IDIAP RESEARCH REPORT

WORDLESS SOUNDS: ROBUST SPEAKER DIARIZATION USING PRIVACY-PRESERVING AUDIO REPRESENTATIONS

Sree Hari Krishnan Parthasarathi Hervé Bourlard
Daniel Gatica-Perez

Idiap-RR-28-2012

SEPTEMBER 2012

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

Wordless Sounds: Robust Speaker Diarization using Privacy-Preserving Audio Representations

Sree Hari Krishnan Parthasarathi *Student Member, IEEE*, Hervé Bourlard *Fellow, IEEE* and Daniel Gatica-Perez *Member, IEEE*

***Abstract*—This paper investigates robust privacy-sensitive audio features for speaker diarization in multiparty conversations: ie., a set of audio features having low linguistic information for speaker diarization in a single and multiple distant microphone scenarios. We systematically investigate Linear Prediction (LP) residual. Issues such as prediction order and choice of representation of LP residual are studied. Additionally, we explore the combination of LP residual with subband information from 2.5 kHz to 3.5 kHz and spectral slope. Next, we propose a supervised framework using deep neural architecture for deriving privacy-sensitive audio features. We benchmark these approaches against the traditional Mel Frequency Cepstral Coefficients (MFCC) features for speaker diarization in both the microphone scenarios. Experiments on the RT07 evaluation dataset show that the proposed approaches yield diarization performance close to the MFCC features on the single distant microphone dataset. To objectively evaluate the notion of privacy in terms of linguistic information, we perform human and automatic speech recognition tests, showing that the proposed approaches to privacy-sensitive audio features yield much lower recognition accuracies compared to MFCC features.**

***Index Terms*—Privacy sensitive audio features, speaker diarization, LP residual, deep neural networks, listening tests.**

I. Introduction

OUR work takes place in the context of analyzing social interactions using multimodal sensors with an emphasis on audio [1]. Towards this we wish to capture conversational and ambient sounds using portable audio recorders. Analysis of conversations can then proceed by modeling speaker turns and durations using speaker diarization.

A key impediment to making progress in the ubiquitous capture of real-life audio is privacy. Recording and storing raw audio would breach the privacy of people whose consent has not been explicitly obtained [2]. While the term "privacy-preserving" or "privacy-sensitive" can have different connotations in different areas of computing, Wyatt et al [3] suggest that the linguistic message in audio is perhaps the most privacy-sensitive information.

One approach to preserving this notion of privacy is to implement an online speaker diarization system directly on the device and store information derived from its output. However, a caveat of this approach is that the set of possible tasks is limited by the output of the diarization system. Other sources of information such as emotion, or the background information are inevitably lost. Another challenge with such a design is the computational limitation imposed by the device [4].

Alternatively, one could store lower-level audio features with the constraint that neither intelligible speech nor lexical content can be reconstructed. Such features are referred to as privacy-sensitive features in the literature [3]. While such audio features may appear to be restrictive, there are different applications that use only the nonverbal cues in speech for the study of social behavior [1].

A further issue inherent to capturing spontaneous conversations using portable recorders is the necessity of speech processing systems, including diarization, to be robust to single distant microphones (SDM). This is in contrast to more conventional speaker diarization systems which work with multiple distant microphones (MDM). In this setting, the long-term scope of our work aims at robust privacy-sensitive audio features enabling conversation and acoustic scene analysis. Our focus in this paper though is on features for speaker diarization in SDM settings, exploring the tradeoff between diarization performance and audio privacy.

Features used in state-of-the-art speaker diarization systems such as [5], in general, are a weighted combination of Mel Frequency Cepstral Coefficients (MFCC) and Time-Delay Of Arrival (TDOA). While such features have been shown to be robust to single distant microphones, Milner et al. [6] show that highly intelligible speech signal can be reconstructed solely from MFCC vectors.

Previous approaches to privacy-sensitive features have focused on either reinterpreting simple, frame-level heuristics for conversation analysis [3], [7], or computing long-term averages of standard features for indexing personal audio logs [2]. However these methods were not proposed for speaker diarization, a choice that is further supported by results in our preliminary experiments.

In this paper, drawing motivation from the source-filter model of speech production, we systematically investigate linear prediction (LP) residual for diarization. Two different representations of LP residual are compared, namely, real-cepstrum and MFCC, with the latter representation yielding better performance. We then study the effect of prediction order on diarization. Next, we explore the combination of residual with subband information (2.5 kHz to 3.5 kHz) and spectral slope. To enforce stricter privacy, we explore obfuscation methods such as local temporal randomization [8]

S.H.K Parthasarathi is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland and Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland e-mail: hari.parthasarathi@idiap.ch

H. Bourlard is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland and Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland e-mail: herve.bourlard@idiap.ch

D. Gatica-Perez is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland and Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland e-mail: gatica@idiap.ch

of residual features.

In addition to LP residual, we propose a supervised residual, obtained by a deep neural network with a bottleneck architecture. We benchmark LP and deep neural network residual against MFCC using the diarization system presented in [9] on the SDM and MDM settings from the NIST RT07 dataset [10]. Experiments show that the proposed features yield performances close to the MFCC features in SDM condition.

The notion of linguistic privacy in audio remains something that is difficult to quantify and evaluate. Studies such as [3] and [11] indicate that the main privacy concerns in audio are reconstructibility of intelligible speech and linguistic information. As ways to evaluate these, we present human speech recognition (HSR) and phoneme recognition studies, with higher recognition accuracy being interpreted as lower privacy. Our studies show that the proposed approaches are more privacy-sensitive than MFCC.

The contributions of this paper are: (a) a systematic investigation of LP residual based features for speaker diarization in SDM and MDM conditions; (b) a deep neural network for privacy-sensitive features; and (c) framework and evaluation of audio privacy in terms of HSR and phoneme recognition. The findings of this paper are that the proposed privacy-sensitive features yield a diarization performance close to the MFCC features on the SDM data, while yielding much stricter privacy in terms of intelligibility and phoneme recognition accuracy.

The rest of the paper is organized as follows. Section II reviews the literature on LP residual and deep neural networks. The overall methodology of this paper is summarized in Section III. A description and an analysis of the proposed features is given in Section IV, while Section V discusses the diarization setup. Parameters selection experiments associated with the proposed features is described in Section VI. Subsequent validations on the RTeval07 dataset are presented in VII. We revisit privacy in Section VIII. Finally, conclusions are drawn in Section IX.

II. Related work

In the introduction we briefly discussed existing work on privacy-sensitive features. In this section, we summarize relevant work in LP residual and deep neural networks.

A. Linear prediction residual

It is generally known that up to two or three formants are required to synthesize intelligible speech or to reconstruct the lexical information [12]. Our approach to preserving privacy is based on adaptively filtering out information about these spectral peaks. This approach is motivated by the source-filter model.

Linear prediction (LP) analysis of speech [13] assumes the source-filter model and it estimates three components, namely an all-pole model, a residual and a gain. The vocal tract response is modeled by the all-pole model, with the model capacity being determined by the prediction order (p). The LP residual, obtained by inverse filtering the speech signal with the all pole model, can be considered to be privacy-preserving. Depending on the prediction order, the LP residual contains mostly information about the excitation source of the speakers [14]. It has been shown that humans can recognize speakers by listening to the LP residual signal [15].

Previous works have exploited the speaker information in LP residual. For example, the residual has been used as a complimentary feature for speaker recognition in [16]. In an earlier work [17], we reinterpreted LP residual as a privacy-sensitive feature for speaker change detection. The choice of the LP order could be interpreted as a tradeoff between privacy and speaker information. The real-cepstral representation of residual was investigated for various prediction orders in combination with subband MFCC and spectral slope. LP residual has also been exploited for speaker recognition in [14] using an autoassociative neural network.

To our knowledge, this is the first work to investigate LP residual for speaker diarization in both single and multiple distant microphone scenarios. A diarization study involving features with respect to single distant microphones is particularly relevant to capturing spontaneous conversations using portable recorders. This setting is in contrast to meeting room speaker diarization tasks which work with audio from multiple distant microphones.

In sensor data research, methods of obfuscating data representations to preserve privacy are well established [8]. Randomization is a form of obfuscating data. We derive motivation from obfuscation methods and hypothesize that, while temporal dynamics of the speech signal is important for its intelligibility, it could be less important for speaker recognition tasks. We analyze local temporal randomization (within 250 ms) of LP residual based features for diarization.

B. Deep neural networks

We briefly review here the relevant literature on deep neural networks as a means to represent phoneme information. In subsequent sections, we describe and exploit privacy-sensitive features derived from a deep architecture.

Multilayer feedforward neural networks with a 3-layer architecture, also called multilayer perceptrons (MLP), have been used for feature extraction in the automatic speech recognition (ASR) community for several years [18], [19]. Recently, deep neural networks, i.e., typically the number of layers being more than three (alternatively, number of hidden layers being more than one), have been receiving attention from both machine learning and speech research community ([20], [21]) due to their ability to represent knowledge compactly and in a principled fashion. The motivation for this has been attributed to results from complexity theory of circuits [22].

Of particular interest to our work are deep neural networks with bottleneck architectures to represent phoneme information. In the field of ASR, deep neural networks with bottleneck architectures recently started to be investigated in the quest towards obtaining better phoneme representation before further processing by a HMM/GMM system [21]. For example, in [21] the output (before the sigmoid nonlinearity) taken from the bottleneck layer of a trained five-layer MLP, was used in a conventional HMM/GMM system to yield promising results.

A key issue in exploiting a deep neural networks is the inherent difficulty in training the weights. A gradient-based optimization starting from random initialization has been reported to get trapped in local optima leading to poor solutions [22]. This was also observed by us in our studies in training neural networks with more than three layers for phoneme recognition on TIMIT, to the extent that deeper networks perform worse than MLPs with one hidden layer.

Two common strategies to address this difficulty are, greedy layer-by-layer training [23], and an autoencoder training [20]. In [24], features derived from the bottleneck layer of a 5-layer deep neural network trained with greedy layer-by-layer method, was shown to yield promising performance for an ASR task on over 100 hours of meeting audio data.

The constraints of privacy in features imply the necessity to capture the complement of phoneme information captured by the bottleneck layer of a 5-layer MLP. In this context, our work exploits features derived from the bottleneck layer of a deep neural network as information that needs to be filtered from the spectrum. In Section IV-B, we describe the proposed method in detail.

III. Our methodology

In this section, we summarize our overall methodology, also illustrated using a block diagram in Figure 1. These blocks are described below.

(a): We begin with a detailed description of the features extracted from LP residual and deep neural networks. Sections IV-A and IV-B describe these features in detail. To gain insight into the features, this is followed by a more formal analysis of the proposed features in terms of mutual information.

(b): Evaluating privacy-sensitive features entails a comparison of diarization performance as well as an evaluation of linguistic privacy. Details of the diarization system, features, datasets, and the baseline performance figures are presented in Section V. Parameter selection experiments associated with the proposed features for diarization is done on the development data (RTeval06) on single and multiple distant microphone data (Section V). Results on evaluation data (RTeval07) is presented in Section VII.

(c): This paper quantifies linguistic privacy using human listening tests and automatic phoneme recognition studies. Section VIII provides further details on the methodology followed and the results obtained using these tests.

IV. Privacy-sensitive features

In this section, we first present the details in deriving the proposed features and follow that by an analysis based on mutual information framework.

A. LP residual based features

We now look at extracting features from LP residual, subband information, and spectral slope.

(a) LP residual: LP residual is extracted every 10 ms, using a hamming window of size 30 ms. The representations of the residual studied are: a real-cepstrum representation

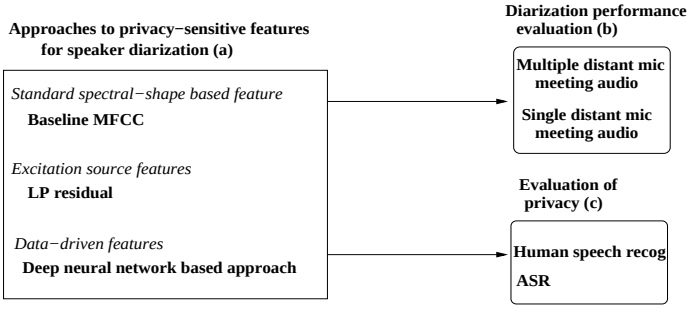

Fig. 1. Block diagram of our approach. A detailed discussion of the figure is provided in Section III.

([16]) with a fixed number of 19 coefficients and a MFCC representation with 19 coefficients. The MFCC representation is computed using HTK [25]. These representations have been fixed at 19 dimensions so as to have the same dimensions as the baseline MFCC features. Feature selection experiments investigating the choice of representation is presented in detail in Section VI. We then 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.

(b) Subband information: Previous studies have shown that the spectral subband from 2500 Hz to 3500 Hz, carries speaker specific information [26]. In our earlier study [17], we exploited the relative importance of the subband 2500 Hz to 3500 Hz over the two neighboring subbands (1500 Hz - 2500 Hz and 3500 Hz - 4500 Hz) for a speaker change detection (SCD) task. We also showed that computing three MFCC coefficients from this subband was better than computing the logarithmic energy from the subband. A further advantage of the MFCC representation is that it decorrelates the filterbank energies and makes it suitable for a Gaussian Mixture Model (GMM) with diagonal covariance matrices.

(c) Spectral shape: Speakers differ from each other in the distribution of spectral energies within their speech signals [27]. 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 [17] we showed that the first cepstral coefficient (c_1) obtained from LP analysis can enhance SCD when combined with LP residual features.

(d) Obfuscation/local temporal randomization: Feature vectors within a block of size ($N = 1, 5, 9, 13$) are shuffled. A uniform pseudo-random 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 $2 \cdot (N - 1)$ frames. In our work, we chose block sizes up to 13 frames since results in [28] indicate that phonetic information in the speech signal up to 230 ms can be exploited for phoneme recognition.

B. Deep neural net based features

We extract the bottleneck features derived from a 5-layer MLP, trained in a greedy layer-by-layer fashion. From [24],

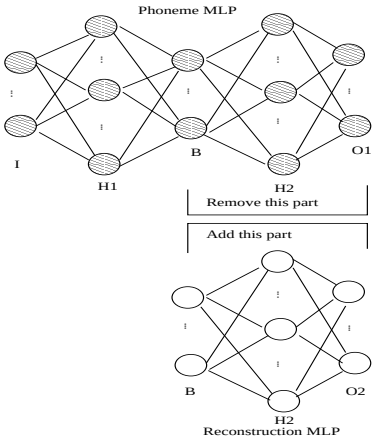

Fig. 2. 5-layer deep neural network with bottleneck architecture. (a) 5-layer phoneme MLP is trained with phoneme targets using cross entropy criterion (b) Keeping weights for the first 2 layers fixed, and removing last 2 layers, a reconstruction MLP is trained for the last two layers with squared error criterion.

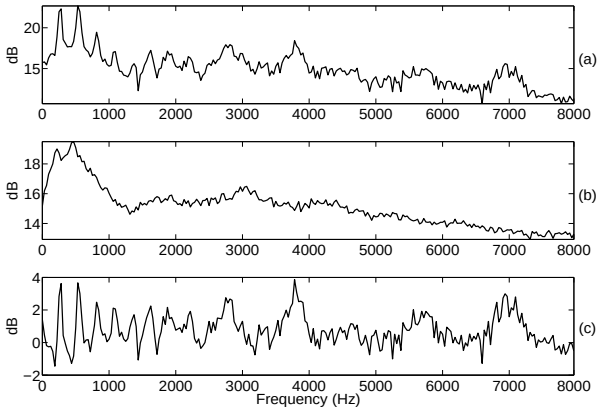

Fig. 3. Steps involved in deep neural network based filtering: (a) Estimated power spectrum of /iy/ sound (b) Reconstructed phoneme spectrum from the bottleneck layer (c) Filtered spectrum

bottleneck features can be considered to capture phoneme information. Using these bottleneck features, we train a second, 3-layer regression neural network to reconstruct the power spectrum: i.e., the second neural network takes the bottleneck features as input and outputs the estimated power spectrum of speech. Assuming a source-filter production model, the reconstructed spectrum is filtered from the original spectrum of the speech signal.

This approach can be viewed as follows: A 5-layer phoneme MLP is trained with phoneme targets using cross entropy criterion. Keeping weights for the first 2 layers fixed, and removing last 2 layers, a reconstruction MLP is trained for the last two layers using squared error criterion. An illustration of this is provided in Figure 2. We now analyze this architecture, before ending this section with an illustration of our deep neural approach.

(a) Phoneme MLP: There are five sets of parameters to the phoneme MLP: the input (I), the first expansion layer (H1), the bottleneck layer (B), the second expansion layer (H2), and the output layer (O1).

Some reasonable choice of inputs to the phoneme MLP include (i) MFCC or PLP; and (ii) DFT square magnitude vectors (obtained from 512 point FFT) as estimated power spectrum. For the sake of limiting the number of experiments, we analyzed both cases without explicit temporal context.

The number of units in first and second expansion layers can be varied independently. In our experiments, the number of nodes in H1 and H2 was kept same. This is done to reduce the number of experiments. Furthermore, experiments in [29], varying the ratio of H1 to H2 did not did not show appreciable difference in ASR performance.

We treat the bottleneck layer as a dimensionality reduction layer, similar to studies such as [21]. Reasonable choices of the number of units in the bottleneck layer can be between 20 to 40 [29]. In our analysis, the output of the bottleneck layer before the sigmoid activation is used. This is similar to other studies.

The output layer of the phoneme MLP represents the phoneme class and we use 39 units with softmax nonlinearity. The phoneme classification network was trained by growing MLPs layer-by-layer on the TIMIT. Cascaded MLPs with 3, 4, and 5 layers are trained using standard back propagation algorithm by minimizing the cross entropy error criterion. We excluded the 'sa' dialect sentences. The TIMIT training data consists of 3000 utterances from 375 speakers and the cross-validation data consists of 696 utterances from 87 speakers. The hand-labeled dataset using 61 labels is mapped to the standard set of 39 phonemes [28].

(b) Reconstruction MLP: There are three sets of parameters to the reconstruction MLP: the input from the bottleneck layer (B), the expansion layer (H3), and the output layer (O2).

The input to the reconstruction MLP is the linear output (i.e., before the nonlinearity) of the bottleneck layer of the phoneme network. No temporal context (1 frame) is used for the second MLP. The number of nodes in the expansion layer (H3) is varied independent of H1 and H2. The primary choice for the output of the reconstruction MLP is the estimated power spectrum (257 coefficients). A further choice such as 19 dimensional MFCC was explored. In either case, the units have a linear activation function, and the MLP is trained on TIMIT train set using standard back propagation algorithm by minimizing the squared error criterion.

(c) Filtering: Filtering is then performed to remove the linguistic information. For the case where the output units are squared magnitude vectors, filtering is performed in this domain. The filtered squared magnitude vector is then converted to an MFCC representation of 19 dimensions. In case of the output units being MFCC, filtering is performed in this domain. These parameters are analyzed in Section VI.

(d) An example: Figure 3 illustrates this process for phoneme /iy/ using an example. Figure 3(a) plots the estimated power spectrum of /iy/. Observe that the broad spectral shape and the spectral details are manifest in the plot. Figure 3(b) shows the reconstructed phoneme spectrum from the bottleneck layer. From the plot, it can be observed that the reconstructed spectrum consists mainly of the spectral shape than the fine, spectral details. Figure 3(c) shows the filtered

spectrum. From this plot, it can be observed that the spectral shape (mainly the first formant) is filtered.

C. Mutual information based analysis

In this section, we present an analysis of the privacy-sensitive features using mutual information. Privacy in audio could be interpreted as a function that maximizes the mutual information (MI) with speakers while minimizing the MI with linguistic information. This framework is discussed next followed by an analysis of the features on TIMIT test data (consisting of 1344 utterances from 168 speakers).

1) MI framework: Given X, a multivariate continuous random variable denoting the log squared magnitude, and S, Q discrete random variables, denoting speaker and phoneme labels respectively, the goal is to find a transformation g that maximizes the function $I(g(X);S) - I(g(X);Q)$.

$$g^* = \arg\max_g I(g(X);S) - I(g(X);Q) \tag{1}$$

This equation is in general difficult to solve without additional constraints or assumptions. Assuming that Q and S are independent[1], the maximum of Eq (1) is reached for:

$$g^*(X) = \tilde{S} \tag{2}$$

where $\tilde{S}$ is a transformation of X that has maximum mutual information with S. A further assumption of a source-filter model of speech production simplifies this to:

$$g^*(X) = \tilde{S} = X - \tilde{X} \tag{3}$$

where $\tilde{X}$ is a transformation of X that has maximum mutual information with Q.

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 ($\tilde{X}$) and it is obtained in an unsupervised fashion as the smoothed spectral envelope. The LP residual naturally becomes $g^*(X)$ in Eq 3.

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

$$Z = \psi(X;\theta_1,\theta_2,\mathcal{D}) \tag{4}$$

where θ_1, θ_2 is the set of all parameters of the MLP (i.e., the weights and the biases) before and after the bottleneck layer respectively, and $\mathcal{D}$ is the training data. Let q_k denote the k^{th} phoneme and $\tilde{P}$ denote the estimated probabilities. The cross-entropy training criterion can be written as:

$$\begin{aligned}
\mathcal{J}(\theta_1,\theta_2) &= -E_X[\sum_k P(q_k|x)\log\tilde{P}(q_k|x)] \\
&= -\int_X p(x)\sum_k P(q_k|x)\log\tilde{P}(q_k|x)dx \\
&= -\int_X \sum_k P(q_k,x)\log\frac{\tilde{P}(q_k|x)\tilde{P}(x)\tilde{P}(q_k)}{\tilde{P}(x)\tilde{P}(q_k)}dx \\
&= -\int_X \sum_k P(q_k,x)[\log\frac{\tilde{P}(q_k,x)}{\tilde{P}(x)\tilde{P}(q_k)} + \log\tilde{P}(q_k)]dx \\
&= I(Q;X) - \sum_k P(q_k)\log\tilde{P}(q_k)
\end{aligned} \tag{5}$$

It can be seen from the above equation that minimum cross-entropy training is equivalent to maximum mutual information training [30]. Let B denote the random variable obtained at output from the bottleneck layer before the nonlinearity. Then,

$$B = \phi(X;\theta_1,\mathcal{D}) \tag{6}$$

where θ_1 is the set of parameters of the MLP up to the bottleneck layer. Furthermore, from data-processing inequality [31],

$$I(X;Q) \geq I(B;Q) \geq I(Z;Q) \tag{7}$$

However, given the constraints of the parameters (θ_1,θ_2), $I(Z;Q)$ is maximized. Similarly, $I(B;Q)$ is maximized for θ_1. This together with the fact that the dimension of the output at the bottleneck (B) is much smaller than that of the dimension of input (X), means that bottleneck (B) serves as a compression of input (X) retaining information that has maximum mutual information with the phonemes (Q).

Therefore, it is reasonable to assume that as the dimension of B is made much smaller than X, other information such as speakers (S) is lost at bottleneck (B). We now consider the second MLP, namely, the reconstruction MLP: i.e., this MLP is trained taking bottleneck output (B) as input and X as the training target, with minimizing the least-squares error cost function. The random variable at the output of this MLP ($\tilde{X}$) is a reconstruction of X and has therefore the same dimension as X. It is, however, reconstructed using B, which has maximum mutual information with Q (and has low MI with S, because of dimensionality reduction at B). Therefore, $\tilde{X}$ can be considered to be an estimate of Q. Inserting $\tilde{X}$ so obtained in Eq (3), we obtain $\tilde{S}$.

2) MI analysis: In practice, we can introduce a variable (λ) in Eq (1) to make it $I(g(X);S) - \lambda \cdot I(g(X);Q)$ and tune this variable for optimal values. Alternatively, we could plot $I(X;Q)$ versus $I(X;S)$ and make more qualitative assessments on the tradeoff between privacy and speaker information, in using these features. In this paper, we take the latter approach. Figure 4 shows such a plot. That is, $I(X;Q)$ versus $I(X;S)$, on the TIMIT test set. A higher $I(X;Q)$ could be interpreted as a feature with lower privacy. Similarly, a feature yielding higher $I(X;S)$ could be interpreted as a better feature for diarization. An ideal privacy-sensitive feature would be in the top-left of this plot.

For estimating the MI with phoneme and speaker labels, we use the following form of MI: $I(X;A) = H(X) - H(X|A)$,

[1]It might be that speakers can have biases towards choices of words and therefore towards phoneme

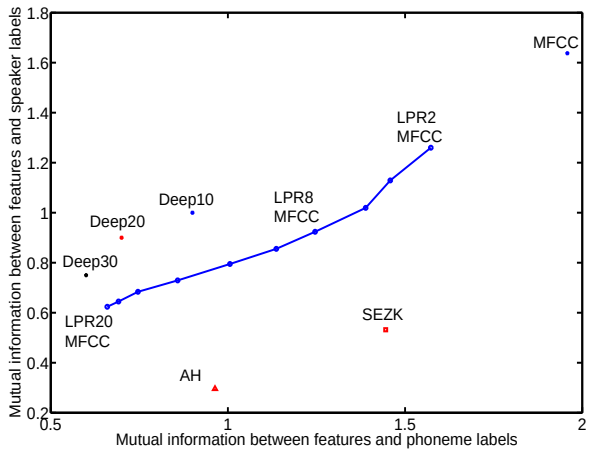

Fig. 4. Plot showing mutual information between the features and phonemes versus mutual information between the features and speakers. LPRx denotes residual features with LP order x. *SEZK* and *AH* denote the features from [7] and [3] respectively. Deepxy refer to deep neural network based features with bottleneck sizes corresponding to xy.

where A denotes either Q or S. To estimate entropies $H(X)$ and $H(X|A)$, we use k-means clustering algorithm to discretize the feature space. The features are then binned and the normalized bin-counts are then used to estimate $I(X;A)$. Model selection on the TIMIT training data is used to identify the number of clusters. Bias correction is performed using the Miller's formula on the estimated mutual information [32].

Figure 4 plots residual and deep neural network features represented using 19 dimensional MFCC. Standard features for diarization such as MFCC have high values of $I(X;Q)$ and $I(X;S)$. For the residual features, it can be observed that as the LP order increases, $I(X;Q)$ and $I(X;S)$ decrease. LP residual, with a prediction order of 8 appears to have much less MI with phoneme than MFCC features.

The figure also illustrates deep neural network features for different bottleneck sizes ($B = 10, 20, 30$). The input and the reconstruction layers are squared magnitude vectors. The expansion layers were fixed at 1000. Furthermore, the filtered output is represented using 19 dimensional MFCC vectors. It can be seen that the deep neural network features yield much lower MI with phoneme labels than does residual while yielding lower mutual information with speaker labels.

Features from [7] and [3], denoted as *SEZK* and *AH*, respectively, are privacy-sensitive but have low speaker information.

V. Diarization Setup

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

A. Baseline diarization system

The baseline system is an ergodic HMM as described in [9]. Each HMM state represents a cluster (speaker). The state emission probabilities are modeled by Gaussian Mixture Models (GMM) with a minimum duration constrain of 3 seconds. The algorithm follows an agglomerative framework, i.e, it starts with a large number of clusters (hypothesized speakers) and then iteratively merges similar clusters until it reaches the best model. After each merge, data are re-aligned using a Viterbi algorithm to refine speaker boundaries. The initial HMM model is built using uniform linear segmentation and each cluster is modeled with a 5 component GMM. The algorithm then proceeds with bottom-up agglomerative clustering of the initial cluster models [33]. At each step, all possible cluster merges are compared using a modified version of the BIC criterion [9].

The diarization system uses 19 dimensional MFCC features and the time delay of arrival (TDOA) features from the beamformed signal. The MFCC vectors are extracted every 10 ms, with a hamming window of size 30 ms, using HTK [25]. Delta and acceleration features are not used.

B. Privacy-sensitive features

The proposed privacy-sensitive features LP residual are compared against the baseline MFCC features by using the diarization system discussed in Section V-A. To summarize Section IV, LP residual is represented using 19 dimensional MFCC features and 19 dimensional real-cepstrum to make the comparison with baseline MFCC features. The subband frequency information between 2.5 kHz to 3.5 kHz is represented by 3 MFCC coefficients. Similarly, spectral slope (SS) is represented using first cepstral coefficient (c_1) obtained from LP analysis. For temporal randomization, features are shuffled with a uniform random number generator for block sizes ($N = 5, 9, 13$). The deep neural network based features are also represented using a 19 dimensional MFCC representation.

C. Datasets

Experiments were performed on NIST RT06 and RT07 evaluation data for Meeting Recognition Diarization task [10], [34]. RT06 evaluation data is used as the development dataset and it contains nine meeting recordings of approximately 30 minutes each. The best set of parameters is then used for benchmarking the proposed features against MFCC features on the RT07 dataset using the baseline diarization system. The evaluation dataset (RT07) contains eight meetings of nearly 43 minutes each. MDM data is obtained by denoising the individual channels using Wiener filter and then beamforming using the BeamformIt toolkit [35]. SDM experiments were performed on randomly selected individual MDM channels.

Speech/nonspeech segmentation is obtained using a forced alignment of the reference transcripts on close talking microphone data using the AMI RT06 first pass ASR models [36]. Since our interest in this paper is in evaluating the privacy-sensitive features for speaker segmentation and clustering, the same speech/nonspeech segmentation is used across all experiments.

D. Baseline performance

The results are reported in terms of Diarization Error Rates (DER). DER is the sum of speech/nonspeech errors and speaker errors. Speech/nonspeech errors is the sum of missed

speech and false alarm speech. For all experiments reported in this paper, we include the overlapped speech in the evaluation.

Table I lists the performance of the baseline diarization system on RT06 MDM and SDM evaluation data. The first 3 columns list the performance of the speech/nonspeech detection system in terms of missed speech, false alarm, and over all speech/nonspeech detection error. The overall speech/nonspeech 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 development MDM data. On RT06 we observe a performance gain of 3.7% on the MDM data over the SDM data.

TABLE I

RT06 evaluation data: The first 3 columns list the performance of the speech/nonspeech detection while the next 2 columns report performance of baseline MFCC features for MDM and SDM.

Evaluation	Miss	FA	Sp/nsp	Spkr err (%) MDM	Spkr err (%) SDM
RT06	6.5	0.1	6.6	**17.1**	**20.8**

VI. Parameter selection on RTeval06

Recall that we use RTeval06 as the development dataset. In Section IV-C, we presented an analysis of the features using MI on the TIMIT test set. In this section we perform parameter selection experiments for the proposed features using the baseline diarization system on RTeval06.

A. LP residual based features

We address three issues in this section: (a) the choice of representation (b) prediction order (c) combination with slope and subband energies.

1) Representations of LP residual: We study the 2 different representations of LP residual using the baseline diarization system described in Section V-A. Figure 5 shows the comparison between the 2 representations on the RT06 MDM evaluation data. It can be observed that MFCC representation yields a better performance for all prediction orders. It is interesting to observe that the gap between the two representations decrease as the prediction order increases. It could be due to MFCC being better able to capture spectral peaks than real cepstrum. From here on, we use MFCC representation of the residual.

2) Prediction order: The effect of LP order on MFCC representation of residual on both MDM and SDM data is presented in Figure 6. Both curves 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.

Let us consider prediction orders between 2 to 6. 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 [12]. Speaker error, therefore, seems to be relatively less affected.

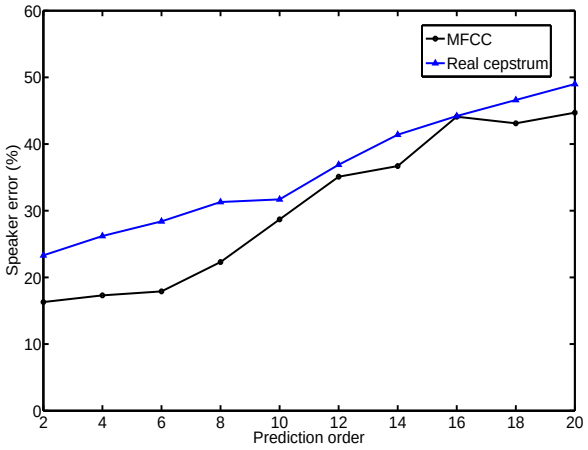

Fig. 5. Comparison between MFCC and real-cepstrum representations of the LP residual on RT06 MDM evaluation data.

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. We note that the vocal tract system is typically characterized by up to five resonances in the 0 to 4 kHz range. An LP order in the range 8 to 12 can model around 3 formants. Since higher order formants carry more speaker information, we note that increasing prediction order beyond 8 results in greater speaker errors.

For the last segment (orders greater than 12), we see a smaller drop in the performance as the order is increased. We note that the LP residual contains both modeling and excitation errors. As the LP order increases beyond 10, the contribution of the error in the residual signal is mainly due to the excitation error component.

It is also interesting to note that residual obtained by 2^{nd} order prediction performs slightly better than the baseline MFCC features in both SDM and MDM cases. Revisiting the performance versus privacy tradeoff, an LP order of 8 seems appropriate for the diarization task, since the first two formants are important for synthesizing an intelligible speech signal [12]. At this prediction order, residual yields a performance of 22.3% on the MDM data while yielding 29.2% on the SDM data.

3) Combination with subband and slope features: The effect of combining LP residual of 8^{th} order in MFCC representation with slope and subband on MDM data is presented in Figure 7. X-axis denotes the weight assigned to LP residual, while y-axis denotes the speaker error. We ran experiments varying the weights in steps of 0.05 starting from 0.05 to 0.95. A weight of 1 denotes that LP residual is not used in combination with the other features.

It can be observed from the plot that for either slope or subband energies, combining residual with weights less than 0.45 yields a lower performance than that achieved with LP residual alone. In general, combination with the subband energies yields a slightly better performance over slope at smaller weights. On the other hand, for weights over 0.4, the plot shows that the difference between slope and subband energies

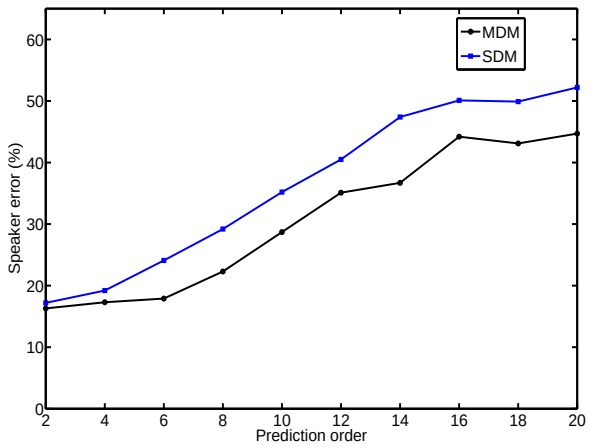

Fig. 6. Using MFCC representation of LP residual, prediction order vs speaker error is illustrated on MDM and SDM conditions of the development dataset (RT06).

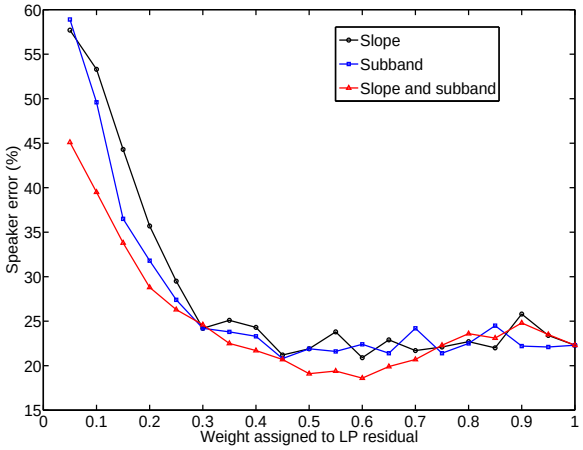

Fig. 7. Combination of LP residual (MFCC representation) with slope and subband. X-axis denotes the weight assigned to LP residual.

may not be significant. For instance, the best combination with spectral slope yields a speaker error of 20.7% at a weight of 0.45, while the best combination with subband energy yields a speaker error of 20.9% at a weight of 0.6.

We note that combining both slope and subband energies yields a consistent gain over combining with either of those features. Furthermore, combining both features with residual yields improvement over residual by itself, for weights between 0.45 to 0.8. The best performance of this combined system is 18.6% at a weight of 0.6. At this configuration, these features yield a promising comparison with the baseline MFCC features (17.1%). It is interesting to note that the diarization system which models the features using Gaussian distributions is suitable for the proposed features as well.

B. Deep neural architecture

We now analyze the parameter selection issues associated with the deep neural architecture, namely, input domain, bottleneck size, and filtering domain.

The phoneme and the reconstruction MLPs were trained on the TIMIT train dataset. Using these MLPs, filtered squared magnitude vectors were obtained on the MDM development data (RT06 eval). MFCC representation was obtained from the squared magnitude vectors and the ICSI diarization system was used to analyze the features.

Figures 8 and 9 illustrate the effect of bottleneck size versus speaker error rates on the development data. The input features are squared magnitude and MFCC vectors, respectively. The size of the reconstruction MLP was varied as well. All the other parameters of the phoneme MLP and the reconstruction MLP were unchanged during the experiments.

1) Squared magnitude input: For the experiments in Figure 8, the input to the phoneme MLPs was 257 dimensional squared magnitude vectors. The output of the reconstruction MLP was 257 dimensional squared magnitude vectors as well. We varied the bottleneck sizes from 10 to 40 in steps of 10. This was repeated for 5 different reconstruction layer sizes from 600 to 1400, in steps of 200. Preliminary experiments indicated that 1000 nodes to be a reasonable choice for the first and the third expansion layers of the phoneme MLP.

From Figure 8, it can be observed that, in general, for all reconstruction layer sizes, a bottleneck layer size of 20 units seems to yield the lowest speaker error rates. When the number of units are higher or lower, the speaker error increases. A similar trend was observed for a 5 layer MLP architecture in [29]. We could infer that a bottleneck size of 20 units is sufficient to capture phoneme information using a bottleneck architecture. With a larger bottleneck, some speaker information could be captured.

Furthermore, the "optimal" size of the expansion layer in the reconstruction MLP is around 800 units. In general, for either more or less number of units, we observe an increase in the speaker errors for the other bottleneck sizes. Intuitively, the reconstruction MLP is trying to reconstruct the input largely with only the phoneme information. Consequently, it is understandable that it requires fewer units (H3) than the first expansion layer (H1) of the phoneme MLP.

We remark that the deep neural network features obtained from the system with a bottleneck size of 20 yields a performance of 16.5% on the MDM development data, which represents a gain of 0.6% over the baseline MFCC features.

2) MFCC input: We now examine Figure 9. For this plot, the input of the phoneme MLP was 19 dimensional MFCC vectors. The output of the reconstruction MLP was 257 dimensional squared magnitude vectors. For these set of experiments, we only investigated 2 different bottleneck sizes: 20 and 40. This was however, repeated for 5 different reconstruction layer sizes from 600 to 1400, in steps of 200. Similar to previous experiments, preliminary experiments indicated that 1000 nodes to be a reasonable choice for the first and the third expansion layers of the phoneme MLP.

For squared magnitude input space, it can be observed that, for all reconstruction layer sizes, a bottleneck layer size of 20 units seems to yield better performance than 40 units. Interestingly, the optimal size of the expansion layer in the

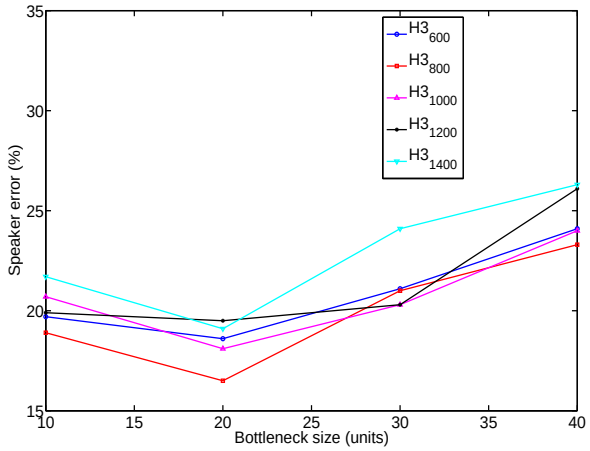

Fig. 8. Performance of the deep neural network on the development data. Bottleneck size (B - in terms of number of units) versus speaker error rates (%) for 5 different reconstruction layer sizes (H3). The input features are squared magnitude vectors.

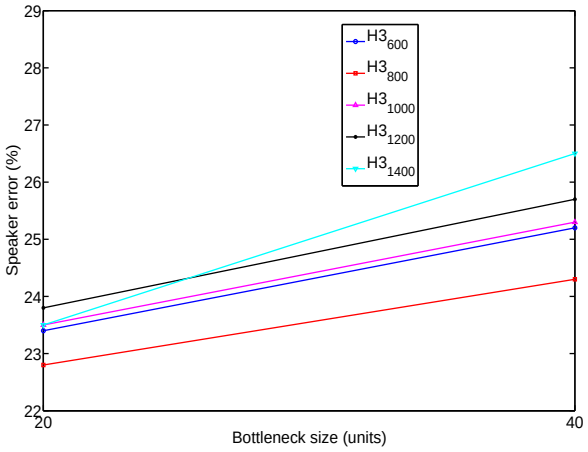

Fig. 9. With input features as MFCC, performance of the deep neural network. Bottleneck size (B - in terms of number of units) versus speaker error rates (%) for 5 different reconstruction layer sizes (H3).

reconstruction MLP again appears to be 800 units.

3) Filtering domain: We performed studies on MFCC being the output of the reconstruction MLP. Unfortunately, the results were not satisfactory. Since the objective of the paper was not to optimize all the parameters of the proposed deep neural architecture, but to analyze the feasibility of the architecture itself, we chose not to delve into the details of why MFCC may not be the optimal filtering domain.

4) Selected deep neural architecture: In conclusion of the analysis in this section, we choose the deep neural architecture with log-squared magnitude input (257-dimensional input), 1000 units for the first expansion layer of the phoneme MLP, 20 units for the bottleneck layer, 1000 units for the second expansion layer of the phoneme MLP, and 800 units for the expansion layer of reconstruction MLP. The output is a 257-dimensional log-squared magnitude input.

VII. Diarization Results on RTeval07

Recall that we use RTeval07 as the evaluation dataset. The results of diarization experiments on MDM and SDM conditions are reported followed by results on phoneme recognition. The relationships suggested by feature analysis is then analyzed.

A. Baseline MFCC

Table II lists the performance of the baseline diarization system RT07 MDM and SDM evaluation data. The performance of the speech/nonspeech detection system on the RT07 evaluation dataset is 3.7%. On RT07 evaluation data, we observe an even higher performance difference for the MFCC features between the SDM and the MDM, with the actual difference being 4.8%.

TABLE II
RT07 evaluation data: The first 3 columns list the performance of the speech/nonspeech detection while the next 2 columns report the performance of baseline MFCC features for MDM and SDM.

Evaluation	Miss	FA	sp/nsp	Spkr err (%) MDM	Spkr err (%) SDM
RT07	3.7	0.0	3.7	**6.4**	**11.2**

B. Comparison with MFCC on RT07 MDM

Table III lists the diarization performance of the privacy-sensitive features against the baseline MFCC features in terms of speaker error in both MDM and SDM conditions. As part of notation, LPR8 denotes LP residual represented using MFCC, while SB denotes subband information from 2.5 kHz to 3.5 kHz and SS denotes spectral slope. DeepNN is used to denote the deep neural architecture summarized in Section VI-B4, whose phoneme and reconstruction MLPs are trained on TIMIT train data.

It can be observed that the baseline MFCC features yield the best speaker errors on MDM and SDM conditions. MFCC features in combination with TDOA features on the RT07 MDM evaluation data yielded a speaker error of 10.9%.

TABLE III
RT07 evaluation data: Performance of 8^{th} order LP residual and deep neural network based features. LPR8 denotes LP residual represented using MFCC. SB denotes subband information from 2.5 kHz to 3.5 kHz, while SS denotes spectral slope.

Features	Spkr err (%) MDM	Spkr err (%) SDM
MFCC (baseline)	6.4	11.2
LPR8	12.9	12.0
LPR8 + SB	11.9	11.9
LPR8 + SS	11.3	12.2
LPR8 + SB + SS	11.0	11.5
DeepNN	14.5	13.9

On the MDM condition, the speaker error of the 8^{th} order LP residual using MFCC representation is about 5% below the baseline. This drop in performance is similar to the drop that was observed on the development data. On the MDM development data, combination with spectral slope

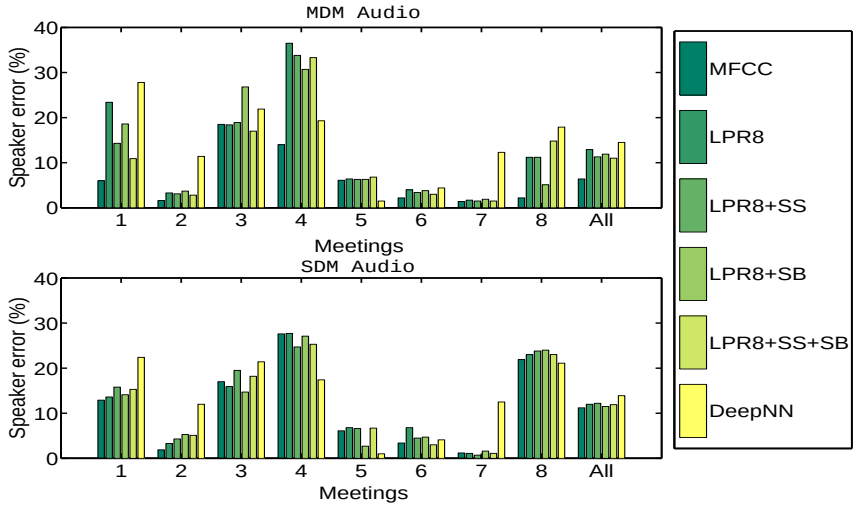

Fig. 10. Meetingwise analysis of the 9 meetings in the RT07 evaluation dataset. The upper plot shows the comparison on the MDM audio while the lower plot shows the comparison the SDM audio. The meeting numbers correspond to the first column in Table IV.

and subband energies yielded a performance gain of nearly 4%, bringing the speaker error close to that of the MFCC features. On the RT07 MDM data, while a gain of nearly 2% is observed, however, the difference with the MFCC features is still 4.6%.

It can be observed from Table III that DeepNN yields a performance of 14.5% on the RT07 MDM data. This represents a performance drop of nearly 8% in comparison with the MFCC features. This is in contrast to the performance of the deep neural network on the development data where it yielded a gain, albeit small, of 0.6% absolute.

C. Comparison with MFCC on RT07 SDM

Table III also lists the performance of the proposed features on the RT07 SDM condition. On SDM condition, however, the performance of the MFCC feature drops significantly. This results in a much smaller difference in speaker error between the MFCC features and the LP residual based features (0.8%). Adding spectral slope to the residual based features does not yield a gain. Similarly, adding subband information between 2.5 kHz to 3.5 kHz does not improve the performance. On the other hand, adding both spectral slope and the subband energies also yields a small gain of 0.5%. The performance of the residual based feature set is robust to SDM conditions and compares well with the baseline MFCC features, yielding a difference of 0.3%.

From Table III, it can be seen that DeepNN yields a performance of 13.9% on the RT07 SDM data. This represents a performance drop of 2.7% in comparison with the MFCC features. The results also show the deep neural network features to be robust to SDM conditions.

D. Meetingwise comparison

We now analyze the RT07 evaluation dataset more closely. Table IV presents a summary statistics of the dataset. The average length of the 8 meetings is 43 minutes. The longest meeting is NIST-20051104-1515 with a length of around 70 minutes, while the shortest meeting is VT-20050408-1500, with a length of 25 minutes. In almost all meetings there are 4 speakers, with the exception of NIST-20060216-1347 and VT-20050408-1500, where there are 6 and 5 speakers, respectively.

TABLE IV
Statistics of the RT07 evaluation dataset.

S.No	Meetings	Length minutes	Speakers	Turns
1	CMU-20061115-1030	41	4	758
2	CMU-20061115-1530	29	4	708
3	EDI-20061113-1500	50	4	873
4	EDI-20061114-1500	48	4	557
5	NIST-20051104-1515	70	4	650
6	NIST-20060216-1347	47	6	630
7	VT-20050408-1500	25	5	508
8	VT-20050425-1000	35	4	726

Figure 10 compares the speaker error rates of the proposed features on RT07 MDM and SDM for each meeting in the RT07 dataset. The upper plot shows the comparison on the MDM audio while the lower plot shows the comparison the SDM audio. There are 9 blocks of results. The 8 meetings in the evaluation dataset correspond to the first eight blocks. The ninth block corresponds to the overall speaker error rate over the entire dataset.

On the MDM dataset, the performance gain of the MFCC features in terms of the overall speaker error rate also translates to gains over individual meetings. However, on meetings that are more than 47 minutes, the performance of the LP residual based feature set compares as well as or better than the MFCC features. It is interesting to note that the best performance of the DeepNN system is on the longest meeting.

On the SDM dataset, the performance of the MFCC features drops substantially over meetings 1, 4, and 8 from Table IV. On the other meetings, MFCC exhibits a more stable behavior. We note that on the SDM dataset, the residual based feature set

compares well with the MFCC features over all the meetings. Furthermore, the performance of the residual based feature set does not drop from MDM to SDM. In fact, some gains can be observed on meetings 1, 4, and 8. On the other hand, while DeepNN system is is still worse than MFCC and the residual based feature set, it yields substantial gains over the longest meeting.

E. Obfuscation method

In Section IV-A, we mentioned another strategy that can be gainfully employed for improving privacy of audio features. In this section, we present speaker error rates of MFCC and LPR8 features that are randomized with block sizes ($N = 1, 5, 9, 13$) on the RT07 MDM evaluation dataset in Table V. In the table, "Randx" is used to denote randomization with block size x frames. We note that randomizing the MFCC features with various block sizes does not change the performance significantly ($\leq 1\%$). Similarly, the performance of the LP residual remains unaffected by local temporal randomization.

TABLE V
Effect of randomization on MFCC and LPR8 on the RT07 MDM dataset. Randx is used to denote randomization with block size of x frames.

Feature	LPR8 (%) Spkr err	MFCC Spkr err
Rand5	13.4	6.7
Rand9	13.8	7.1
Rand13	13.7	6.8

VIII. Analysis of privacy

So far we have investigated LP residual and deep neural network based features for speaker diarization. We now proceed to make an analysis of the privacy aspects.

To our knowledge, quantitatively analysis of audio features for privacy has not been studied before in the literature. Wyatt et al. [3] and [11] indicate that the main concerns with respect to privacy in audio are the reconstructibility of an intelligible speech signal and of the linguistic information. In this paper, we explore two possible ways to analyze this notion of privacy: human speech recognition rates (HSR) of the synthesized speech from the privacy-sensitive features and automatic speech recognition (ASR) rates using the privacy-sensitive features. ASR accuracies are generally reported in the literature using phoneme recognition rates or word recognition rates. Since the latter is more complex for assessing privacy due to the differences in vocabulary sizes, dictionaries, and language models, we prefer phoneme recognition studies.

A. Analysis using human speech recognition

One way to assess privacy in audio is to estimate the intelligibility of speech synthesized from features. In the field of HSR, one aspect of the test is whether the vocabulary is open set or closed set. Another aspect of these studies is whether one tests on individual units such as nonsense syllables or on fully-formed sentences. Furthermore, fully-formed sentences could be semantically meaningful sentences such as conversations, news, phonetically confusable sentences, or semantically unpredictable sentences

TABLE VI
20 semantically unpredictable sentences in the dataset.

No.	Sentence
1	The dust leaned through the broad hat.
2	The task joined the staff that coped.
3	The pure word cleaned the mind.
4	When does the flow guide the blue front?
5	Use the length or the export.
6	The youth knelt with the fresh state.
7	The road dared the growth that slipped.
8	The large wine blamed the store.
9	How does the thing cut the true wall?
10	Bear the truth and the pool.
11	The foot gazed under the dead spring.
12	The suspect mixed the pain that crept.
13	The nice block paid the blood.
14	Why does the jazz hit the brown bar?
15	Bite the book and the stress.
16	The health went down the dark square.
17	The dog built the wife that walked.
18	The good man marked the tree.
19	Where does the post need the poor race?
20	Export the son or the firm.

In this study, we used open set, semantically unpredictable sentences (SUS) [37]. This is done so that the test evaluates only the acoustic aspect of intelligibility instead of the cognitive aspect of prediction. SUS are usually constructed from simple grammatical templates.

1) HSR setup: For our experiments, we used the 20 SUS from EMIME bilingual database [38], with a vocabulary size of 88 words. The list of sentences is given in Table VI. In this database, there are 7 female and 7 male native english speakers with different accents. We chose one female and one male speaker, resulting in 10 sentences being spoken by female and 10 being spoken by male speakers. The speech from the close talking microphone, sampled at 22 kHz, was downsampled to 16 kHz.

We generated the following features from this audio: (a) baseline MFCC features; (b) MFCC representation of 8^{th} order LP residual; and (c) MFCC representation of deep neural network features. Upon reconstruction[2], we now have audio from the 3 sets of features for each of the 20 sentences. Since our pool of listeners were mostly non-native in english, we added the raw waveform as the 4^{th} set of audio (or 4^{th} *system*) for the 20 sentences. This is done to estimate the upper bound of performance that can be achieved by non-native listeners.

Because we expected few listeners (and eventually had 27), in the tradeoff between reasonable estimates of intelligibility versus repeating each sentence as few times as possible, we chose the following strategy: we divided the 80 utterances (20 sentences $\times$ 4 systems) into 2 groups of 40 each. Each group of 40 utterances were obtained with a Latin square design to maximize coverage of the four systems and the 20 sentences. In order that listeners do not get used to a predetermined sequence of audio from the 4 systems, we randomized the sequences in both groups. Each listener was assigned to one

[2]We obtained a noise-excited reconstruction from MFCC using the RASTAMAT library: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

of the two groups and she/he listened to 40 utterances with 10 utterance from each system. Each listener, therefore, listened to each sentence twice.

A web-based application was setup so that listeners could listen using their headphones or speakers. After listening, they had to type-in the sentences they heard. They could complete the task in multiple sessions. Listeners were asked to restrict the number of times they could listen to an utterance to a maximum of 5 times. If an utterance was not intelligible after 5 listening tests, they were asked to type "Not intelligible". Out of the 27 listeners who did the test, one was a native english listener.

2) HSR experiments: Before scoring, we preprocessed the listeners' typed-in responses. This was done to ensure that spelling mistakes or punctuation marks do not show up as errors in intelligibility. For example, some listeners used ellipsis or "?" to indicate words they missed. These were removed from the responses. We used the HResults tool [25] to score the number of words correctly recognized. This is the ratio of number of correct words to the total number of words.

The results of scoring the features are listed in Table VII. In addition, we also obtained an ordering of listeners according to the percentage of words correctly recognized. In Table VII, the two rows correspond to the performance of the 4 systems scored over all the listeners, or scored only over the top 10 best performing listeners. The four columns indicate performance corresponding to the 4 systems: (a) raw waveform; (b) reconstruction from MFCC; (c) reconstruction from MFCC representation of 8^{th} order LP residual; and (d) reconstruction from MFCC representation of deep neural network features.

TABLE VII
HSR performance of the 4 systems over all the listeners or over the top 10 best performing listeners. The four columns indicating performance correspond to raw waveform, reconstruction from MFCC, from MFCC representation of 8^{th} order LP residual, and from MFCC representation of DeepNN features, respectively.

	Wav	MFCC	LPR_8^{MFCC}	$DeepNN^{MFCC}$
Total	85.2	71.3	***13.7***	**6.8**
Top-10	91.8	79.4	***28.9***	**16.9**

It can be seen that for both sets of listeners (total, and top-10), listening to the raw waveform yielded the best performance. Reconstruction from MFCC also yielded very good intelligibility, i.e., around 71% for all the listeners and around 79% intelligibility for the top-10 listeners. In general, listening to speech reconstructed from the MFCC representation of 8^{th} order LP residual appears much less intelligible, with around 50 $\%$ to 60 $\%$ drop in intelligibility. This could partially be due to the loss of the first formant, which carries more linguistic information [12]. In addition, there is a further loss in information from LP residual by representing it using MFCC. Deep neural network based features yield the lowest intelligibility, yielding around 7% intelligibility over all listeners and around 17% over the top-10 listeners.

Furthermore, since listeners listen to each sentence twice, some listeners reported that this led to them performing better on systems having lower intelligibility (having already listened to a cleaner version before). On the other hand, the two sequences corresponding to the utterances for each group were randomized and therefore there is no systematic bias towards privacy-sensitive or the non privacy-sensitive systems.

B. Analysis using automatic phoneme recognition

Another approach to assessing linguistic privacy is to study automatic phoneme recognition accuracies for privacy-sensitive and MFCC features. In our experiments, phoneme recognition studies were performed on TIMIT database. Experiments were conducted excluding the 'sa' dialect sentences. The training data consists of 3000 utterances from 375 speakers, cross-validation data consists of 696 utterances from 87 speakers, and the test data set consists of 1344 utterances from 168 speakers. The phoneme set corresponds to the standard set of 39 units [28].

1) Phoneme recognition system: Features are mean/variance normalized across the training data set. A three layered MLP is used to estimate the phoneme posterior probabilities. MLP consists of 1000 hidden units, and 39 output units with softmax nonlinearity, representing the phoneme classes. The input layer uses a temporal context of 9 frames on the features generated at a frame rate of 100 Hz. For all the features studied (baseline MFCC, LP residual with MFCC representation, deep neural network features with MFCC representation), the input to the MLP was 13-dimensional MFCC with delta and acceleration coefficients. The MLP is trained using standard back propagation algorithm by minimizing the cross entropy error criterion. The phoneme recognition experiments are performed using the hybrid HMM/MLP system reported in [18]. The phoneme sequence is decoded using the Viterbi algorithm, where each phoneme is represented by a left-to-right, 3-state HMM, enforcing a minimum duration of 30 ms. The emission likelihood in each of the three states is the same, and is derived from the output of the MLP.

2) Phoneme recognition experiments: Figure 11 plots the recognition accuracies with respect to increasing LP orders using the phoneme recognition system. It can be observed that as the LP order increases the recognition accuracies drop. 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 lower order formants generally carry more linguistic information, one could expect the performance to drop when the LP order is increased.

From Figure 11, we observe that the LP residual with a prediction order of 8, yields around 15% lower phoneme recognition accuracy in comparison with the MFCC features. We remark that the phoneme recognition experiments using simple features proposed in [7], namely, spectral flatness, energy, zero-crossing rate, and kurtosis (*SEZK*)and the features proposed in [3], namely, autocorrelation and relative-spectral entropy (*AH*), with delta and acceleration coefficients, and with a 9 frame context, yielded accuracies of 40.8% and 31.2% respectively. The performance of an 8^{th} order LP residual lies between that of the simple features and the MFCC (68.2%).

Phoneme recognition experiments using the MFCC representation of deep neural network features yielded 48.7%. This

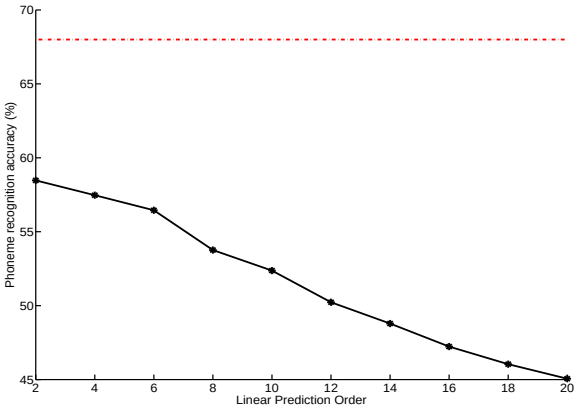

Fig. 11. *Phoneme recognition accuracy for the residual based features various LP orders on TIMIT. The x-axis shows the LP order while the y-axis shows the phoneme accuracy in (%).*

recognition accuracy is much lower than that of 8^{th} order LP residual.

We then performed recognition experiments for the obfuscation method on 8^{th} order LP residual. We note here that randomization can be performed for (a) only test data; or (b) both train and test data with different seeds. The difference between the two stems from the fact that in the second case, the MLP has been trained with noisy targets. While randomized training (29.3%) improves the performance marginally over clean training (28.2%), we still observed a substantial drop in phoneme recognition performance over residual itself. Although our HSR experiments in the previous section showed that reconstructing speech from MFCC representation of 8^{th} order LP residual is unintelligible, this result suggests that randomization can be used to enforce further privacy.

IX. Summary and Conclusion

In this paper we presented two different approaches to privacy-sensitive audio features for robust speaker diarization, namely, LP residual based and and deep neural network based. We systematically investigated both sets of features for speaker diarization in single and multiple distant microphone conditions. The SDM scenario, however, is more relevant to a portable audio recorder scenario. The notion of audio privacy is interpreted as linguistic privacy. We now summarize our key conclusions.

3) LP residual: We studied two different strategies to represent the LP residual, with the MFCC representation of the residual yielding superior performances for all prediction orders. Additionally, we explored the combination of residual with subband information from 2.5 kHz to 3.5 kHz and spectral slope. Although residual features performed slightly less than the conventional MFCC features, we observed that residual features are less affected by the change from MDM to SDM scenarios. Furthermore, residual features proved to be more privacy-sensitive than MFCC features in terms of lower intelligibility and phoneme recognition accuracy.

4) Deep neural network: We utilized a greedy, layer-by-layer trained deep neural network for representing the phoneme information in the short-term spectrum of the signal. A second MLP was utilized to reconstruct the spectrum. The reconstructed spectrum was used as a phoneme filter. In terms of diarization performance, this approach performed slightly worse than the LP residual based approach. However, these features proved to be more privacy-sensitive then residual features. Future work on this approach will investigate improvements such as training the deep MLP on meeting data.

5) Putting privacy and diarization together: We attempted to quantify the abstract notion of privacy in audio through phoneme recognition and intelligibility studies. On the one hand, standard spectral features such as MFCC yielded, not surprisingly, good linguistic reconstruction. Proposed approaches to privacy-sensitive audio feature extraction yielded substantially lower linguistic performance compared to the MFCC features.

While the diarization performance of the LP residual features are similar to the baseline MFCC features, the performance of the deep neural network based features were about 2% lower than MFCC features. 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 [39].

6) Future Work: Nonverbal cues in audio have been explored in developing computational models of face-to-face human behavior. However, most work done in this domain are from meeting room audio. Our future work will utilize the privacy-sensitive audio features in this paper to capture real-world audio. Patterns of speech/nonspeech detection, diarization, and indoor/outdoor classification can then be used to analyze social interactions.

Acknowledgment

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.

References

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

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

[3] D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, "A privacy-sensitive approach to modeling multi-person conversations," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2007.

[4] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "SoundSense: scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the International Conference on Mobile Systems, Applications, and Services*, 2009.

[5] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proceedings of Workshop on Classification of Events, Activities, and Relationships and the Rich Transcription Meeting Recognition*, 2008.

[6] B. P. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 24 – 33, 2007.

[7] 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 *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[8] J. Krumm, "A survey of computational location privacy," *Personal and Ubiquitous Computing*, vol. 13, pp. 391–399, August 2009.
[9] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2003.
[10] "http://www.nist.gov/speech/tests/rt/rt2007/spring/."
[11] D. Wyatt, T. Choudhury, and J. Bilmes, "Conversation detection and speaker segmentation in privacy-sensitive situated speech data," in *Proceedings of Interspeech*, 2007.
[12] R. Donovan, "Trainable speech synthesis," Ph.D. dissertation, Cambridge University, 1996.
[13] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of IEEE*, vol. 63, pp. 561–580, 1975.
[14] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243–1261, 2006.
[15] T. C. Feustel, G. A. Velius, and R. J. Logan, "Human and machine performance on speaker identity verification," *The Journal of the Acoustical Society of America*, vol. 83, pp. 169–170, 1989.
[16] P. Thevenaz and H. Hugli, "Usefulness of the LPC- residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145–157, 1995.
[17] S. H. K. Parthasarathi, M. Magimai.-Doss, D. Gatica-Perez, and H. Bourlard, "Speaker change detection with privacy-preserving audio cues," in *Proceedings of International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction*, 2009.
[18] H. Bourlard and N. Morgan, *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Publishers, 1994.
[19] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2000.
[20] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504 – 507, 2006.
[21] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2007.
[22] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, pp. 1–40, 2009.
[23] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2006.
[24] J. Frankel, D. Wang, and S. King, "Growing bottleneck features for tandem ASR," in *Proceedings of Interspeech*, 2008.
[25] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge University Press, 2000.
[26] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques." *Speech Communication*, vol. 5, pp. 183 – 197, 1986.
[27] F. K. Soong and A. K. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 871 – 879, 1988.
[28] J. Pinto, G. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP based hierarchical phoneme posterior probability estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2, pp. 225–241, 2011.
[29] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2008.
[30] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Proceedings of Advances in Neural Information Processing Systems*, 1990.
[31] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley, 1991.
[32] G. A. Miller, "Note on the bias of information estimates," *Information Theory and Psychology*, pp. 95–100, 1954.
[33] S. Chen. and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA speech recognition workshop*, 1998.
[34] "http://www.nist.gov/speech/tests/rt/rt2006/spring/."
[35] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," in *http://www.icsi.berkeley.edu/xanguera/BeamformIt*, 2006.
[36] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The AMI meeting transcription system: progress and performance," in *Proceedings of Workshop on Machine Learning for Multimodal Interaction*, 2006.
[37] C. Benoit, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, pp. 381 – 392, 1996.
[38] M. Wester, "The EMIME bilingual database," The University of Edinburgh, Tech. Rep., 2010.
[39] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization." *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.