IDIAP RESEARCH REPORT

PRIVACY-SENSITIVE AUDIO FEATURES FOR SPEECH/NONSPEECH DETECTION

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

Idiap-RR-12-2011

MAY 2011

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

Privacy-Sensitive Audio Features for Speech/Nonspeech Detection

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

***Abstract*—The goal of this paper is to investigate features for speech/nonspeech detection (SND) having low linguistic information from the speech signal. Towards this, we present a comprehensive study of privacy-sensitive features for SND in multiparty conversations. Our study investigates three different approaches to privacy-sensitive features. These approaches are based on: (a) simple, instantaneous feature extraction methods; (b) excitation source information based methods; and (c) feature obfuscation methods such as local (within 130 ms) temporal averaging and randomization applied on excitation source information. To evaluate these approaches for SND, we use multiparty conversational meeting data of nearly 450 hours. On this dataset, we evaluate these features and benchmark them against standard spectral shape based features such as Mel Frequency Perceptual Linear Prediction (MFPLP). Fusion strategies combining excitation source with simple features show that comparable performance can be obtained in both close-talking and far-field microphone scenarios. As one way to objectively evaluate the notion of privacy, we conduct phoneme recognition studies on TIMIT. While excitation source features yield phoneme recognition accuracies in between the simple features and the MFPLP features, obfuscation methods applied on the excitation features yield low phoneme accuracies in conjunction with SND performance comparable to that of MFPLP features.**

***Index Terms*—Privacy sensitive features, speech/nonspeech detection**

I. Introduction

THE work reported in this paper takes place in the context of modeling face-to-face interaction patterns using multimodal sensor data [1]. Our work aims to help represent and infer the interactions among people in various formal, semi-formal, and informal settings. Towards this goal, we wish to capture spontaneous, multiparty conversations using portable audio recorders and supplement it with other rich contextual information such as location, movement, and proximity.

For the above purposes, recording and storing raw audio could breach the privacy of people whose consent has not been explicitly obtained [2]. One way to address this privacy issue is to store audio features instead of raw audio, such that neither intelligible speech nor lexical content can be reconstructed [3]. While such audio features may appear to be restrictive, there are different applications that with success use only the nonverbal cues in speech for the study of social behavior [4]. We refer to these features as privacy-sensitive (or privacy-preserving) features. The term "privacy-sensitive" can have different connotations in different areas of computing. Instead of coining a new term, in our case, we decided to follow its use as originally proposed in the speech community by Wyatt et al [3].

As an alternative to storing such audio features, one can directly implement an online speech/nonspeech detection (SND) and a speaker diarization system on a portable device and store information based on the output. A caveat of this method though is that the set of possible tasks using such a *high-level* information is then limited by the output of the diarization system. For example, other sources of information, not including the verbal information, such as emotion, language, location, and the background acoustic scene information are inevitably lost. Another challenge concomitant with such a design choice is the computational limitation imposed by the portable device. Towards this end, a sound sensing framework is proposed for the limited resources available on the Apple iPhone [5].

An issue inherent to capturing spontaneous conversations using portable recorders is the necessity of speech processing techniques, including feature extraction methods, to be relatively robust to microphone distances from speakers. This is in contrast to more conventional speech processing tasks which work either with close-talking or farfield microphones, where the distances are either uniformly close or uniformly far. Furthermore, considering the portability of the recorders and the mobility that it provides the wearer, the features also need to be robust to changes in the ambient environment.

In this context, the full scope of our work aims at investigating robust privacy-sensitive features for tasks such as SND, speaker change detection (SCD) and speaker diarization towards enabling the development of systems for conversation and acoustic scene analysis. Our focus in this paper is on investigating privacy-sensitive audio features for SND, exploring the tradeoff between SND performance and privacy.

One of the more challenging applications of SND is in the context of segmenting meeting room recordings [6] for automatic speech recognition (ASR) and speaker diarization. State-of-the-art systems for such tasks in general use SND systems based on spectral-shape based features. As examples, the ICSI meeting room diarization system [7] and the AMIDA

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

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

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

M. Magimai.-Doss is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland e-mail: mathew@idiap.ch

2009 meeting transcription system [8] use SND based on Mel Frequency Cepstral coefficients (MFCC) and Mel Frequency Perceptual Linear Prediction (MFPLP) features [9] respectively. While such features have been shown to be robust, a potential issue with such features is that both an intelligible speech signal and the lexical content can be reconstructed.

Previous approaches to privacy-sensitive features have focused on either reinterpreting simple, frame-level heuristics ([3], [10]) in the context of conversation analysis or computing long-term averages of standard spectral features such as MFCCs ([2]) in the context of referencing and indexing large personal audio logs. Benchmarking [10] the two sets of simple, frame-level heuristics (henceforth called *simple features*) revealed that the performance of these privacy-sensitive features with explicit temporal modeling is comparable to the standard spectral features such as MFPLP, that do not have the privacy constraint. Our subsequent study focusing on the robustness of these features [11], however, found that there could be a small gap in performance between the privacy-sensitive and the non privacy constrained features in *mismatched* conditions.

In this paper, we investigate two new approaches to privacy-sensitive representations of audio for SND along with the simple features studied in [3] and [10]: (a) excitation source information based methods; and (b) feature obfuscation based methods such as local temporal averaging or randomization. To evaluate these approaches for SND, we use multiparty meeting data of nearly 450 hours. On this dataset, these approaches are then analyzed on close-talking and far-field microphone scenarios, and benchmarked against standard MFPLP features.

The notion of privacy in audio remains something that is difficult to quantify and evaluate. Measures of usability of corrupted speech segments [12] could be interpreted as means to evaluate privacy, with high usability corresponding to low privacy. More recently, studies such as [3] and [13] indicate that the main privacy concerns in audio are the reconstructibility of the linguistic information or an intelligible speech. In this paper, we present phoneme recognition studies as means to evaluate this notion of privacy, with higher recognition accuracy being interpreted as lower privacy. Features such as MFPLP can be considered to be less privacy-sensitive since these features yield state-of-the-art phoneme recognition performance. Similarly, simple features could be interpreted as being more privacy-sensitive. Combinations with excitation source based features and feature obfuscation methods provide privacy comparable to [3] and [10], while yielding state-of-the-art SND performance.

The rest of the paper is organized as follows. Section II reviews the literature on privacy-sensitive features and linear prediction residual based features. Section III summarizes our approach. The dataset definition and the annotations, including the dataset protocol involving the matched, the mismatched, and the cross-validation setups are provided in Section IV. Section V discusses the implementation and the notational details of the SND system, comprising the features, the classifier, and the combination techniques. Parameter selection experiments are discussed in Section VI. We discuss the SND performance and revisit the privacy-sensitive aspects of the features in Sections VII and VIII. Finally, we draw some conclusions in Section IX.

II. Related work

In this section, we present the most relevant work in privacy sensitive audio features and in LP residual.

A. Privacy-sensitive features

One approach to privacy-sensitive audio cues relies on storing certain statistical properties, such as long-term averages (of the order of a minute) of the short-term spectral-based features [2]. This approach was shown to be effective in scene analysis tasks for referencing large personal audio logs [2], [14]. Since speech is perhaps the most informative content in audio logs, and conversation analysis requires a finer temporal resolution of features, in [15] short-term features based on autocorrelation were proposed for robust speech/nonspeech segmentation. These features are meant to be used for detecting speech segments and making them unintelligible before storage.

In the case of automatic conversation analysis, [16] is probably among the earliest work on features. Here a dyadic conversation analysis is performed using nonverbal cues based on short-term autocorrelation and relative spectral entropy. These features were studied with respect to robustness to noise, robustness to microphone distance, and robustness to environment. Work reported in [3] and [13] applied these features to privacy-sensitive, multiparty conversation detection and modeling. The focus of these studies was mainly on modeling the dependencies between speakers in conversations.

More recently in [10], we reinterpreted four other classical short-term SND features as privacy-sensitive features. These features are energy [6], [17], zero crossing rate [6], [17], spectral flatness [18], and kurtosis [6]. We also benchmarked both the sets of privacy-sensitive features ([3], [10]) against standard spectral features (MFPLP) used in [9]. Furthermore, the efficacy of the temporal context for these features was shown, with increases in temporal context yielding improvements in performance. A context of about 500 ms was shown to yield performance comparable to standard spectral features in *matched* conditions.

Motivated by the fact that real-life conversations are often recorded in various environments, the robustness of the privacy-sensitive features were evaluated in *mismatched* conditions against the standard spectral features in [11]. Explicitly modeling the temporal context was shown to be useful in mismatched conditions as well. Further analysis showed that in mismatched conditions, there is a small gap in performance in comparison with the spectral features.

In this paper, to bridge this gap we investigate two new approaches to privacy-sensitive features: (a) linear prediction (LP) residual; and (b) feature obfuscation methods such as local temporal randomization and averaging of features; Obfuscation methods have been used previously in other aspects of privacy in sensor data research [19]. We apply these techniques for privacy in audio.

B. Linear prediction residual

This section begins with a reinterpretation of LP residual as a privacy-sensitive feature. Subsequently, related work on processing the LP residual is examined.

1) Privacy-sensitive reinterpretation: It is generally known that up to two or three formants are required to synthesize intelligible speech or to reconstruct the lexical information [20]. 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 [21].

Linear prediction (LP) analysis of speech [18] assumes the source-filter model and it estimates three components: (a) an all-pole model; (b) a residual; and (c) 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. This approach to privacy-sensitive features was adopted for speaker change detection in [22].

2) Related work on LP residual: Depending on the prediction order, the LP residual contains mostly information about the excitation source of the speakers [23]. It has been shown that humans can recognize speakers by listening to the LP residual signal [24]. Previous works have exploited this. For example, the LP residual has been used as a complimentary feature for speaker recognition in [25], while [23] exploits speaker information in the LP residual at segmental levels (10 - 30 ms) using an autoassociative neural network.

Another property of LP residual is that it has been shown to be relatively robust to additive noise [26]. The Hilbert envelope of the LP residual is processed in [26] using covariance analysis and the periodicity property of this signal was then used in a voice activity detection task.

The importance of long temporal context (≈250 ms) for spectral-shape based features such as MFCC is well known for ASR [27]. This has also been exploited for SND in [9]. In this paper, we investigate whether information at such temporal scales exist in LP residual.

Our work extends these previous works in several ways. Unlike [25] we use LP residual independent of the all-pole model parameters. Secondly, in contrast to [26] and [23] we investigate and then exploit long temporal context in LP residual. A systematic investigation of the LP residual for various prediction orders is conducted for SND. The robustness of the LP residual in farfield microphone data is then evaluated. To the best of our knowledge, the present paper is the first work that exploits LP residual in a privacy-sensitive SND scenario.

III. Our approach

Before we present the SND system and the results, we summarize our overall approach. Figure 1 illustrates this using a block diagram. These blocks are described below.

(a): Evaluating privacy-sensitive features for speech detection entails a comparison of SND performance as well as an evaluation of linguistic privacy. To evaluate SND we construct the scenario using multiparty meeting data, namely the NIST [28], AMI [29], and ICSI [30] databases. Section IV discusses the SND datasets in more detail.

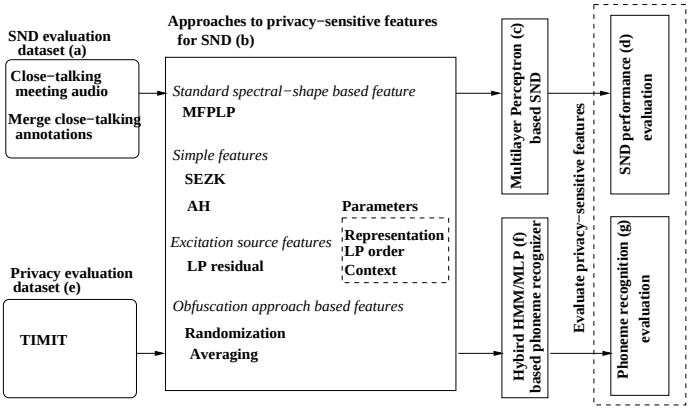

Fig. 1. Block diagram of our approach. A detailed discussion of the figure is provided in Section III.

(b): Privacy-sensitive and the standard spectral features (MFPLP) are derived from these datasets. Some issues with LP residual are the choice of parameters, namely, its representation, the LP order, and the temporal context. Section VI describes parameter selection experiments with these features, their combinations, and the notations in detail.

(c,d): A separate multilayer perceptron (MLP) classifier is trained for each feature set for the speech/nonspeech classification task, similar to [9]. This allows us to compare the privacy-sensitive features with the reference MFPLP features, by way of eliminating the effects of the classifier. MLP classifier is also useful in studying the effect of temporal context. Section V-B provides more details on the MLP classifier, while Section VII presents the SND results.

(e,f,g): The notion of linguistic privacy is quantified using phoneme recognition studies on the TIMIT dataset. These experiments are performed with the hybrid HMM/MLP system [31]. The trained MLP classifiers used for SND are different from the ones used in the hybrid HMM/MLP system. The phoneme recognition results with these features are provided in Section VIII.

IV. SND dataset

An issue in comparing the features is a lack of standard datasets, due to privacy concerns. For this study, we used the scenario that was constructed in [10]. We likened the audio collected by subjects wearing portable audio recorders to a meeting room scenario captured using close-talking microphones. In contrast to the traditional meeting room applications where, given the close-talking microphone signal, the interest generally lies in the speech segments of the wearer ([9], [6]), in conversation analysis, speech segments that are spoken by the other speakers are also of interest. As a consequence of this, crosstalk segments in the meeting room tasks are now considered as speech segments.

A. Dataset and Annotations

The dataset and annotations were used from our setup in [10]. The audio data consists of individual close-talking microphone recordings from meetings. Groundtruths are then derived by merging the speech-activity annotations for the individual microphones, that are closer than a fixed time

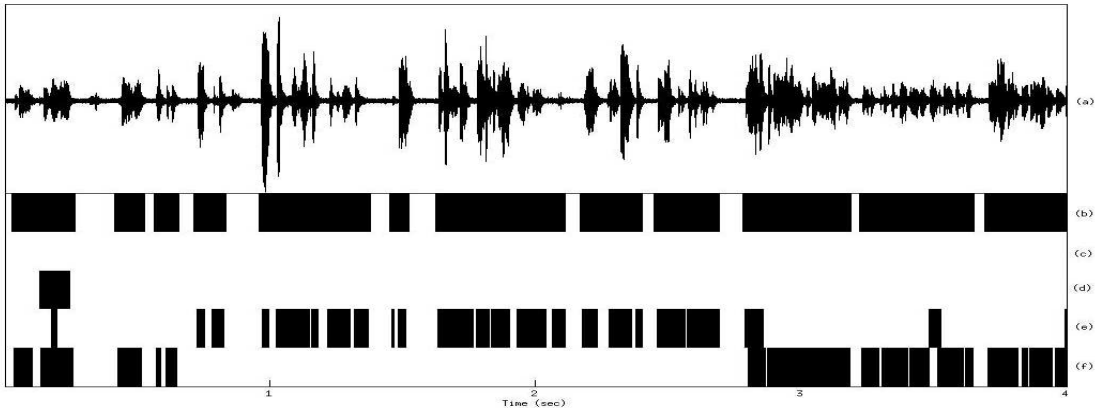

Fig. 2. A close-talking microphone recording of a meeting segment with the speech/nonspeech annotations on the four close-talking microphone channels and their merged annotation. Dark regions indicate speech segments and light regions indicate nonspeech segments. (a) close-talking microphone recording (b) merged annotation using the annotations from all the channels (c) Speaker 1's annotation - appears to be silent in this segment (d) Speaker 2's annotation for the same meeting segment with respect to her microphone recording (e) Speaker 3's annotation for the same meeting segment with respect to her microphone recording - in this case, the signal in Fig. 2(a) was used to produce this annotation (f) Speaker 4's annotation for the same meeting segment with respect to her microphone recording.

interval of 100 ms. Since manual annotations are not consistent ([32]), forced-alignment was used to derive the annotations for the individual microphones. More details on the forced-alignment procedure used to derive the annotations can be found in [9].

A figure illustrating this merging process for a close-talking microphone recording of a meeting room speech segment is shown Figure 2. Each speaker's SND annotation for that meeting segment is done with respect to whether the speaker spoke or not during that segment. During this meeting segment, speakers 1 and 2 (Figure 2(c), (d)) appear to be mostly silent. The annotation corresponding to the wearer of this microphone is shown in Figure 2(e). The merged groundtruth using the process discussed above is shown in Figure 2(b).

The close-talking microphone recordings, sampled at 16kHz, were obtained from NIST [28], AMI [29], and ICSI [30] meeting room data. The total data adds up to 100 hours of meeting speech spanning 120 meetings. The actual amount of individual close-talking recordings adds up to nearly 450 hours with NIST, AMI and ICSI contributing 52, 50 and 350 hours respectively. The training data from NIST, AMI and ICSI amounted to 9, 15 and 48 hours respectively. Using the groundtruth defined above, the overall ratio of nonspeech to speech was around 1:4.2. The amount of near-field speech is considerably less than the amount of far-field speech, with overall ratio of nonspeech: near-field speech:far-field speech being 1.4: 1: 4.8.

B. SND dataset protocol

Using the dataset described earlier, we construct matched, mismatched, and cross-validation conditions. The notations for these conditions are described in Table I. Numbers inside brackets denote the number of hours and the numbers outside denote the notation for that particular dataset.

For a training dataset x and a test dataset y from the table, we use the notation $\{N \text{ or } A \text{ or } I\}\{N \text{ or } A \text{ or } I\}xy$, where N, A, and I correspond to NIST, AMI, and ICSI datasets respectively. The 3 matched setups on NIST, ICSI, and AMI used in [10] are NN14, AA25, and II36 respectively. Similarly the 6 mismatched setups used in [11] are NA15, NI16, AN24, AI26, IN34, and IA35 respectively. The cross-validation setups are AI23 and IA32.

TABLE I
Train and test datasets for matched, mismatched and cross-validation experiments. Numbers in the brackets denote the number of hours and the numbers outside denote the notation for that dataset.

Features	NIST	AMI	ICSI
Train	1 (9)	2 (15)	3 (48)
Test	4 (52)	5 (50)	6 (350)

V. SND SYSTEM

As part of the experimental setup, all SND systems have been constrained to have access to audio from one channel only. This section discusses the implementation and the notational details of the features, followed by the MLP classifier. Combinations of classifiers and features are discussed next.

A. Features

All the features are extracted by pre-emphasizing the signal and then using a 25 ms analysis window with a 10 ms shift.

1) Simple features: The first set of simple features are spectral flatness (*S*), energy (*E*), zero-crossing rate (*Z*), and kurtosis (*K*) [10]. In our implementation of short-term spectral flatness, it is derived as the ratio of the energy of the LP model error (residual) to the energy of the original signal [18]. The energy feature is implemented as short-term log-energy of the signal, while kurtosis feature is implemented as the short-term signal kurtosis. We use *SEZK* and *EZK* to denote the set of all four features and the set of three features respectively.

The features proposed in [3] and [16] are the non-initial maximum of the normalized autocorrelation, the number

of autocorrelation peaks, and the relative spectral entropy. The relative spectral entropy feature is implemented as the Kullback-Leibler divergence between the normalized power spectrum of the current frame and a normalized average of the power spectra of the previous 500 frames [16]. Let *AH* denote this feature set.

Based on our previous works ([10], [11]), the temporal context is fixed at 51 frames and the features are augmented with their first and second derivatives. The dimensionalities of *SEZK*, *EZK*, and *AH* for each frame are 12, 9, and 9 respectively.

2) Linear prediction residual based features: We now look at some issues in using LP residual as features.

(a) Choice of representation of the LP residual: The representations of the residual studied are: a real-cepstrum representation ([25]) with a fixed number of 12 coefficients along with c_0 and a MFPLP representation with 12 coefficients along with c_0. The MFPLP representation is computed using HTK [33]. These features are augmented with delta and acceleration coefficients. Feature selection experiments investigating the choice of representation is presented in detail in Section VI. In either representation, with delta and acceleration coefficients, the dimensionality of the LP residual features for each frame is 39. Delta and acceleration coefficients of LP residual yielded a small gain in performance on the cross-validation data.

(b) 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.

(c) 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).

3) Temporal obfuscation approach: The two obfuscation methods studied are:

(a) 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 (equivalently $2 \cdot (N-1) \cdot 10$ 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.

(b) Local temporal averaging: Feature vectors within block of size ($N = 1, 5, 9, 13$) are averaged. These methods are applied to MFPLP and LP residual based features.

4) Spectral-shape based features (MFPLP): The 12 MFPLP coefficients along with c_0 are computed using HTK. In addition, log-energy and signal kurtosis are extracted. Delta and acceleration coefficients are then appended. In [9], these features were augmented with a set of cross-channel based features. Since we use each microphone channel independently, we drop the cross-channel based features, while we retain all the other features. We use the notation *MFPLP* to denote this feature set. The total dimensionality of this feature set for each frame is 45.

B. MLP based SND Classifier

A separate MLP classifier was trained on each feature set for speech/nonspeech targets based on the groundtruth definition described in Section IV. The minimization of cross-entropy was used as the training criterion. 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 features and MFPLP features were identified by model selection in our previous studies ([10]). For the LP residual features these experiments were conducted on the cross-validation set. These results are summarized in Table II.

TABLE II
Number of input and hidden units for each MLP.

Features	Input	Hidden
Simple features	$51 \times$ dim of feature	200
LP residual features	$\{1, 31, 51, 101\} \times$ dim of feature	50
LP residual with simple features	$51 \times$ dim of feature	100
MFPLP features	$31 \times$ dim of feature	50

C. Classifier and feature combinations

Classifier combination techniques [35] typically combine either the decisions made by the individual classifiers or assign a weight to each classifier's evidence to exploit complementary information. These weights can be either estimated statically or dynamically. In our experiments, we explored one static and one dynamic classifier combination technique: (a) Averaging the weights - static weighting (b) Inverse entropy weighting - dynamic weighting. However, from our earlier paper [11] and from the experiments performed for this paper, it was observed that averaging the weights performed consistently better. So in this paper, for the sake of clarity, we only present our studies on averaging the weights.

Feature-level combinations are also studied to investigate the possibility of exploiting the correlation between features. To this end, feature-level combinations of the LP residual based features and the simple features are investigated.

D. Notations

For the discussions that follow in the remainder of the paper, the notations for the feature sets, the MLP based SND systems, and the combinations are summarized in Table III. In the table, the notation $F(x)$ stands for an MLP based system trained for a feature (or a feature set) *x*. For example, $F(E_1)$ is an MLP based system trained on energy with no context but with delta and acceleration coefficients. Similarly, $F(EZK_{51})$ is an MLP based system trained on energy, zero-crossing rate, and kurtosis with 51 frame context and with delta and acceleration coefficients. To explicitly indicate feature-level combinations of simple features with LP residual based features, we use the notation: $F(x, y)$. For example, $F(LPR8_{51}, EZK_{51})$ denotes a feature-level combination of the individual features *LPR8* and *EZK* using 51 frame context.

TABLE III
Glossary of notations and their definitions.

Notations	Dim	Definition
Feature sets		
EZK	9	energy, zero-crossing rate, and kurtosis (with delta and acceleration coefficients).
SEZK	12	spectral flatness, energy, zero-crossing rate, and kurtosis (with delta and acceleration coefficients).
AH	9	non-initial maximum of the normalized autocorrelation, the number of autocorrelation peaks and the relative spectral entropy (with delta and acceleration coefficients).
MFPLP	45	MFPLP representation of signal with energy and kurtosis and with delta and acceleration coefficients.
LPR8	39	MFPLP representation of 8^{th} order LP residual with delta and acceleration coefficients.
MLP based SND systems based on (individual and combinations of features):		
$F(E_1)$	3	energy with no context (with delta and acceleration coefficients).
$F(Z_1)$	3	zero-crossing rate with no context (with delta and acceleration coefficients).
$F(K_1)$	3	kurtosis with no context (with delta and acceleration coefficients).
$F(E_{51})$	153	energy using 51 frame context (with delta and acceleration coefficients).
$F(Z_{51})$	153	zero-crossing rate using 51 frame context (with delta and acceleration coefficients).
$F(K_{51})$	153	kurtosis using 51 frame context (with delta and acceleration coefficients).
$F(EZK_1)$	9	*EZK* features using no context (with delta and acceleration coefficients).
$F(EZK_{51})$	459	*EZK* features using 51 frame context (with delta and acceleration coefficients).
$F(SEZK_{51})$	612	*SEZK* features using 51 frame context (with delta and acceleration coefficients).
$F(AH_{51})$	459	*AH* features using 51 frame context (with delta and acceleration coefficients).
$F(LPR8_{51})$	1989	MFPLP representation of LP residual of prediction order 8 using 51 frame context (with delta and acceleration coefficients).
$F(LPR8_{51}, EZK_{51})$	2448	MFPLP representation of LP residual and *EZK* features using 51 frame context (with delta and acceleration coefficients).
$F(LPR8_{51}, SEZK_{51})$	2601	MFPLP representation of LP residual and *SEZK* features using 51 frame context (with delta and acceleration coefficients).
$F(MFPLP_{31}, \overline{DA})$	403	MFPLP representation of signal with 31 frame context and without delta and acceleration coefficients.
$F(MFPLP_{31})$	1209	MFPLP representation of signal with 31 frame context and with delta and acceleration coefficients.
$F(MFPLP_{31}, EK, \overline{DA})$	465	MFPLP representation of signal with 31 frame context with energy and kurtosis without delta and acceleration coefficients.
$F(MFPLP_{31}, EK)$	1395	MFPLP representation of signal with 31 frame context with energy and kurtosis and with delta and acceleration coefficients.
$C(LPR8_{51}, EZK_{51})$	1989, 459	combination of $F(LPR8)$ and $F(EZK_{51})$ using equal weights with 51 frame context.
$C(LPR8_{51}, SEZK_{51})$	1989, 612	combination of $F(LPR8)$ and $F(SEZK_{51})$ using equal weights with 51 frame context.
$C(LPR8_{51}, AH_{51})$	1989, 459	combination of $F(LPR8)$ and $F(AH_{51})$ using equal weights with 51 frame context.
$C(LPR8_{51}, EZK_{51}, AH_{51})$	1989, 459, 612	combination of $F(LPR8)$, $F(EZK_{51})$, and $F(AH_{51})$ using equal weights with 51 frame context.
$F(LPR8_{51}^{Ax})$	1989	averaged MFPLP representation of *LPR8* features over a block of x frames using 51 frame context.
$F(LPR8_{51}^{Rx})$	1989	randomized MFPLP representation of *LPR8* features over a block of x frames using 51 frame context (both train and test).
$F(LPR8_{51}^{Cx})$	1989	randomized MFPLP representation of *LPR8* features over a block of x frames using 51 frame context (only test data).

We use $C(x, y)$ to denote a system obtained by combining the output of individual MLP systems based on features x and y using classifier combination. For example, the system $C(LPR8_{51}, EZK_{51})$ does a classifier combination of the individual systems $F(LPR8_{51})$ and $F(EZK_{51})$.

E. SND evaluation measure

For evaluation of SND, we use the area under the receiver operating characteristics (AROC) curve as a metric to evaluate speech detection, as in [6], [10], [11] . The receiver operating characteristics (ROC) curve is plotted by varying the detection-threshold on the posterior probability estimates provided by the SND MLP. A value of 50% for the AROC indicates a random performance and value of 100% indicates a perfect classification. Furthermore, this measure was selected so that the evaluation measure is not biased towards a prior distribution of speech and nonspeech.

VI. Parameter selection for feature extraction based on linear prediction residual

We now conduct studies on the parametrization of LP residual: (a) choice of representation of LP residual; (b) LP prediction order; and (c) effect of temporal context on LP residual. These studies were performed on the cross-validation set, namely, AI23 and IA32. The optimal hyperparameters are fixed for later studies in Section VII and VIII.

A. Representation of LP residual

We study the 2 choices of representations of LP residual discussed in Section V-A2: MFPLP and cepstral representation. Figure 3 shows the comparison between the 2 representations with two different temporal contexts - no context and 51 frames context on the AI23 dataset. It can be observed that MFPLP representation yields a better performance with both temporal contexts. This trend was observed on IA32 dataset as well.

B. Prediction order

We now focus on the MFPLP representation in Figure 3 and investigate the choice of LP order. As the prediction order increases, the all pole model approximates the envelope of the short-time power spectrum better. Consequently, we see a drop in the performance for SND as the prediction order is

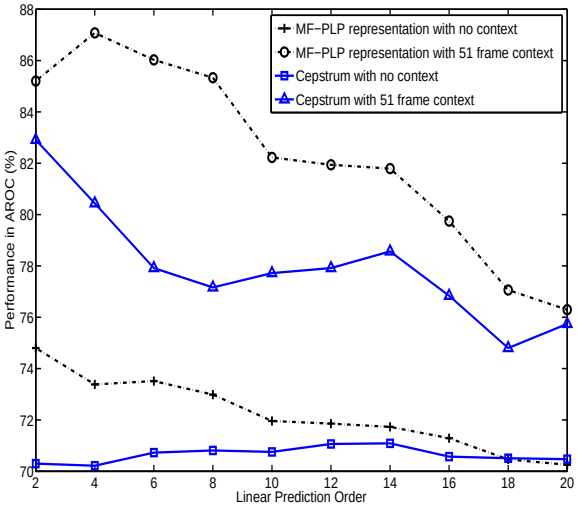

Fig. 3. *Choice of representation of linear prediction residual on AI23 dataset. The two representations of the residual studied in this paper are cepstrum and MFPLP. The x-axis is the linear prediction order and y-axis is the SND performance in area under the receiver operating curve (AROC). This figure also compares the two representations with two different temporal contexts - no context and 51 frame context.*

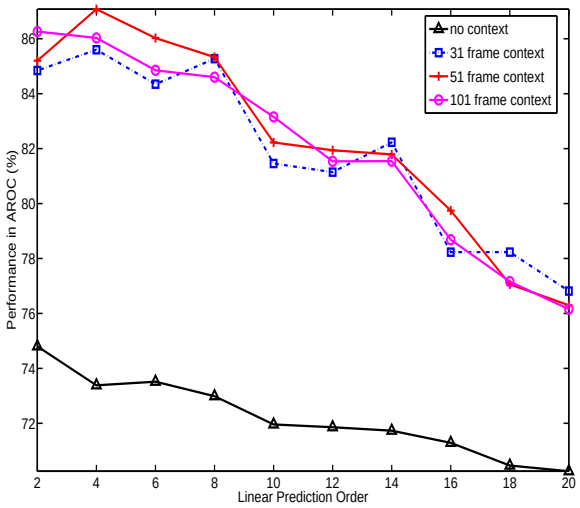

Fig. 4. *Effect of temporal context on the MFPLP representation of linear prediction residual on AI23 dataset. This plot shows four different temporal contexts - no context, 31 frame context, 51 frame context, and 101 frame context. The x-axis is the linear prediction order and y-axis is the SND performance in area under the receiver operating curve (AROC).*

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.

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 14 can model between 2 to 5 formants. Revisiting the performance versus privacy tradeoff, an LP order of 8 seems appropriate for the SND task, since the first two formants are important for synthesizing an intelligible speech signal [20].

C. Temporal context

Figure 4 compares the performances when the temporal context of the LP residual features is increased. This plot shows four different temporal contexts - no context, 31 frame context, 51 frame context, and 101 frame context. A substantial gain in performance can be observed when the temporal context is increased from 1 frame to 31 frames. In general, there is a small gain for most LP orders when the context is increased from 31 frames to 51 frames. An increase in context from 51 frames to 101 frames does not yield any gain. For $F(MFPLP_{31}, EK)$, on the other hand, we observed that the performance saturates at around 31 frames. This observation is consistent with studies in [9]. These trends were observed on IA32 dataset as well.

D. Selected parameters

To conclude this section, we fix the values of the following hyperparameters: (a) LP residual representation is MFPLP; (b) LP order is 8; and (c) Temporal context is 51 frames.

VII. SND Performance Results

This section presents the results for simple features and excitation source features on matched and mismatched conditions. Further analyzes are performed on close-talking and far-field microphone recording scenarios. Feature-level and classifier-level combinations are also investigated. The next section VIII discusses phoneme recognition results to quantify privacy. As a means to enforce stricter privacy on excitation source features in terms of phoneme recognition rates, we then discuss the obfuscation methods.

The results are reported in Table IV for NIST, AMI, and ICSI meeting data. In the discussion that follows, N, A, and I refer to NIST, AMI and ICSI datasets. $A \rightarrow B$ refers to the system being trained on a dataset A and being tested on a dataset B. The dataset protocol, mentioned in Section IV-B, is also mentioned for the respective columns in the table.

The second column lists the overall performance of each system. We observe that the combination of simple features yields benefit over individual systems, with exception of the addition of spectral flatness to $F(EZK_{51})$ [11]. LP residual based systems yield better performance than simple features. However, combinations of simple features with LP residual yield substantial gain in performance. For example, the best performing simple feature based system, $C(EZK_{51}, AH_{51})$, yields 83.4% while the best performing system with LP residual, $C(LPR8_{51}, AH_{51})$, yields 86.3%. Furthermore, we see that this system gives comparable or better performance than $F(MFPLP_{31}, EK)$ (85.0%). We note that the addition of delta and acceleration features, in addition to energy and kurtosis, yields gains to $F(MFPLP_{31}, \overline{DA})$.

We now further analyze the features in both matched and mismatched conditions.

TABLE IV

Performance of features (in percentage of area under ROC) with a context of 51 frames, in matched and mismatched conditions. The second column lists the overall performance of each system. N, A, and I refer to NIST, AMI, and ICSI datasets. $A \rightarrow B$ refers to the system being trained on a dataset A and being tested on a dataset B. The table is grouped into blocks of privacy-sensitive and non privacy-sensitive features. For each column and for each block, the best performance is highlighted in bold. The dataset protocol, mentioned in Section IV-B, is also mentioned for the respective columns in the table.

Features	All datasets	N NN14	A AA25	I II36	N→A NA15	N→I NI16	A→N AN24	A→I AI26	I→N IN34	I→A IA35
	Performance Summary	Matched conditions			Mismatched conditions					
Privacy-sensitive features										
$F(EZK_{51})$	80.5	84.1	90.8	82.0	82.0	75.5	86.0	80.3	82.5	86.7
$F(SEZK_{51})$	79.4	84.0	91.5	81.5	79.7	71.5	86.7	80.6	83.6	87.2
$F(AH_{51})$	81.3	83.3	90.3	85.7	86.0	75.7	85.3	78.9	83.6	88.1
$C(EZK_{51}, AH_{51})$	**83.4**	86.0	91.5	86.2	87.2	78.1	87.5	82.7	85.0	89.1
$F(LPR8_{51})$	84.8	83.0	90.9	89.0	84.5	79.6	83.4	85.3	83.3	87.8
$F(LPR8_{51}, SEZK_{51})$	84.7	86.7	91.3	88.9	85.6	79.0	86.7	84.1	**86.0**	87.6
$F(LPR8_{51}, EZK_{51})$	85.2	86.1	91.1	89.5	84.2	80.2	86.6	84.5	84.9	88.4
$C(LPR8_{51}, SEZK_{51})$	85.0	86.8	**92.1**	88.3	86.9	79.2	87.0	85.3	85.5	89.3
$C(LPR8_{51}, EZK_{51})$	85.4	86.7	91.8	88.6	87.2	81.1	86.9	84.6	85.1	89.1
$C(LPR8_{51}, AH_{51})$	**86.3**	86.1	91.8	**89.8**	88.4	**82.0**	86.4	**86.1**	85.1	89.8
$C(LPR8_{51}, EZK_{51}, AH_{51})$	86.0	**87.5**	92.0	88.9	**88.7**	81.8	**87.8**	85.4	**86.0**	**90.0**
Non privacy-sensitive features										
$F(MFPLP_{31}, \overline{DA})$	81.8	83.0	**91.6**	89.8	82.9	65.6	85.5	**86.8**	84.3	89.7
$F(MFPLP_{31})$	83.6	83.4	91.4	**90.7**	85.3	71.5	85.1	86.4	85.0	**90.2**
$F(MFPLP_{31}, EK, \overline{DA})$	83.0	**84.6**	91.1	87.9	84.9	73.5	**86.5**	84.8	84.3	88.4
$F(MFPLP_{31}, EK)$	**85.0**	84.5	**91.6**	89.9	**87.4**	**77.2**	86.1	86.3	**85.3**	90.0

A. Analysis on matched conditions

From Table IV, it can be seen that the performance of the LP residual based SND system with a context of 51 frames, denoted by $F(LPR8_{51})$ is slightly less than $F(EZK_{51})$, $F(SEZK_{51})$, $F(AH_{51})$ and $F(MFPLP_{31}, \overline{DA})$ for the NIST dataset. On the AMI dataset, all the features are comparable. Whereas, for the ICSI dataset, the LP residual is significantly better (at least 3%) than EZK, $SEZK$ and AH and it is comparable to $F(MFPLP_{31}, \overline{DA})$.

Next, we consider the feature combination studies. Table IV shows that on matched conditions, $F(LPR8_{51}, SEZK_{51})$ and $F(LPR8_{51}, EZK_{51})$ yield superior performance in comparison with $F(EZK_{51})$, $F(SEZK_{51})$, and $F(AH_{51})$. These systems are comparable with the systems based on $MFPLP$ on all the three datasets.

Combining either *AH* or *EZK* features with residual based features through classifier combination scheme yields similar results. In matched conditions combining both *AH* and *EZK* with the residual based features through classifier combination methods does not yield consistent improvements over combinations with just one of the feature sets.

We now analyze the performance of $MFPLP$ features. It can be noted that the addition of delta and acceleration coefficients or energy and kurtosis to $F(MFPLP_{31}, \overline{DA})$ does not increase the performance significantly. In matched conditions it appears that simple spectral based system $F(MFPLP_{31}, \overline{DA})$, is sufficient for state-of-the-art performance.

Finally, the best performance for the privacy-sensitive features on the NIST, AMI, and ICSI datasets are 87.5%, 92.1%, and 89.8% respectively. The best performances achieved by the non privacy-sensitive features on the same datasets are 84.6%, 91.6%, and 90.1% respectively. We see that both sets of features are comparable on matched conditions.

B. Analysis on mismatched conditions

For the mismatched conditions, it can be seen that the LP residual based SND system is generally better than $F(EZK_{51})$ and $F(SEZK_{51})$. The comparison with $F(AH_{51})$ and $F(MFPLP_{31}, \overline{DA})$ is more mixed for $F(LPR8_{51})$.

Combining LP residual with *SEZK* at feature-level yields a small, if any, gain in performance. Comparison between $F(LPR8_{51}, EZK_{51})$ and $F(LPR8_{51}, SEZK)$ systems yields mixed results. Similar to matched conditions, $F(LPR8_{51}, SEZK_{51})$ and $F(LPR8_{51}, EZK_{51})$ yield superior performance in comparison with $F(EZK_{51})$, $F(SEZK_{51})$, and $F(AH_{51})$.

In contrast to feature combination methods, the classifier combination methods typically yield a bigger and a more consistent gain. Furthermore, from Table IV, we observe that, similar to feature combinations, $C(LPR8_{51}, EZK_{51})$ yields mixed results in comparison with $C(LPR8_{51}, SEZK_{51})$. This shows that the addition of the spectral flatness measure does not add significant complementary information to the classifier. Combining either *AH* or *EZK* features with residual based features through classifier combination schemes yields similar results.

Unlike the matched conditions, combining *AH* with residual based features appears to be better than combining *EZK* with residual based features through classifier combination methods. For example, on the A→I column, $C(LPR8_{51}, AH_{51})$ yields a performance of 86.1% while $C(LPR8_{51}, EZK_{51})$ yields a performance of 84.6%. Furthermore, unlike the matched conditions, combining all the three privacy-sensitive systems through classifier combination methods, yields, in general, a more consistent gain in performance than combining just two of them.

Regarding the performance of the spectral-shape based fea-

TABLE V

SND performance analysis (in percentage of area under ROC) in matched and matched conditions (AA25, II36). The table is grouped into blocks of privacy-sensitive and non privacy-sensitive features. For each column and for each block, the best performance is highlighted in bold. The dataset protocol, mentioned in Section IV-B, is also mentioned for the respective columns in the table.

	AMI AA25		ICSI II36		A→I AI26		ICSI dataset
	Matched conditions				Mismatched conditions		Performance Summary
Features	Close-talk	Far-field	Close-talk	Far-field	Close-talk	Far-field	Far-field
Privacy-sensitive features							
$F(E_1)$	89.7	74.6	92.4	68.4	92.0	68.4	68.4
$F(K_1)$	89.6	75.7	92.7	68.2	90.8	68.5	68.4
$F(Z_1)$	64.3	61.9	54.9	51.0	57.0	51.9	51.5
$F(E_{51})$	94.1	86.0	96.0	73.5	92.8	72.5	73.0
$F(K_{51})$	94.3	86.9	95.5	74.4	92.2	72.5	73.5
$F(Z_{51})$	88.9	80.0	81.8	67.1	82.0	61.0	64.1
$F(EZK_1)$	90.3	77.8	92.0	70.4	89.7	71.0	70.7
$F(EZK_{51})$	95.3	90.2	95.0	79.4	93.4	78.0	78.7
$F(AH_{51})$	94.9	89.8	96.1	84.1	91.6	77.2	80.7
$F(LPR8_{51})$	95.1	90.4	96.1	87.8	94.5	83.7	85.8
$C(LPR8_{51}, AH_{51})$	96.0	91.4	**97.5**	**88.5**	**96.1**	**84.3**	86.4
$C(LPR8_{51}, EZK_{51}, AH_{51})$	**96.2**	**91.7**	97.4	87.5	**96.1**	83.6	85.6
Non privacy-sensitive features							
$F(MFPLP_{31}, \overline{DA})$	**95.1**	91.3	**95.4**	85.3	**95.4**	**85.3**	85.3
$F(MFPLP_{31}, EK)$	94.8	**91.4**	**95.4**	**88.9**	93.2	85.1	87.0

tures, it can be noted that the addition of delta and acceleration coefficients to MFPLP coefficients yields a more consistent gain than in the matched condition case. Adding energy and kurtosis, also in general, yields improvements. The addition of delta and acceleration in conjunction with energy and kurtosis also yields a consistent gain in performance.

C. Analysis on close-talking and far-field microphones

In order to gain better understanding, we further analyze the features with respect to close-talking and far-field microphones. In general, we expect the close-talking data in the matched conditions to be the easiest, while far-field data in the mismatched conditions to be the hardest. This was done not only to evaluate the privacy-sensitive features in these conditions, but also to investigate if the performance gains due to temporal context and due to feature/classifier combinations are consistent under all conditions.

To perform this analysis, the two-class groundtruth for SND on the test set was split into a three-class groundtruth: close-talking speech, far-field speech, and nonspeech. Close-talking groundtruths corresponding to the close-talking microphones were used for generating the three-class groundtruths. ROC curves are plotted for $\{$close-talking speech, nonspeech$\}$ and $\{$far-field speech, nonspeech$\}$, and the area under the ROC (AROC) is computed.

1) Analysis on matched conditions: The results are listed in Table V. It can be observed from the table that for all single features such as energy, zero-crossing, and kurtosis, the increase in performance due to an increase in context is more significant in the far-field case than the close-talking case. As an example, for energy, due to the increase in temporal context, the gain in performance is nearly 12% in the far-field case on the AMI dataset, whereas, for the close-talking case, the gain due to increase in context is less than 5%. Similar trends can also be observed for ICSI dataset for the single features. Furthermore, even when no context is used, combinations of single features yield a bigger gain for the far-field case than the close-talking case.

Next, we analyze the performance of systems based on spectral-shape based features. As we had noticed in previous experiments [10], in comparison with AMI meetings, ICSI meetings were recorded in a larger meeting room with speakers being farther apart. This results in the signal-to-noise ratio (SNR) of the speech signal of a speaker who is farther from a close-talking microphone to be lower. We had hypothesized that spectral features such as $MFPLP$ handle this case more effectively. This is indeed observed to be true when we compare $F(MFPLP_{31}, \overline{DA})$ (85.3%) with $F(EZK_{510})$ (79.4%) and $F(E_{510})$ (73.5%) using the far-field scoring, for ICSI dataset in the matched conditions.

We also observed that $F(LPR8_{51})$, while performing at similar performance levels to $F(EZK_{510})$ on AMI near and far-field evaluations performs significantly better when evaluated on ICSI far-field dataset. Furthermore, we see that the LP residual has complementary information compared to $F(AH_{51})$. Also, LP residual performs similar to $MFPLP$ features on the ICSI dataset using the far-field scoring in matched conditions. Lastly, in matched conditions, for the far-field scoring, combining both *AH* and *EZK* with residual based features through classifier combination methods does not yield consistent improvements over combinations with just *AH*.

2) Analysis on mismatched conditions: Table V also presents the results for the far-field and the close-talking cases in mismatched conditions. From the table, we observe a similar trend for a single feature such as energy, wherein there is an increase in performance due to an increase in context for the far-field scenario. But it is interesting to note that when there is no context, the performance of $F(EZK_1)$ is similar or slightly worse than $F(E_1)$ and $F(K_1)$ for the close-talking case, while it is better than all the three single features for the far-field scenario. On the other hand, when there is a temporal support

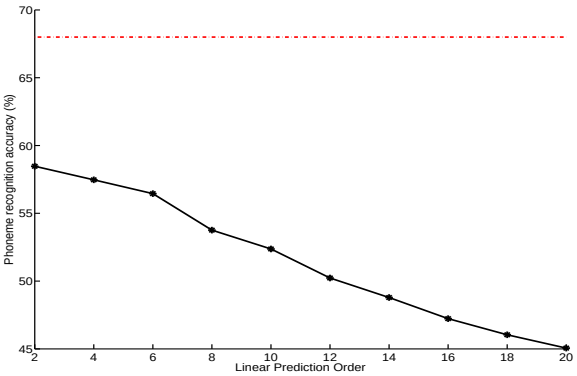

Fig. 5. *Phoneme recognition accuracy for the residual based features various LP orders on TIMIT. The x-axis shows the LP order while the y-axis shows the phoneme accuracy in (%). Phoneme recognition accuracy of reference MFPLP features is shown as a red dotted line.*

of 51 frames, $F(EZK_{51})$ is consistently better than energy based and kurtosis based systems for both the close-talking and the far-field case.

We observe that $F(LPR8_{51})$, while performing at slightly better levels than $F(EZK_{510})$ on close-talking evaluations, performs significantly better when evaluated on far-field data. This along with the observations in the matched conditions case, strongly suggests that excitation based features $F(LPR8_{51})$ are robust not only with respect to distance, but also robust with respect to mismatched ambient conditions. This result is supported by robustness studies on LP residual such as [26].

In the A$\rightarrow$ I mismatched scenario we have chosen for this table, the spectral-shape based features yield the best performances in both close-talking and far-field scenarios. We have omitted the other mismatched conditions since the trends were similar. LP residual features in combination with simple features, show performance comparable to $MFPLP$ features in other far-field scenarios.

VIII. Revisiting Privacy

So far we have investigated simple features and LP residual based features. Before we investigate the temporal obfuscation approach, we briefly revisit privacy. To the best of our knowledge, quantitatively benchmarking audio features for privacy has not been studied before in the literature. Some possible ways to benchmark linguistic privacy in audio features could be: (a) human speech recognition rates of the synthesized speech from the privacy-sensitive features (b) subjective assessments of the privacy-sensitivity of features by human subjects (c) automatic speech recognition rates using the privacy-sensitive features. Since synthesizing speech using simple features is not trivial, we prefer ASR studies for quantifying privacy. ASR accuracies are generally reported in the literature using phoneme recognition rates or word recognition rates. The latter is more complex for assessing privacy due to the differences in vocabulary sizes, dictionaries, and language models.

TABLE VI
Phoneme recognition accuracy(%) for MFPLP and LP residual of order 8 for different randomization and averaging block sizes. Linear prediction residual is shown as LPR. Randomization can be performed for (a) only test data - second column or (b) both train and test data with different seeds - next two columns.

Block size (N)	LPR	LPR	MFPLP	LPR
	Clean train Randomized test	Randomized train Randomized test		Averaging
1	53.8	53.8	68.0	53.8
5	42.3	44.1	63.7	50.7
9	33.7	35.1	55.0	45.1
13	28.0	29.1	46.1	39.8
EZK (no randomization): 40.8				
AH (no randomization): 31.2				

A. Dataset for phoneme recognition

Phoneme recognition studies were performed on TIMIT database (4.3 hours), sampled at 16kHz. 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 hand-labeled dataset using 61 labels is mapped to the standard set of 39 phonemes [34].

B. 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, 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 [31]. 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.

C. Privacy as phoneme error rate

Figure 5 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 5, LP residual for a prediction order of 8 has a phoneme recognition accuracy of 53.8%. We remark that the phoneme recognition experiments using simple features, *EZK* and *AH* features, with delta and acceleration coefficients, and a 9 frame context, yielded accuracies of 40.8% and 31.2%

TABLE VII

Performance of averaged/randomized test features (in percentage of area under ROC) with a context of 51 frames, in mismatched conditions. The second column lists the overall performance of each system. N, A, and I refer to NIST, AMI, and ICSI datasets. $A \rightarrow B$ refers to the system being trained on a dataset A and being tested on a dataset B. Rx denotes randomization with a block size $N = x$. The table is grouped into blocks of reference, averaged, and randomized (both train and test or test alone) features. For each column and for each block, the best performance is highlighted in bold. The dataset protocol, mentioned in Section IV-B, is also mentioned for the respective columns in the table.

Features	All datasets	N NN14	A AA25	I II36	N→A NA15	N→I NI16	A→N AN24	A→I AI26	I→N IN34	I→A IA35
	Performance Summary	Matched conditions			Mismatched conditions					
$F(LPR8_{51})$	84.8	83.0	90.9	89.0	84.5	79.6	83.4	85.3	83.3	87.8
$C(LPR8_{51}, AH_{51})$	**86.3**	86.1	91.8	89.8	88.4	**82.0**	86.4	86.1	85.1	89.8
$C(LPR8_{51}, EZK_{51}, AH_{51})$	86.0	**87.5**	**92.0**	88.9	**88.7**	81.8	**87.8**	85.4	**86.0**	**90.0**
$F(MFPLP_{31}, EK)$	85.0	84.5	91.6	**89.9**	87.4	77.2	86.1	**86.3**	85.3	**90.0**
Averaged features										
$F(LPR8_{51}^{A5})$	84.4	82.4	90.8	89.5	84.8	78.6	82.2	84.9	82.0	87.5
$F(LPR8_{51}^{A9})$	84.2	81.4	90.7	89.2	84.5	78.6	81.5	84.9	81.6	87.1
$F(LPR8_{51}^{A13})$	83.9	81.3	90.4	89.1	83.5	78.6	80.8	83.9	81.3	87.3
$C(LPR8_{51}^{A13}, AH_{51})$	**85.9**	**85.6**	**91.7**	**89.9**	**88.1**	**81.2**	**85.9**	**85.4**	**84.6**	**89.6**
Randomized {train + test} condition										
$F(LPR8_{51}^{R5})$	85.0	83.0	90.7	89.3	83.3	80.6	82.9	85.2	83.0	87.5
$F(LPR8_{51}^{R9})$	83.9	82.3	90.4	88.4	82.9	78.9	81.8	84.3	82.1	87.1
$F(LPR8_{51}^{R13})$	83.3	81.3	90.1	88.0	82.3	77.8	81.5	84.0	81.4	86.2
$C(LPR8_{51}^{R13}, AH_{51})$	**85.7**	**85.6**	**91.5**	**89.4**	**87.8**	**81.0**	**86.0**	**85.6**	**84.5**	**89.3**
Clean train condition + randomized test condition										
$F(LPR8_{51}^{C5})$	84.8	83.1	90.5	89.2	84.4	79.9	82.5	85.0	83.1	87.8
$F(LPR8_{51}^{C9})$	84.0	82.3	90.2	88.6	83.7	78.8	81.9	84.1	82.5	87.4
$F(LPR8_{51}^{C13})$	83.1	81.5	89.9	87.8	82.9	77.7	81.1	83.3	81.8	87.0
$C(LPR8_{51}^{R13}, AH_{51})$	**85.6**	**85.6**	**91.5**	**89.5**	**87.8**	**80.9**	**85.9**	**85.2**	**84.6**	**89.6**

respectively. The performance of an 8^{th} order LP residual (53.8%) lies between that of the simple features and the MFPLP features (68.0%).

D. Enforcing stricter privacy requirements

Table VI lists the phoneme recognition accuracies for obfuscation methods on LP residual and MFPLP features for different block sizes. We note here that randomization can be performed for (a) only test data - second column in the table or (b) both train and test data with different seeds - next two columns in the table. The difference between the two stems from the fact that in the second case, the MLP has been trained with noisy targets. It can be observed that randomized training improves the performance and that as the block size N for randomization increases, the performances of LP residual and MFPLP decrease.

Similarly, we observe from the table that local averaging also provides privacy through a decrease in phoneme recognition accuracies as a function of block size, with randomization providing correspondingly lower phoneme accuracies than averaging. For example, LP residual with 13-frame averaging yields 39.8% while LPR with 13 randomization yields 29.1%.

Furthermore, LP residual with a randomization block size of 13 yields a phoneme accuracy of 29.1%, which is much lower than *EZK* and is also lower than *AH*. This shows that while linear prediction (with varying prediction orders) provides a degree of control in the allowing linguistic information (privacy), another approach to control the linguistic information can be exploited through temporal randomization or averaging. From the table, it can be seen that obfuscation methods on LP residual yields lower phoneme recognition accuracies than on spectral-shape based features. For this reason, we investigate these methods (randomization, averaging) on LP residual in the next section.

E. Analysis of SND performance for obfuscation methods in matched condition

Table VII reports results for obfuscation methods. For the sake of quick reference, we repeat the results of the following SND systems from Table IV: $F(LPR8_{51})$, $C(LPR8_{51}, AH_{51})$, $C(LPR8_{51}, EZK_{51}, AH_{51})$, and $F(MFPLP_{31}, EK)$. We now summarize the SND performance under three categories in matched conditions.

(a) Averaging features: Both train and test sets are locally averaged with various block sizes N. It can be observed that for averaging with block sizes N equal to 5, 9, or 13 frames, there is a small drop in performance in comparison with the case where there is no averaging. $C(LPR8_{51}^{A13}, AH_{51})$, denoting the classifier combination of the system trained on a MFPLP representation of 8^{th} order LP residual with 13 frame averaging and the system trained on $F(AH_{51})$, is comparable with the state-of-the-art system, $F(MFPLP_{31}, EK)$.

(b) Randomized {train + test} conditions: In this case, we train randomized features with the correspondingly synchronized groundtruths. The train and test datasets are randomized with different seeds. It can be observed that for 8^{th} order LP residual with a randomization size N equal to 5 or 9 frames, there is no appreciable difference in performance in comparison with no randomization. On the other hand, for a randomization size of 13 frames, there is a small drop in performance.

(c) Clean train condition + randomized test condition: In this case, we use the trained MLP nets on the original unrandomized features with the corresponding unrandomized

groundtruths and test them on the randomized test data. On the NIST and ICSI datasets, there is a drop of about 1.5%, which is not substantial in comparison with the performance drop observed in phoneme recognition. Furthermore, in both this case and in the previous case, combination with *AH* features yields state-of-the-art performance.

F. Analysis of SND performance for obfuscation methods in mismatched condition

From the Table VII, the performance of the features in mismatched conditions for the obfuscation methods is analyzed under the same categories.

(a) Averaging: It can be observed that for averaging with block sizes N equal to 5, 9, or 13 frames, there is a small drop in performance in comparison with the case where there is no averaging, except for the A$\rightarrow$N case where there is a drop of 2.6%.

(b) Randomized {train + test} conditions: Unlike the matched case, for a randomization size of 13 frames, there is, in general, a performance drop of about 2%. On the other hand, the drop in performance for the combination with *AH* features is small (less than 1% in all cases).

(c) Clean train conditions + randomized test conditions: Like the matched case there is a performance drop of little more than 2% in many cases. However, combination with *AH* yields comparable performances to unrandomized case (less than 1% in most cases).

Comparing randomization in SND and phoneme recognition, to normalize the advantage that a larger temporal context provides SND in the randomization case, we increased the temporal context of features for phoneme recognition experiments to as much as 51 frames (performed model selection again for this setup). This only decreased the phoneme recognition accuracies. We therefore conclude that randomization affects phoneme recognition much more (around 30%) than it does SND (around 2%).

The second column lists the overall performance of each system. We observe that, for obfuscation methods in general, there is a drop in SND performance for LP residual features. However, this drop in performance is small: for example, 13 frame averaging yields a drop in performance by 0.9%, and a 13 frame randomization yields a drop in performance by 1.7%. For the LP residual systems combined with simple features, this drop is even lesser.

IX. Final Discussion and Conclusion

Our study investigated three different approaches to privacy-sensitive features for speech/nonspeech detection (SND). These approaches are based on: (a) simple, instantaneous feature extraction methods (b) excitation source information based methods (c) local feature obfuscation methods such as temporal averaging and randomization. To evaluate these features, we used the multiparty conversational meeting data of nearly 450 hours. On this dataset, we evaluated these features and benchmarked them against state-of-the-art spectral shape-based features (MFPLP), on matched and mismatched conditions. To gain further insights, the results were then analyzed for close-talking and far-field microphone scenarios. To quantify the notion of privacy, we conducted phoneme recognition studies on TIMIT. Our investigations suggest the following.

1) Simple features: We evaluated the robustness of two sets of simple privacy-sensitive features: (a) energy, zero crossing rate, spectral flatness measure, and kurtosis. (b) Autocorrelation and spectral entropy based features. Explicitly modeling the temporal context is useful in matched and mismatched conditions. For all single features such as energy, zero-crossing, and kurtosis, the increase in performance due to an increase in temporal context is more significant in the far-field case than the near-field case. Furthermore, combinations of single features yield a bigger gain for the far-field case than the close-talking case. Our studies also show that state-of-the-art performance, comparable to MFPLP features, can be achieved by these simple features for the close-talking scenario.

2) Excitation source information: Characterizing the excitation source information using LP residual, we showed that exploiting temporal support of up to 51 frames can yield significant gains in the performance. The residual based feature, while performing at only slightly better levels than simple features on close-talking evaluations, performs significantly better when evaluated on far-field data. We also observed that excitation based features are robust not only with respect to distance, but also with respect to mismatched conditions. Fusion strategies combining LP residual with simple features show that state-of-the-art performance can be obtained in both matched and mismatched conditions, on close-talking and far-field microphone scenarios.

3) Local temporal randomization and averaging: We investigated the use of local temporal randomization and averaging (up to 130 ms) on the LP residual features. These approaches caused a small drop in SND performance. However, combinations of the randomized or averaged features with simple features yield state-of-the-art SND performance at stricter privacy requirements, defined in terms of phoneme recognition accuracies. These approaches can also be applied to MFPLP features. However, it was noted that it would yield higher phoneme recognition accuracies.

4) Putting privacy and SND performance together: We quantified privacy in audio through phoneme recognition studies on TIMIT. On the one hand, standard spectral features such as MFPLP yielded, not surprisingly, state-of-the-art phoneme recognition accuracies. On the other hand, simple features yielded much lower phoneme recognition accuracies. LP residual based features yielded phoneme recognition accuracies in between the simple features and the standard spectral features, with the LP order determining the actual performance. Local feature obfuscation methods such as temporal randomization or averaging caused a substantial fall in phoneme recognition performance, with randomization yielding lower phoneme accuracies. SND performance, on the other hand, was relatively unaffected by the temporal obfuscation methods. While it is known that the information in the temporal dynamics of the speech signal can be exploited for phoneme recognition, however, for SND the combination of results showing the importance of temporal context and the relative insensitivity

to randomization leads to the conclusion that there is perhaps more information in the statistics of the frames in the temporal support than in the actual temporal dynamics.

5) Future Work: We remark that there is a balance between privacy, SND performance, and computational load. While LP residual and obfuscation based approaches yield SND performance comparable to state-of-the-art MFPLP features, these features incur an extra computational load to ensure stricter privacy. We would like to assess this computational load on a portable device.

In this paper, we have proposed phoneme recognition to investigate the complex issue of assessing privacy in audio. Complementary social acceptability studies are needed to determine reasonable norms on measured phoneme accuracy.

Our earlier work [22] investigated excitation source based features for a privacy-sensitive speaker change detection task. Preliminary experiments applying these methods to speaker diarization show promising performances for both excitation and obfuscation approaches.

ACKNOWLEDGMENT

This work was supported by the Swiss National Science Foundation through the projects MULTImodal Interaction and MULTImedia Data Mining (MULTI) and the National Centres of Competence in Research (NCCR) IM2.

REFERENCES

[1] D. Gatica-Perez, "Analyzing group interaction in conversations: A review," in *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006.

[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] T. Choudhury and S. Basu, "Modeling conversational dynamics as a mixed-memory markov process," in *Proceedings of Advances in Neural Information Processing Systems*, 2004.

[5] 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.

[6] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 84–91, 2005.

[7] 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.

[8] T. Hain, L. Burget, J. Dines, P. Garner, A. E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "The AMIDA 2009 meeting transcription system," in *Proceedings of Interspeech*, 2010.

[9] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *Proceedings of Interspeech*, Pittsburgh, USA, 2006.

[10] 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.

[11] ——, "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.

[12] R. E. Yantorno, "A study of the spectral autocorrelation peak valley ratio (SAPVR) as a method for identification of usable speech and detection of co-channel speech," Speech Processing Lab, Temple University, Tech. Rep., 2000.

[13] D. Wyatt, T. Choudhury, and J. Bilmes, "Conversation detection and speaker segmentation in privacy-sensitive situated speech data," in *Proceedings of Interspeech*, 2007.

[14] D. P. W. Ellis and K. Lee, "Features for segmenting and classifying long-duration recordings of personal audio," in *Proceedings of Workshop on Statistical and Perceptual Audio Processing*, 2004.

[15] K. Lee and D. P. W. Ellis, "Voice activity detection in personal audio recordings Using autocorrelogram compensation," in *Proceedings of Interspeech*, 2006.

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

[17] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 201–212, 1976.

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

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

[20] R. Donovan, "Trainable speech synthesis," Ph.D. dissertation, Cambridge University, 1996.

[21] G. Fant, *Acoustic Theory of Speech Production*. Mouton, Haag, 1960.

[22] 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.

[23] 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.

[24] 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.

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

[26] K. S. R. Murty, B. Yegnanarayana, and S. Guruprasad, "Voice activity detection in degraded speech using excitation source information," in *Proceedings of Interspeech*, 2007.

[27] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 587–589, 1994.

[28] J. S. Garofolo, C. D. Laprun, M. Michel, V. M. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus," in *Proceedings of International Conference on Language Resources and Evaluation*, 2004.

[29] J. Carletta, S. Ashby, S. Bourban, M. Guillemot, M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma, "The AMI meeting corpus," in *Proceedings of Workshop on Machine Learning for Multimodal Interaction*, 2005.

[30] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[31] H. Bourlard and N. Morgan, *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[32] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The Rich Transcription 2006 Spring Meeting Recognition Evaluation," in *http://www.itl.nist.gov/iad/mig//publications/storage_paper/RT06SResults-v07.pdf*, 2006.

[33] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge University Press, 2000.

[34] J. Pinto, G. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP based hierarchical phoneme posterior probability estimator," *To appear in IEEE Transactions on Audio, Speech, and Language Processing*, 2010. [Online]. Available: http://www.idiap.ch/~jpinto/pubs/hierarchy.pdf

[35] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

Sree Hari Krishnan Parthasarathi (S '10) is a 4th year PhD student at the Swiss Federal Institute of Technology at Lausanne (EPFL) and Idiap Research Institute. His doctoral work studies privacy-sensitive audio features in the context of modeling social interactions. He obtained his master's degree from IIT Madras in 2007, with a thesis on the robustness of group delay based methods for speech processing. Prior to his master's, he was with the research division of Honeywell (Bangalore). His research interests include speech processing, machine learning, and data-driven methods in the context of social computing.

Daniel Gatica-Perez (S'01, M'02) received the B.S. degree in Electronic Engineering from the University of Puebla, Mexico in 1993, the M.S. degree in Electrical Engineering from the National University of Mexico in 1996, and the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, in 2001, receiving the Yang Research Award for his doctoral work. He is now a senior researcher at Idiap Research Institute, Martigny, Switzerland, where he directs the Social Computing group. His recent work has developed statistical methods to analyze small groups at work in multisensor spaces, populations using cell phones in urban environments, and on-line communities in social media. He currently serves as Associate Editor of the IEEE Transactions on Multimedia, Image and Vision Computing, Machine Vision and Applications, and the Journal of Ambient Intelligence and Smart Environments. He is a member of the IEEE.

Hervé Bourlard received the Electrical and Computer Science Engineering degree and the Ph.D. degree in Applied Sciences both from "Faculté Polytechnique de Mons", Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute (www.idiap.ch), Full Professor at the Swiss Federal Institute of Technology at Lausanne (EPFL), and Director of a National Center of Competence in Research in "Interactive Multimodal Information Management" (IM2, www.im2.ch). Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI, www.icsi.berkeley.edu) in Berkeley, CA, he is now a member of the ICSI Board of Trustees.

His main research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, augmented multi-party interaction, and distant group collaborative environments.

H. Bourlard is the author/coauthor/editor of 4 books and over 250 reviewed papers (including one IEEE paper award) and book chapters. He is an IEEE Fellow (since 2000) "for contributions in the fields of statistical speech recognition and neural networks". He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of IEEE ICASSP 2002, General Chairman of Interspeech 2003) and on the Editorial Board of several journals (e.g., past co-Editor-in-Chief of "Speech Communication").

Over the last 20 years, Herv Bourlard has initiated and coordinated numerous international research projects, as well as multiple collaborative projects with industries. He also initiated several start-up companies, and is the recipient of a few prestigious entrepreneurship awards.

Mathew Magimai Doss (S '03, M'05) received the B.E. in Instrumentation and Control Engineering from the University of Madras, India in 1996; the M.S. by Research in Computer Science and Engineering from the Indian Institute of Technology, Madras, India in 1999; the PreDoctoral diploma and the Docteur ès Sciences (PhD) from École Polytechnique Fédérale de Lausanne (EPFL), Switzerland in 2000 and 2005, respectively. From April 2006 till March 2007, he was a postdoctoral fellow at International Computer Science Institute, Berkeley, USA. Since April 2007, he has been working as a research scientist at Idiap Research Institute, Martigny, Switzerland. His research interests include speech processing, automatic speech and speaker recognition, statistical pattern recognition, and artificial neural networks.