LP Residual Features for Robust, Privacy-Sensitive Speaker Diarization

Sree Hari Krishnan Parthasarathi[1,2], Hervé Bourlard[1,2], Daniel Gatica-Perez[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Switzerland
{sparta,bourlard,gatica}@idiap.ch

Abstract

We present a comprehensive study of linear prediction residual for speaker diarization on single and multiple distant microphone conditions in privacy-sensitive settings, a requirement to analyze a wide range of spontaneous conversations. Two representations of the residual are compared, namely real-cepstrum and MFCC, with the latter performing better. Experiments on RT06eval show that residual with subband information from 2.5 kHz to 3.5 kHz and spectral slope yields a performance close to traditional MFCC features. As a way to objectively evaluate privacy in terms of linguistic information, we perform phoneme recognition. Residual features yield low phoneme accuracies compared to traditional MFCC features.

Index Terms: LP residual, privacy-sensitive, diarization

1. Introduction

This work takes place in the context of analyzing a wide range of spontaneous audio captured by a portable recorder. However, one of the biggest obstacles facing this field concerns privacy: recording and storing raw audio would breach the privacy of people whose consent has not been obtained. Wyatt et al [1] suggest that the linguistic message in the signal is the privacy-sensitive information. One approach, therefore, is to store features instead of raw audio such that neither an *intelligible speech* signal nor the *lexical content* can be reconstructed [1]. These features are called as *privacy-sensitive features* in this paper.

Another approach is to implement an online speaker diarization system on the portable device and store information derived from its output. A caveat is that the set of possible tasks is then limited by the output of the diarization system. Furthermore, this design is constrained by the computational limitations of the device.

We consider the former approach, namely, storing privacy-sensitive features. Analysis of social interactions can then proceed (but not limited to) by modeling the speaker turns. For instance, speaker turns from an offline diarization system using these features can be employed to deduce the type of conversation, recognize roles, and to detect dominance [2].

A further constraint is the necessity of features to be robust to audio captured by portable, single distant microphones. In this setting, this paper focuses on robust features having low linguistic information for diarization (*who spoke when*), an area that is relatively unexplored in the field of conversation analysis.

State-of-the-art diarization systems [3] use features such as Mel Frequency Cepstral Coefficients (MFCC). While MFCC have been shown to be robust for diarization, it is possible to reconstruct a highly intelligible speech signal and perform state-of-the-art automatic speech recognition (ASR) from MFCC. Previous approaches to privacy-sensitive features have mostly focused on reinterpreting simple, frame-level heuristics ([1, 4]) or computing long-term averages of standard features [5]. However these methods were not proposed for speaker diarization, a choice further supported by our preliminary experiments.

Drawing motivation from the source-filter model, our approach to privacy is based on adaptively filtering formants required to synthesize intelligible speech [6]. To this end, we present linear prediction (LP) residual as a privacy-sensitive feature. Two different representations of the residual signal are analyzed: real-cepstrum and MFCC. Combination of residual with subband information and spectral slope is studied. We then present a systematic investigation of residual for diarization in single and multiple distant microphone (SDM, MDM) conditions. Experiments using the ICSI system [3] on the NIST RT06 evaluation [7], show that the proposed features yield a performance close to MFCC features.

To the best of our knowledge, benchmarking audio features for privacy has not been studied before, and it remains something that is difficult to quantify. To this end, we present phoneme recognition as one way to quantify privacy, with higher accuracy interpreted as lower privacy. We show that proposed features yield lower phoneme recognition accuracies than MFCC features. Furthermore, informal experiments suggest that synthesizing speech from MFCC representation of 8^{th} order residual is not sound intelligible. We also explore obfuscation methods such as local temporal randomization (within 130 ms) of the features as a means to provide stronger privacy.

In the next section, we analyze residual using mutual information (MI). The diarization system is described in Section 3. Diarization results are provided in Section 4 before revisiting privacy in Section 5. Conclusions are drawn in Section 6.

2. Analysis of residual features

We begin with a motivation for residual before proceeding to an MI analysis of the two issues concerning feature extraction from residual signal: choice of representation and prediction order.

It is generally known that up to three formants are required to synthesize intelligible speech or reconstruct lexical information [6]. Our approach to privacy, motivated by source-filter model of speech production, is based on adaptively filtering the spectral peaks. LP analysis of speech assumes the source-filter model and it estimates 3 components (a) an all-pole model, representing the vocal tract (b) a residual, representing the excitation source of the speaker (c) a gain, correlating with the energy of the signal. Residual can therefore be considered to be privacy-sensitive.

We now take a detour to interpret privacy-sensitive features for diarization as maximizing the MI with speakers while minimizing the MI with phonemes. A feature having higher MI with phonemes could be considered as worse in terms of pri-

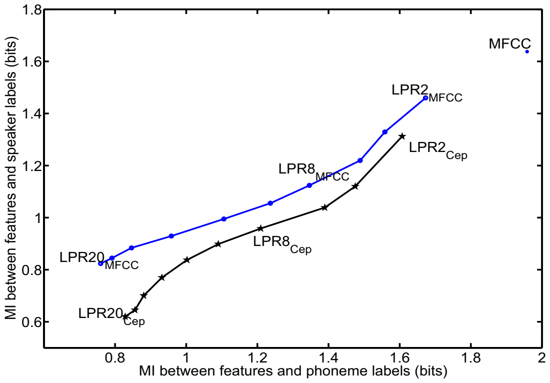

Figure 1: Plot showing MI between residual features and phonemes vs MI between residual features and speakers. LPRx denotes residual with prediction order x. The subscript denotes real-cepstrum or MFCC. Lone point denotes baseline MFCC.

vacy, while a feature having higher MI with speakers could be considered as better for diarization. Formally, given X, a random variable denoting the short-term power spectrum of the signal, and S, Q random variables, denoting the speaker and phoneme labels respectively, we seek a transformation g that optimizes $I(g(X); S) - I(g(X); Q)$. We view residual features as a transformation $(g_{p,\theta}(.))$ parametrized by prediction order (p) and choice of representation θ.

Experiments were performed on the TIMIT train set (3696 utterances with 462 speakers and 39 phonemes) to analyze the effect of p and θ. MI is estimated on the discretized feature space (using k-means algorithm) and then using Miller's formula to correct the bias [8]. The MI analysis in this section is subsequently validated in Sections 4, 5.

2.1. Representing the residual signal

We study two representations (θ) of the residual signal obtained from short-term LP analysis: real-cepstrum as in [9] and MFCC of residual signal. The representations have an advantage that the dimensions are decorrelated, facilitating GMM modeling with diagonal covariances. We chose 19 dimensions for these representations to match the baseline MFCC feature dimension.

Figure 1 shows a comparison between the two representations by plotting MI with phonemes versus MI with speakers on TIMIT train set. The baseline MFCC feature in the plot yields the highest MI with phoneme and the speaker labels.

As prediction order increases, MI with phonemes and speakers decrease for both representations. But the MFCC representation yields higher MI with both speaker and phoneme labels for all values of the prediction order. As the prediction order increases from 2 to 20, the real-cepstrum yields a bigger fall in MI with speakers than the MFCC representation (0.7 bits versus 0.6 bits). On the other hand, MI with phonemes shows a reverse trend: i.e., as the prediction order increases from 2 to 20, the real-cepstrum yields a fall of 0.75 bits, while the MFCC representation yields a fall of 0.9 bits. Based on these observations we choose the MFCC representation over the real-cepstrum.

2.2. Analysis of LP order

From Figure 1, we notice that the drop in MI with phonemes for the MFCC representation is as much for an increase in prediction order from 2 to 8 (or 10), as it is for an increase in prediction order from 10 to 20. A similar trend is observed for the MI with speakers for the MFCC representation.

A prediction order of 8 seems appropriate since the first two formants are important for synthesizing intelligible speech [6]. Although the 8^{th} order residual signal can be intelligible, informal experiments synthesizing speech from the MFCC representation of an 8^{th} order residual does not sound intelligible. This is due to a further loss in information by representing the residual using MFCC.

3. Diarization system

This section discusses the baseline system, features, datasets and the performance measure used to evaluate the features.

3.1. Baseline system

The baseline ICSI system is based on ergodic HMM with the emission probabilities being modeled by GMM. The algorithm follows an agglomerative framework with a minimum duration constraint of 3 sec. After each merge, data are re-aligned using a Viterbi algorithm. Initial HMM is built using uniform segmentation. Cluster merges are compared using modified Bayesian Information Criterion (BIC) [3].

This system uses 19 dimensional MFCC features with time delay of arrival (TDOA) features. MFCC is extracted every 10 ms, with a hamming window of size 30 ms using HTK. Delta and acceleration features are not used. In our experiments, the TDOA features are not used.

3.2. Privacy-sensitive features

The proposed features are residual features in combination with the subband frequency information between 2.5 kHz to 3.5 kHz (SB) and the spectral slope (SS).

The speech signal is pre-emphasized and then analyzed with a hamming window of length and shift 30 ms and 10 ms, respectively. The residual signal obtained from an 8^{th} order predictor is then represented using 19 dimensional MFCC features. This is based on the analysis in Section 2.

Previous studies have shown that the spectral subband from 2500 Hz to 3500 Hz carries speaker information [10]. We compute three MFCC coefficients from this subband. Speakers differ in the distribution of spectral energies [11]. For instance, the spectrum of female speakers show a steeper slope than male speakers. Spectral slope (SS) is a way to characterize this. We use the first cepstral coefficient (c_1) obtained from LP analysis as a measure of SS.

Obfuscation methods have been used previously in other aspects of privacy in sensor data research [12]. Here, obfuscation is achieved through shuffling feature vectors within non-overlapping blocks of frames ($N = 5, 9, 13$). A uniform random number generator is used for this.

3.3. RT06 evaluation dataset

Experiments were performed on NIST RT06 evaluation data for meeting recognition diarization task [7]. It contains nine meeting recordings of approximately 30 minutes each. For the MDM dataset, individual channels were Wiener filtered and beamformed using the BeamformIt toolkit [3]. SDM experiments were performed on individual MDM channels yielding the worst performance. Speech/nonspeech detection (SND) was obtained using forced alignment of the reference transcripts on close talking microphone data. The same SND is used across all experiments since our interest is in evaluating the features for speaker segmentation.

The results are usually reported in terms of Diarization Error Rates (DER), which is the sum of SND errors and speaker errors. SND errors is the sum of missed speech and false alarm. Since our objective is to compare features for speaker segmentation, we focus on the speaker errors.

4. Diarization results on RT06eval

The diarization results of the proposed privacy-sensitive and the baseline MFCC features on RT06 evaluation data (MDM and SDM) are reported.

4.1. Baseline system

Table 1 lists the performance of the baseline diarization system. The first 3 columns list the performance of the SND system in terms of missed speech, false alarm, and the overall SND error. We note that the overall SND error rate over all the files on the RT06 evaluation dataset is 6.6%.

The next two columns list the performance of the baseline MFCC features in terms of the speaker error for both the MDM and the SDM scenarios. As expected, MFCC features perform better on the beamformed MDM data with a difference of 3.7%.

Table 1: *Performance of the baseline MFCC.*

Feature	Miss	FA	sp/nsp	Spkr (%) MDM	Spkr (%) SDM
MFCC	6.5	0.1	6.6	**17.1**	**20.8**

4.2. Performance of privacy-sensitive features

Table 2 compares the speaker errors of the proposed features against the baseline MFCC features.

It can be observed that the baseline MFCC features yield the best speaker errors for MDM and SDM conditions. In MDM condition, the speaker error of 8^{th} order LP residual using MFCC representation (denoted by LPR-8) is about 5% below the baseline. In SDM condition the difference with the MFCC features is more (around 9%).

Adding either spectral slope or subband information between 2.5 kHz to 3.5 kHz to LPR-8 improves the performance in both conditions, with improvement being more significant in the SDM case. Combining both spectral slope and subband information with LPR-8 features yields a bigger improvement than combining with either of them, with SDM condition benefiting more. In both conditions, the combined system compares reasonably with the baseline MFCC features, with a difference of 2%. Combination weights were not tuned: i.e., when residual is combined with SB or SS (or both), residual is assigned 50% of the weight.

As a reference, the simple privacy-sensitive features proposed in [1] and [4] yielded performances nearly 30% below the baseline MFCC features. Diarization experiments using 8^{th} order residual on MDM data confirm MI studies that the MFCC representation yields nearly 9% lower errors than cepstrum.

Table 2: *Speaker errors of the features on SDM and MDM.* 8^{th} *order residual is denoted by LPR-8, subband information by SB, and spectral slope by SS. Dimensionality of the features are listed for reference.*

Features	Dimension	Spkr (%) MDM	Spkr (%) SDM
MFCC (baseline)	19	**17.1**	**20.8**
LPR-8	19	22.3	29.2
LPR-8 + SB	22	21.9	26.0
LPR-8 + SS	20	21.8	28.6
LPR-8 + SB + SS	23	**19.1**	**22.2**

4.3. Effect of prediction order on diarization

The effect of LP order on residual on diarization is presented in Table 3. Similar to the MFCC features, the proposed features perform better on the MDM data for all prediction orders.

Results on MDM and SDM data exhibit similar behaviors, which can be analyzed separately in 3 relatively distinct regions: smaller drop in performance for increases in prediction orders from 2 to 6, followed by a more dramatic drop in performance for prediction orders between 8 to 12, and then again a smaller drop afterward.

An increase from 2 to 6 results in a drop of 1.6% in the MDM case. This could be due to the loss of the first formant, which carries more linguistic information [6]. For LP orders between 8 to 12, an increase in the LP order results in a bigger drop in performance. For instance, an increase in LP order from 8 to 10 results in a drop of nearly 6% in MDM and 5% in SDM. It is interesting to note that an LP order of 2 yields better results than the baseline in Table 1. This could be due to 2^{nd} order LP filter removing the slope, boosting higher frequencies. An LP order in the range 8 to 10 can model around 3 formants. Since higher formants carry more speaker information, increasing LP order beyond 8 results in greater speaker errors.

Table 3: *Effect of LP order in MDM and SDM conditions.*

LP order	Spkr err (%) MDM	Spkr err (%) SDM
2	16.3	17.2
4	17.3	19.2
6	17.9	24.1
8	**22.3**	**29.2**
10	28.7	35.2
12	35.1	40.5
14	36.7	47.4
16	44.2	50.1
18	43.1	49.9
20	44.7	52.2

For the last segment (orders greater than 12), we see a smaller drop in the performance as the order is increased. Since residual contains both modeling and excitation errors. Beyond a prediction order of 10, the contribution of the error in the residual is mainly due to the excitation component.

5. Revisiting privacy

This section revisits the issue of assessing privacy. Some methods to evaluate the linguistic notion of privacy are, human speech recognition of speech synthesized from features, and ASR using the privacy-sensitive features. This paper presents phoneme recognition to evaluate privacy.

5.1. Phoneme recognition

The phoneme recognition experiments were performed using hybrid HMM/MLP. A separate 3-layered MLP is trained for each set of features. MLPs consist of 1000 hidden and 39 output units (phoneme classes). The inputs use a temporal context of 9 frames, with delta and acceleration. Each MLP is trained using back propagation algorithm by minimizing the cross entropy criterion. The phoneme sequence is decoded using Viterbi algorithm, where each phoneme is represented by a 3-state HMM. The emission likelihood of the states is same, and is derived from the MLP. Further details can be found in [13].

Figure 2 plots the recognition accuracies with respect to increasing LP orders using this system. It can be observed that

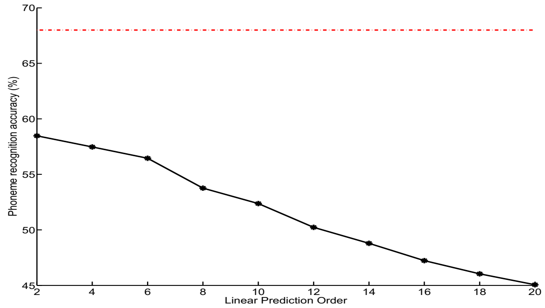

Figure 2: *Phoneme recognition: The x-axis shows the LP order while the y-axis shows the phoneme accuracy in (%).*

as the LP order increases the recognition accuracies drop. An increase in LP order by 2 can allow an extra complex conjugate pole pair to be modeled, possibly modeling an extra formant. Since lower order formants generally carry more linguistic information, one could expect the performance to drop when the LP order is increased.

The performance of the baseline MFCC is shown as a dotted red line. The performance of 8^{th} order residual (53.8%) is much lower than MFCC features (68.0%). Phoneme recognition using simple features [4, 1], yielded accuracies of 40.8% and 31.2% respectively. Performance of 8^{th} order LP residual lies between simple features and MFCC features. Combining residual with subband information improves phoneme recognition by 3%, while adding spectral slope does not improve it.

5.2. Effect of randomization

Table 4 presents the diarization performance of MFCC and LPR-8 feature vectors shuffled within non-overlapping blocks of count ($N = 5, 9, 13$). Randomizing MFCC features does not change the diarization performance significantly ($\leq 1\%$). Similarly, the performance of residual is unaffected by local temporal shuffling.

On the other hand, randomizing 8^{th} order residual with a block size of 13 yields a phoneme accuracy of 29.1%, which is comparable to phoneme accuracies obtained for simple features [1] and [4], which yield 31.2% and 40.8% respectively.

Table 4: *Speaker error on RT06 MDM (%): randomization of LPR-8 and MFCC for 3 blocks of feature vectors.*

Size	LPR-8	MFCC
5	23.2	17.8
9	24.0	18.9
13	23.7	18.3

5.3. Putting privacy and diarization together

Apart from phoneme recognition, we also did informal listening tests on speech synthesized from MFCC representation of 8^{th} order residual. These do not sound intelligible. This is due to a further loss in information from residual by representing it using MFCC. Future work will study comprehensive listening tests to assess privacy.

On the other hand, the diarization performance of the proposed features are 2% below the baseline MFCC features on SDM and MDM conditions. However, the effect of a 2% drop in diarization performance on socially relevant tasks such as dominance estimation have been shown to be minimal, if any [14].

6. Conclusion

We investigated residual for speaker diarization on MDM and SDM conditions in privacy-sensitive settings. Using mutual information, we compared two representations of residual. Combining residual with subband information and spectral slope yielded a diarization performance close to traditional MFCC features on RT06eval. Diarization performance was sensitive to LP order. As a means to quantify the abstract notion of privacy, we conducted phoneme recognition studies. Experiments showed that residual features yield low phoneme accuracies in comparison with MFCC features. Overall, our study suggests that privacy-sensitive features, clearly needed for ethical recording of real conversations, are feasible and competitive. Informal experiments synthesizing speech from MFCC representation of 8^{th} order residual sound unintelligible. Future work will incorporate comprehensive listening tests to validate this.

7. Acknowledgements

This work was supported by the Swiss National Science Foundation through the projects Intelligent Cognitive Systems (ICS) and the National Centres of Competence in Research (NCCR) IM2.

8. References

[1] D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, “A privacy-sensitive approach to modeling multi-person conversations,” in *Proc. of IJCAI*, 2007.

[2] D. Gatica-Perez, “Automatic nonverbal analysis of social interaction in small groups: A review,” *Image and Vision Computing*, vol. 27, pp. 1775–1787, 2009.

[3] X. Anguera, “Robust Speaker Diarization for Meetings,” Ph.D. dissertation, Universitat Politcnica de Catalunya, 2006.

[4] S. H. K. Parthasarathi, M. Magimai.-Doss, H. Bourlard, and D. Gatica-Perez, “Evaluating the robustness of privacy-sensitive audio features for speech detection in personal audio log scenarios,” in *Proc. of ICASSP*, 2010.

[5] D. P. W. Ellis and K. Lee, “Accessing minimal impact personal audio archives,” *IEEE Multimedia*, vol. 13, pp. 30–38, 2006.

[6] R. Donovan, “Trainable speech synthesis,” Ph.D. dissertation, Cambridge University, 1996.

[7] “http://www.nist.gov/speech/tests/rt/rt2006/spring/.”

[8] G. A. Miller, “Note on the bias of information estimates,” *Information Theory and Psychology*, pp. 95–100, 1954.

[9] P. Thevenaz and H. Hugli, “Usefulness of the LPC- residue in text-independent speaker verification,” *Speech Communication*, vol. 17, pp. 145–157, 1995.

[10] S. Furui, “Research on individuality features in speech waves and automatic speaker recognition techniques.” *Speech Communication*, vol. 5, pp. 183 – 197, 1986.

[11] 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*, vol. 36, pp. 871 – 879, 1988.

[12] J. Krumm, “A survey of computational location privacy,” *Personal and Ubiquitous Computing*, vol. 13, pp. 391–399, August 2009.

[13] J. Pinto, G. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, “Analysis of MLP based hierarchical phoneme posterior probability estimator,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, pp. 225–241, 2011.

[14] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, “Estimating Dominance in Multi-Party Meetings Using Speaker Diarization.” *IEEE Trans. on Audio, Speech, and Language Processing*, 2011.