# An Efficient Speaker Diarization using Privacy Preserving Audio Features Based of Speech/Non Speech Detection

S.Sathyapriya M.phil[1], A.Indhumathi M.Phil , Ph.D[2]

[1] *Mphil Scholar, Department of Computer Science, Dr.S.N.S College of Arts and Science, Saravanampatti, Tamilnadu, India*

[2] *Associate professor, Department of Computer Science, Dr.S.N.S College of Arts and Science, Saravanampatti, Tamilnadu, India.*

*Abstract*— **Privacy-sensitive audio features for speaker diarization in multiparty conversations: i.e., a set of audio features having low linguistic information for speaker diarization in a single and multiple distant microphone scenarios is a challenging research field in now-a-days. Existing system used a supervised framework using deep neural architecture for deriving privacy-sensitive audio features. In proposed system Patterns of speech/nonspeech detection (SND) is utilized for privacy-sensitive audio feature to capture real-world audio. SND and diarization can then be used to analyze social interactions. In this research privacy preserving audio features has been investigated instead for recording and storage that can respect privacy by minimizing the amount of linguistic information, whereas achieving modern performance in conversational speech processing tasks. Certainly, the main contribution of the proposed system is the achievement of state-of-the-art performances in speech/nonspeech detection and speaker diarization tasks using such features, which we refer to, as privacy-sensitive. In addition a comprehensive analysis of these features has been provided for the two tasks in a variety of conditions, such as indoor (predominantly) and outdoor audio. To objectively evaluate the notion of privacy, the proposed system use automatic speech recognition tests, with higher accuracy in either being interpreted as yielding lower privacy.**

*Keywords*— **Speech/Nonspeech Detection, Diarization,Privacy-sensitive features, deep neural networks, LP residual.**

## I. Introduction

Speech is acoustic signal which contains information of idea that is formed in speaker's mind. Speech is bimodal in nature [1] [2]. Speech processing can be performed at different three levels. Signal level processing considers the anatomy of human auditory system and process signal in form of small chunks called frames [3]. In phoneme level processing, speech phonemes are acquired and processed.

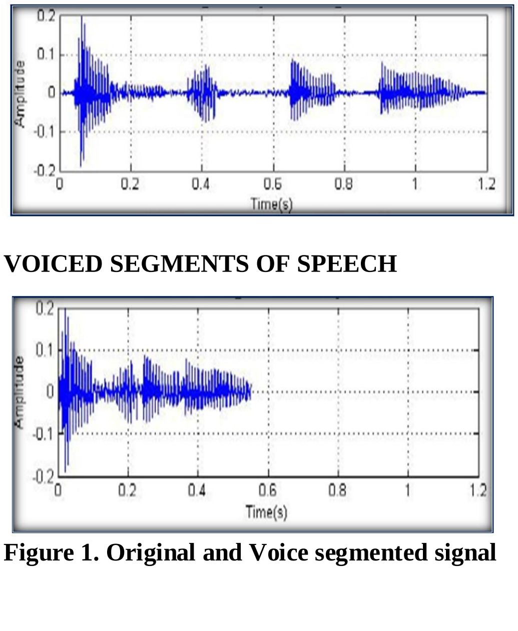
The major work of speech recognition has been explored to analyse the social interactions using multimodal sensors with an emphasis on audio.

Analysis of conversations can then carried by modeling the speech/speaker activities produced by a speaker diarization system. The objective of speaker diarization is to segment an audio recording into speaker-homogeneous regions whereas the output of a diarization system may appear to be restrictive, there are a growing number of applications that model the speech/speaker activities for learning productivity and personal health.

For instance, [4] presents numerous case studies investigating organizational productivity using such measures, executed on wearable devices. In the medical community, portable device were used to record features [3] from which, among others, speech activity was extracted to study physical and mental health.

For speech recognition the acoustical parameters of spoken signal have been investigated, and being able to be labelled into two types of processing field: First group corresponds to spectral based parameters and another group is dynamic time series. Very popular spectral based parameter used in recognition method is the Mel Frequency Cepstral Coefficients called MFCC. Modern diarization systems [5], uses features obtained from the spectral shape such as Mel Frequency Cepstral Coefficients (MFCC). Whereas these features are widely strong to SDM, Milner *et al.* [6] show that largely intelligible speech can be reconstructed from MFCC has been shown in fig1. The following figure1 shows the original signal and voice segmented signal of speech.

**ORIGINAL SPEECH SIGNAL**



**VOICED SEGMENTS OF SPEECH**

**Figure 1. Original and Voice segmented signal**

## II. RELATED WORK

Most relevant work in privacy sensitive audio features, MFCC Features and in LP residual is discussed in the following literature.

### A. PRIVACY-SENSITIVE FEATURES

In [7] Ellis presented an approach to privacy-sensitive audio cues which relies on storing certain statistical properties, such as long-term averages of the short-term spectral based features . This approach was effective in scene analysis tasks for referencing large personal audio logs [7], [8]. But speech is possibly the most informative content in audio logs, and conversation analysis needs a finer temporal resolution of features.

In [9] S. Basu presented an earliest work of features in the case of automatic conversation analysis. Here a dynamic conversation analysis is carried out using nonverbal prompts based on short-term autocorrelation and relative spectral entropy. The underlying features were considered with respect to robustness to noise, robustness to environment and robustness to microphone distance.

### B. LP RESIDUAL

In [10] J. Makhoul presented a Linear prediction (LP) analysis of speech which assumes the source-filter model and then approximates three components: (a) an all-pole model; (b) a gain and (c) a residual.

In [11] S. H. K. Parthasarathi presented a method, methods of obfuscating data to preserve privacy used randomization in sensor data research. Convinced by obfuscation methods, we conjecture that temporal dynamics of the speech signal has been conjectured by its intelligibility, it could be less for speaker recognition tasks.

### C. MFCC FEATURES

In [12] Modern diarization systems were presented and features are derived from the spectral shape such as Mel Frequency Cepstral Coefficients (MFCC). Whereas these features are relatively robust to SDM.

Previous approaches in [13][14]is presented to derive the privacy sensitive features which have focused on either reinterpreting simple, frame-level heuristics for estimating speech activity in conversations or computing long-term averages of standard features for indexing audio logs .

However these methods were not proposed for diarization, a choice that is further supported by results in our initial experiments.

## III. EXISITNG SYSTEM

In existing, Speaker Diarization is composed of three stages. Privacy sensitive features such as LP Residual features, DNN features were used. The following features were used for Diarization purpose.

### A. LINEAR PREDICTION RESIDUAL FEATURES

In this, features can be derived from Linear Prediction (LP) residual, subband information, and spectral slope.

#### 1) LP residual

LP residual is extracted every 10 ms, using a hamming window of size 30 ms.Two representations of the residual studied are: Real-cepstrum and MFCC. These representations have been fixed at 19 dimensions to have the same dimensions as the baseline MFCC features.

#### 2) Subband information

Earlier studies have shown that the spectral subband, 2500 Hz to 3500 Hz, carries speaker specific information. For speaker change detection (SCD),the subband can be represented by using three MFCC. An MFCC representation decorrelates the filterbank energies and makes it suitable for a Gaussian Mixture Model (GMM) with diagonal covariance matrices. To compute subband MFCC, we employed HCopy : it bandlimits the signal between 2500 Hz to 3500 Hz, and distributes the four filterbank channels equally on the mel scale such that the lower cutoff of the first filter is at 2500 Hz and the upper cutoff of the fourth filter is at 3500 Hz. Three cepstral coefficients are then calculated from the four values using Discrete Cosine Transform (DCT).

#### 3) Spectral shape:

Spectral slope (SS) is a way to characterize spectral energy distribution, with the spectrum of female speakers tending to show a steeper slope than male speakers.

### B. DNN FEATURES:

The aim of the proposed approach is to model the peaks in the spectral envelope that tend to carry linguistic information. For this the spectral envelope is reconstructed from a phoneme representation. The reconstructed envelope is then filtered to obtain a residual (similar to LP residual), which is represented using MFCC.

## IV. PROPOSED WORK

Aiming at discriminating speech and non-speech segments from a given audio signal, speech/non-speech detection (SND) is crucial for speech signal processing applications.
The proposed work extends the existing features by considering the Features patterns of Speech/non speech. By using these features diarization of social network can be analysed. Speech features are extracted from recorded speech of a male or female speaker and compared with templates available in database. This feature can be parameterized by using PLP.
The following features are privacy preserving feature which supports SND:

LP order**:** We study LP residual by varying the prediction orders from 2 to 20. The choice of the LP order presents a tradeoff between privacy and SND performance.

Temporal context**:** The efficacy of temporal context for LP residual with respect to the SND task is studied by varying the temporal support from no-context (1 frame) to 101 frames (with 50 frames for both left and right context).

### A. *OBFUSCATION:*

1) Local temporal randomization:

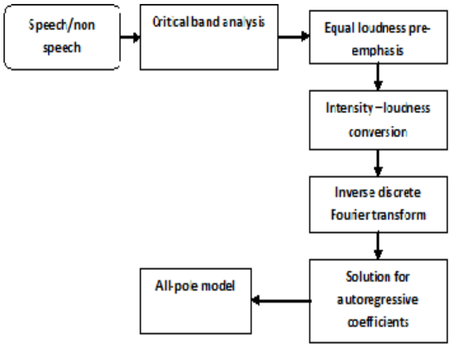
Feature vectors within a block of size (N = 1, 5, 9, 13) are shuffled. A uniform pseudorandom number generator was used to shuffle the frames in the block. It can be noted that a randomization of N frames could result in two successive frames being separated by 2 · (N −1) frames (equivalently 2· (N −1) ·10 ms). We chose block sizes up to 13 frames because results in [34] indicate that phonetic information in the speech signal up to 230 ms can be exploited for phoneme recognition.

2) Local temporal averaging:

Feature vectors within block of size (N = 1, 5, 9, 13) are averaged. These methods are applied to and LP residual, DNN and SND based features.

3) Parameter selection using PLP

The Perceptual Linear Prediction PLP model developed by Hermansky. PLP models the human speech based on the concept of psychophysics. PLP discards irrelevant information of the speech and thus improves speech recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of human auditory system.


Speech/non speech
Critical band analysis
Equal loudness pre-emphasis
Intensity–loudness conversion
Inverse discrete Fourier transform
Solution for autoregressive coefficients
All-pole model

Detailed steps of PLP computation is shown in figure 2.

**Figure 2: The block diagram of plp is shown**

### B. *MUTUAL INFORMATION BASED ANALYSIS*

*1) LP residual:*

In the case of LP, an independent source-filter model assumption is part of the modeling. The all-pole model can be reinterpreted as an estimate of the phoneme information $(\breve{X})$ and it is obtained in an unsupervised fashion as the smoothed spectral envelope. The resulting LP residual becomes $g^{*}(X)$.

*2) Deep neural network filter*

An alternative is to train a data driven filter that yields$\breve{X}$, given X as input. Let us consider a 5-layer MLP for phoneme classification, with bottleneck architecture. Let X denotes the input, and let Z denote the random variable at the output of the MLP. Then

$z=\psi(X;\theta_1,\theta_2,D)$

where $\theta_1,\theta_2$ is the parameter set of MLP

D denotes training data.

*3) SND Filter*

A separate MLP classifier was trained on each feature set for speech/non speech $(\theta_S,\theta_{NS})$ targets based on the ground truth definition g*. The minimization of cross entropy E was used as the training criterion S. All the features are normalized to zero-mean and unit variance at the input of the MLP using the global means and variances estimated on the training data. The number of hidden and input units in the MLP classifier trained for simple and LP based features and was identified by model selection.

## V. EXPERIMENTAL RESULTS

### A. *ANALYSIS OF PRIVACY*

So far, we have investigated LP residual and DNN features and SND features for speaker diarization. We now proceed to analyze privacy. The two methods to analyse the privacy is by using following notions:

- Human speech recognition (HSR) rates
- Automatic speech recognition (ASR) rates

The proposed work can be experimentally verified using Automatic Speech recognition rates

*1) Automatic speech recognition (ASR) rates*

The following features from this audio: DNN with LP, DNN-with Perceptual LP, DNN-SND Reconstruction yields audio from the 3 sets of features for each of the 20 sentences.

Experimental studies were performed on TIMIT database (4.3 hours), sampled at 16kHz. Experiments were conducted excluding the 'sa' dialect sentences. The training data holds of 3000 utterances from 375 speakers, cross justification data consists of 696 utterances from 87 speakers, and the test data set consists of 1344 utterances from 168 speakers. The hand-labeled dataset using 61 labels is planned to the standard set of 39 phonemes.

Comparative graph for DNN with LP and DNN with PLP is shown

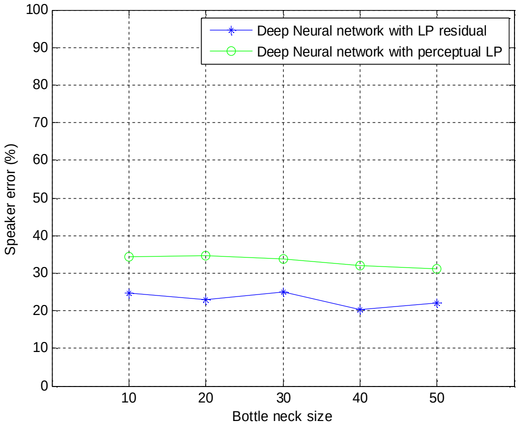


Figure 3.Speaker error Vs Bottle neck size

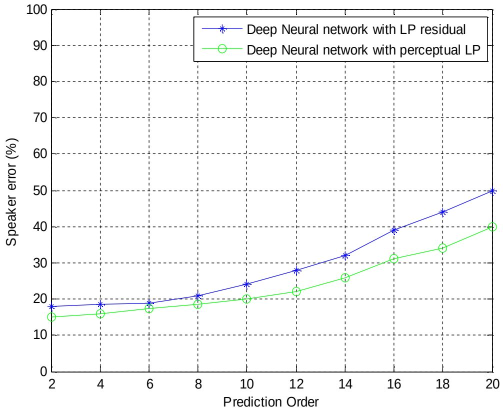


Figure 4: Speaker error Vs Prediction Order

Thus the above graph in figure 3 and 4 shows that it provides less speaker error when Speech and Non speech pattern of features is used in DNN-LP and DNN-PLP.

## VI. CONCLUSION

The proposed research investigated Patterns of speech/non speech detection and diarization can then be used to analyze social interactions three different approaches to privacy sensitive features. In this research three different approaches to privacy-sensitive audio features for robust speaker diarization, namely, LP residual, DNN based and speech/non speech detection (SND). These features have been investigated for speaker diarization in single and multiple distant microphone conditions. The conception of audio privacy was interpreted as the linguistic message, and methods to assess them in terms of phoneme recognition and different speeches from the speakers. Experimental result provides better accuracy in terms of producing less speaker error for spontaneous increase in bottleneck sizes and prediction order.

## REFERENCES

[1] Syed Ayaz Ali Shah, Azzam ul Asar, S.F. Shaukat, “Neural Network Solution for Secure Interactive Voice Response,” World Applied Sciences Journal 6 (9): 1264-1269, ISSN 1818-4952, 2009

[2] Corneliu Octavian Dumitru, Inge Gavat, “A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language,” International Symphosium ELMAR, 07-09 June, 2006, Zadar, Croatia.

[3] 6. Goranka Zoric, “Automatic Lip Synchronization by Speech Signal Analysis,” Master Thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Oct-2005

[4] D. Olguin-Olguin and A. Pentland, “Sensor-based organisational design and engineering,” *Int. J. Organisational Design and Eng.*, vol. 1, pp. 5–28, 2010.

[5] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” in *Proc. Workshop Classification of Events, Activities, and Relationships and the Rich Transcript. Meeting Recognit.*, 2008.

[6] B. P. Milner and X. Shao, “Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 24–33, Jan. 2007

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

[8] D. P. W. Ellis and K. Lee, “Features for segmenting and classifying longduration recordings of personal audio,” in *Proceedings of Workshop on Statistical and Perceptual Audio Processing*, 2004.

[9] S. Basu, “Conversational scene analysis,” Ph.D. dissertation, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science, 2002.

[10] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of IEEE*, vol. 63, pp. 561–580, 1975

[11] S. H. K. Parthasarathi, M. Magimai.-Doss, H. Bourlard, and D. Gatica- Perez, “Investigating privacy-sensitive features for speech detection in multiparty conversations,” in *Proceedings of Interspeech*, 2009.

[12] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” in *Proc. Workshop Classification of Events, Activities, and Relationships and the Rich Transcript. Meeting Recognit.*, 2008.

[13] D.Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, “A privacy-sensitive approach to modeling multi-person conversations,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2007.

S. H. K. Parthasarathi, M. Magimai-Doss,H. Bourlard, andD.Gatica- Perez, “Evaluating the robustness of privacy-sensitive audio features for speech detection in personal audio log scenarios,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4474–4477.