 12th order LP analysis was used as a measure of the spectral slope. Three dimensional MFCC feature and log-energy representations of three different subbands, namely, 1500 Hz - 2500 Hz, 2500 Hz - 3500 Hz, and 3500 Hz - 4500 Hz were investigated. The proposed features (up to 21 dimensions) form

the input to the SCD system described in Section 3.2.

3.5 Evaluation measure

The performance of an SCD system is evaluated based on the two types of errors. A Type-I error is said to occur if the system does not detect a speaker change point within a window. We have used the same size of window as done in [3], i.e., a window of size 1 second. A Type-II error occurs when a speaker change point is detected but it does not exist in the reference. The Type I and II errors are also evaluated as precision (P) and recall (R) respectively. These are defined as:

$$\mathrm{P} = \frac{\text{number of changes found correctly}}{\text{total number of changes found}} \cdot 100 \quad (\%) \quad (1)$$

$$\mathrm{R} = \frac{\text{number of changes found correctly}}{\text{total number of changes}} \cdot 100 \quad (\%) \quad (2)$$

In order to compare the performance of different systems, the F-measure is used and is defined as

$$\mathrm{F} = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \quad (\%) \quad (3)$$

A higher F-measure indicates a better performance.

4. RESULTS

The results of all the experiments on the privacy-sensitive features and the baseline MFCC features are reported in Tables 1, 2, and 3 on the HUB-4 1997 evaluation set using precision (P), recall (R), and F-measure (F). In the discussion that follows, LPR-x denotes the 16th order real cepstrum of the residual of LP order x, SS denotes the spectral slope estimated using cepstral coefficient (c_1), MFCC(a - b) denotes the subband MFCC coefficients from a kHz to b kHz, and E(a - b) denotes the subband log-energy value from a kHz to b kHz. The findings of the study are summarized as follows.

4.1 Performance of privacy-sensitive features

Table 1 compares the performance of the privacy-sensitive features with baseline full-band MFCC features. It can be observed that adding either spectral slope or the subband MFCC to the LP residual cepstrum increases the performance (F-measure). We note that combining spectral slope with LP residual features yields a performance as good as the baseline MFCC features. Combining all the three privacy-sensitive features gives a slight improvement over the baseline MFCC features. It is interesting to note that the SCD system which models the features using Gaussian distributions is suitable for the proposed features as well.

Table 1 shows that baseline MFCC features provide a balance between precision and recall. On the other hand, using residual features by itself yields a higher amount of recall at a lower precision. The addition of subband MFCC to LP residual increases the recall at the same level of precision. Whereas, combining spectral slope with residual features increases the precision. Finally, we observe that combining all the three features results in a more balanced segmentation.

4.2 Representing subband information

In this section, we investigate (a) the optimal subband, and (b) a representation of subband information for SCD. Table 2 shows the performance of three non-overlapping frequency bands represented with MFCC and log-energy values. We note that with either subband MFCC values or with subband log-energies, the suband 2500 Hz to 3500 Hz yields the best performance. This corroborates with earlier studies [8].

Table 1: ***Complementarity of information in LPR, SS and FB: LPR-x denotes the real cepstrum of LP residual of order x, SS denotes spectral slope, and MFCC(a - b) denotes MFCC values from a kHz to b kHz. The best performance by MFCC baseline is highlighted in bold and italics while the best performances by privacy-sensitive features are highlighted in bold. The dimensions of the 4 feature vectors are 12, 17, 18, 20 and 21 respectively.***

Features	P (%)	R (%)	F (%)
MFCC (Baseline)	63.00	64.47	***63.72***
LPR-4	57.98	67.38	61.31
LPR-4 + SS	67.60	60.78	64.00
LPR-4 + MFCC (2.5 - 3.5)	57.14	69.13	62.57
LPR-4 + SS + MFCC (2.5 - 3.5)	66.60	63.50	**65.01**

Further, we note that the ranking of the three subbands in terms of performance is the same for both subband MFCC and subband log-energy representations. The table also reveals that the subband MFCC representation is a better representation than the subband log-energy representation. In fact, from Tables 1 and 2 it can be observed that the addition of log-energy value brings down the performance.

Table 2: ***Representing subband information: LPR-x denotes the real cepstrum of the LP residual of order x, SS denotes spectral slope, MFCC(a - b) denotes MFCC values from a kHz to b kHz, and E(a - b) denotes log-energy values from a kHz to b kHz. The first 3 feature vectors have a dimensionality of 21 while the next 3 have a dimensionality of 19.***

Features	P (%)	R (%)	F (%)
Representing subband information with MFCC			
LPR-4 + SS + MFCC (1.5 - 2.5)	65.68	60.58	63.02
LPR-4 + SS + MFCC (2.5 - 3.5)	66.60	63.50	**65.01**
LPR-4 + SS + MFCC (3.5 - 4.5)	65.19	60.00	62.48
Representing subband information with log-energy			
LPR-4 + SS + E (1.5 - 2.5)	62.27	58.64	60.40
LPR-4 + SS + E (2.5 - 3.5)	62.23	61.75	**61.99**
LPR-4 + SS + E (3.5 - 4.5)	59.43	61.17	60.29

4.3 Effect of LP order

In this section we present our investigation on the effect of increasing the LP order. From Table 3, it can be observed that increasing the LP order leads to a decrease in the performance up to a prediction order of 10.

We note that an increase in LP order by 2, can allow an extra complex conjugate pole pair to be modeled, possibly modeling an extra formant. Since higher order formants in general, carry more information about speakers, we can expect the performance to drop when LP order is increased.

On the other hand, increasing the LP order beyond 10, results in an increase in the performance. To explain this, we note that the LP residual contains both modeling and excitation errors. As the LP order increases beyond 10,

Table 3: ***Effect of LP order in LPR: LPR-x denotes the real cepstrum of the LP residual of order x, SS denotes spectral slope, and MFCC(a - b) denotes MFCC values from a kHz to b kHz. All feature vectors have a dimensionality of 21.***

Features	P (%)	R (%)	F (%)
Even linear prediction order			
LPR-4 + SS + MFCC (2.5 - 3.5)	66.60	63.50	**65.01**
LPR-6 + SS + MFCC (2.5 - 3.5)	63.62	58.06	60.71
LPR-8 + SS + MFCC (2.5 - 3.5)	63.41	55.53	59.21
LPR-10 + SS + MFCC (2.5 - 3.5)	60.84	50.68	55.30
LPR-12 + SS + MFCC (2.5 - 3.5)	61.47	52.04	56.36
LPR-14 + SS + MFCC (2.5 - 3.5)	59.91	54.56	57.10
Odd linear prediction order			
LPR-5 + SS + MFCC (2.5 - 3.5)	65.39	63.11	**64.23**
LPR-7 + SS + MFCC (2.5 - 3.5)	64.59	56.31	60.17
LPR-9 + SS + MFCC (2.5 - 3.5)	62.01	52.62	56.93

the contribution of the modeling error in the residual signal decreases while the contribution of the excitation error remains constant. In this case, the residual can be likened to modeling the excitation source, which contain speaker information [11]. Experiments performed with LP order approaching 40, showed performance saturating around 60%.

In comparison with an increase in LP order by 2, an increase LP order by 1 does not lead to a big drop in performance. For example increasing LP order from 4 to 5 leads to a drop of only 0.78%. An LP order of 4 can model up to one complex conjugate pole pair, whereas an LP order of 5 can model an extra real pole. Therefore, the performance does not drop much when the LP order is increased from 4 to 5.

5. DISCUSSION

In this paper, we investigated features for the privacy-sensitive SCD task. Wyatt et al [13], on the other hand, approach the privacy-sensitive speaker segmentation task by extracting simpler features, and focusing on extensive modeling.

While this paper utilizes the real cepstral representation of the LP residual, a number of other representations are possible. In [12], a mel cepstrum representation of the LP residual cepstrum was utilized as a complimentary feature to LPCC for speaker identification. On the other hand, Dhananjaya et al [6] used the LP residual directly (without any transformation) for a dyadic SCD task. There may be other feasible representations such as group delay function of the residual signal [10], or features based on glottal flow [11].

We note that the cepstral order of the residual was fixed at 16. However, it would be reasonable to expect the cepstral order to be inversely related to the LP order. For instance, a higher LP order tends to model more formants. Consequently, fewer cepstral coefficients may be sufficient when a high LP order is used.

We investigated two different representations of subband information. Subband MFCC representation yielded better results. However, alternate representations such as spectral linear prediction can possibly be used to characterize the subband [14].

6. CONCLUSIONS

In this paper we investigated a set of privacy-sensitive features for SCD. These features are linear prediction residual cepstrum, subband MFCC with a bank of 4 filters, and spectral slope. Using F-measure as a evaluation measure on the HUB-4 1997 evaluation set, experiments showed that the performance of the proposed privacy-sensitive features is comparable or better than that of the state-of-the-art full-band spectral-based MFCC features. In addition, it was shown that SCD performance was sensitive to LP order. Overall, our study suggests that privacy-preserving approaches, clearly needed for ethical recording of real conversations in the wild, are feasible and competitive.

7. ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation through the projects MULTImodal Interaction and MULTImedia Data Mining (MULTI2) and the National Centres of Competences in Research (NCCR) IM2. The authors would like to thank Dr. Jitendra Ajmera (Toshiba Corporate Research and Development Center) for his help and support.

8. REFERENCES

[1] D. Wyatt, T. Choudhury and H. Kautz. Capturing spontaneous conversation and social dynamics: a privacy sensitive data collection effort. In *Proc. of ICASSP*, 2007.

[2] D. Gatica-Perez. Analyzing Group Interaction in Conversations: a Review. In *Proc. of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006.

[3] J. Ajmera, I. McCowan and H. Bourlard. Robust Speaker Change Detection. *IEEE Signal Processing Letters*, 2004.

[4] R. Donovan. Trainable speech synthesis. *PhD Dissertation, Cambridge University*, 1996.

[5] P. Thevenaz and H. Hugli. Usefulness of the LPC- residue in text-independent speaker verification. *Speech Communication*, pages 145–157, 1995.

[6] N. Dhananjaya and B. Yegnanarayana. Speaker change detection in casual conversations using excitation source features. *Speech Communication*, pages 153–161, 2007.

[7] F. K. Soong and A. K. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. on Acoustics Speech and Signal Processing*, pages 871–879, 1988.

[8] S. Furui. Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, 1986.

[9] J. Makhoul. Linear prediction: A tutorial review. *Proc. of the IEEE*, 1975.

[10] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. on Speech and Audio Processing*, pages 325–333, 1995.

[11] M. D. Plumpe and T. F. Quatieri and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, pages 569–586, 1999.

[12] J. He et al. On the use of features from prediction residual signals in speaker identification. In *Proc. of Eurospeech*, 1995.

[13] D. Wyatt et al. A Privacy-sensitive approach to modeling multi-person conversations. In *Proc. of IJCAI*, 2007.

[14] J. Makhoul. Spectral linear prediction: properties and applications. *IEEE Trans. on Acoustics Speech and Signal Processing*, pages 283–296, 1975.