INTERSPEECH 2009 BRIGHTON

# Investigating Privacy-Sensitive Features for Speech Detection in Multiparty Conversations

*Sree Hari Krishnan Parthasarathi*[1,2], *Mathew Magimai.-Doss*[1], *Hervé Bourlard*[1,2], *Daniel Gatica-Perez*[1,2]

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

## Abstract

We investigate four different privacy-sensitive features, namely energy, zero crossing rate, spectral flatness, and kurtosis, for speech detection in multiparty conversations. We liken this scenario to a meeting room and define our datasets and annotations accordingly. The temporal context of these features is modeled. With no temporal context, energy is the best performing single feature. But by modeling temporal context, kurtosis emerges as the most effective feature. Also, we combine the features. Besides yielding a gain in performance, certain combinations of features also reveal that a shorter temporal context is sufficient. We then benchmark other privacy-sensitive features utilized in previous studies. Our experiments show that the performance of all the privacy-sensitive features modeled with context is close to that of state-of-the-art spectral-based features, without extracting and using any features that can be used to reconstruct the speech signal.

**Index Terms**: Multiparty Conversation, Privacy-sensitive features, Speech detection.

## 1. Introduction

Recently, there has been a growing interest in capturing spontaneous, multiparty conversations, also referred to as personal audio logs, using portable recording devices [1, 2]. The main interest here lies in the analysis of social interactions. However, capturing raw audio could breach the privacy of people whose consent has not been explicitly obtained. An approach to overcome this is to store features instead of audio, such that neither intelligible speech nor lexical content can be reconstructed from these features [2]. This is what we call privacy-sensitive features in this paper [2].

There are different applications that only use nonverbal cues in speech for the study of social behaviors. For example, [3] used nonverbal cues for analyzing dyadic conversations. More recently, a privacy-sensitive approach was adopted to analyze spontaneous multi-person conversations [4].

An important pre-processing step in conversational analysis is to perform speech detection. State-of-the-art speech/nonspeech detection (SND) systems such as [5] utilize spectral-based features. However, with these features both speech and lexical content can be reconstructed. Privacy-sensitive, instantaneous (frame-level) features for modeling conversations have been used in [4, 3]. These features are based on short-term autocorrelation and spectral entropy. Long-term spectral averages have also been used as features for speech segmentation in personal audio recordings [1].

In this paper, we investigate four different, classical short-term features for speech detection by temporal processing of the audio signal (i.e., without estimating the spectrum). These features are energy [6, 7], zero crossing rate [6, 7], spectral flatness [8], and kurtosis [7]. In addition to these four features, we also systematically study the features proposed earlier in [4, 3]. Previous studies have mostly modeled the instantaneous values of these features. In this work, we also model the temporal context of these features. One of the goals of this study is to approach the performance of a state-of-the-art SND system proposed in [5], but only using privacy-sensitive features.

A key challenge in comparing features is a lack of standard datasets, due to privacy concerns. To this end, we describe a way of constructing a scenario close to the personal audio log scenario using multiparty conversational meeting data. On the meeting data setup, we analyze the above features, their combinations, and compare the performance with a state-of-the-art spectral-based feature, namely mel frequency PLP (MF-PLP) coefficients.

Our studies show that energy emerges as the best privacy-sensitive feature among the four studied features without temporal context modeling. The combination of the four features leads to an improvement in the SND performance. Modeling the temporal context yields improvements for all privacy-sensitive features, including the features from [4, 3]. Kurtosis emerges as the best privacy sensitive feature among the four when temporal context is modeled. Furthermore, the combination of the four features with context modeling, or of the features described in [4, 3] can yield performance comparable to the state-of-the-art spectral based features.

The rest of the paper is organized as follows. A definition of the dataset and annotations is provided in Section 2. Section 3 discusses the proposed system in terms of features, classifier, and evaluation measure. The description of the results and the discussion is provided in Sections 4 and 5, respectively. Finally, we draw some conclusions in Section 6.

## 2. Definition of data and annotations

Personal audio logs are collected by subjects wearing portable audio recorders. The placement of the microphone is similar to that of a lapel microphone used in recording meeting room conversations [2]. We identify this to a meeting room scenario captured using lapel microphones. In the context of meeting room applications such as automatic speech recognition and speaker diarization, given the lapel microphone signal, the interest generally lies in the speech segments of the wearer. Previous speech detection studies, such as [5, 7] on meeting data, have focused on this aspect. However, in conversation analysis, speech segments that are spoken by any speaker are also of interest. As a consequence of this, crosstalk in the meeting task is part of speech.

For our study, the lapel microphone recordings from meeting room datasets are used. Unlike previous studies [5, 7], where "individual" lapel ground truth is used to train the SND system, we train the SND system using the ground truth obtained by merging the speech segments from individual lapel ground truths that are closer than a fixed time interval (100ms). These individual lapel ground truths were defined and obtained from [5].

We began our experiments using lapel microphone recordings from NIST [9], AMI [10], ISL [11] and ICSI [12] meeting room data. Initial experiments revealed that due to the directional nature of the microphones used in ISL recordings, this data is unsuitable for speech detection for speakers other than the wearer. Consequently, in our subsequent experiments, we used only NIST, AMI, and ICSI datasets. The training, testing and cross-validation (CV) data from these datasets are identical to the system described in [5]. In all, the total data add up to 100 hours of speech spanned over 120 meetings. And using the ground truth defined above, the overall ratio of nonspeech to speech was 1:4.2.

# 3. Description of SND system

In this section a brief description of the privacy-sensitive and the state-of-the-art features is provided. The classifier used is then discussed. Finally, the evaluation measure used is also detailed.

All the features are extracted by first pre-emphasizing the signal and then by using a rectangular analysis window of length and shift 25 ms and 10 ms, respectively. In addition, we augment these basic features with their first and second derivatives.

## 3.1. Privacy-sensitive features

1. Short-term energy (E): Studies have shown that short-term energy has been one of the most important features for speech detection [6, 7]. Furthermore, studies have shown that long-term information in energy can be exploited for speech detection [13].
2. Short-term zero crossing rate (Z): The zero crossing rate at a frame-level has been a popular feature for voiced/unvoiced/nonspeech classification [6, 7].
3. Short-term spectral flatness measure (S): The short-term spectrum of the nonspeech signal such as wideband noise can be expected to be flatter than the short-term spectrum of the speech signal. Thus a measure of spectral flatness can be useful for SND. By normalizing the spectrum, and viewing it as a probability mass function, entropy can be used as a measure of the flatness of spectrum. Linear prediction analysis can also be used to derive an efficient flatness measure of speech without explicitly estimating the spectrum [8]. The flatness measure is derived as the ratio of the energy in the model error (residual) to the energy in the original signal. We investigate the measure of spectral flatness obtained using the latter approach.
4. Short-term kurtosis (K): Kurtosis is derived from the fourth order moment of a distribution and it measures its "peakedness". Speech samples have been shown to have a flatter distribution and kurtosis has been shown to be useful in detecting speech [7].
5. Features proposed in [3, 4] for privacy-sensitive speech detection (AH) are the non-initial maximum of the normalized autocorrelation, the number of autocorrelation peaks and the relative spectral entropy.

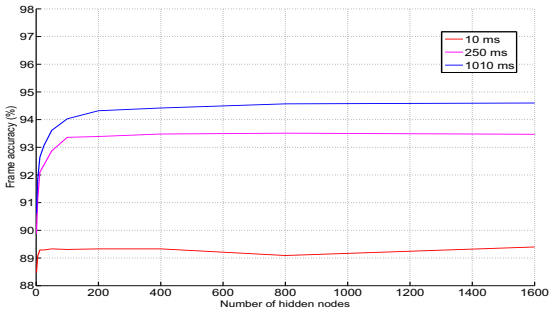


Figure 1: *Number of hidden units vs frame accuracy on NIST cross-validation data for three different temporal contexts.*

## 3.2. Reference spectral-based features

The reference spectral-based features (that is, non privacy-sensitive) are taken from a state-of-the-art SND system [5]. The features consist of 12 MF-PLP coefficients (computed using HTK), and first cepstral coefficient $c_0$, with their delta and acceleration coefficients, in addition to energy and kurtosis. In [5], these 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.

## 3.3. Classifier

The features are analyzed using a multi-layer perceptron (MLP) classifier. The MLP is trained for speech/nonspeech classes based on the ground truth definition described in Section 2, using two output units and minimizing the cross-entropy criterion.

In order to study the effect of temporal context modeling for privacy-sensitive features, we vary the temporal context at the input of the MLP from 0 to 401 frames (i.e., from 10 ms to 4010 ms). Varying the number of MLP hidden units from 1 to 1600 clearly shows (on CV data) that about 200 units yields optimal performance, independently of the privacy-sensitive feature being used. Fig. 1 illustrates this for feature combination (S + E + Z + K). Based on this observation, we use 200 hidden units for all the privacy-sensitive features.

The reference features were analyzed with a trained MLP using 31 frame context (310 ms) as the input layer and 50 units in the hidden layer, as done in [5]. The selection of the training, testing, and CV data follows the procedure described in [5]. Finally, 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. It is noted that all the features were augmented with delta and acceleration features because experiments on CV data showed that using first and second derivatives improved the performance of all the features.

## 3.4. Evaluation measure

For evaluation, we use the area under the receiver operating characteristics (ROC) curve as a metric to evaluate speech detection, as in [7] . The ROC curve is plotted by varying the detection-threshold on the posterior probability estimates provided by the MLP. A value of $50\%$ for the area under ROC 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.

Table 1: *Performance (in percentage of area under ROC) on NIST data for all studied features. The best performance by AH is highlighted in bold and italics while the best performances by EZK and SEZK are highlighted in bold.*

| context(ms) | E | Z | S | K | EZ | EZK | SEZK | AH |
|---|---|---|---|---|---|---|---|---|
| 10 | 73.5 | 67.3 | 70.6 | 51.9 | 74.8 | 75.4 | 73.4 | 74.9 |
| 250 | 79.8 | 78.1 | 79.8 | 78.8 | 81.0 | 82.2 | 80.9 | 83.0 |
| 510 | 80.1 | 78.8 | 80.5 | 81.5 | 82.6 | **83.1** | 81.2 | ***83.3*** |
| 1010 | 81.1 | 78.8 | 80.1 | 81.9 | 81.6 | 82.6 | **81.4** | 82.7 |
| 2010 | 79.7 | 78.3 | 79.6 | 80.7 | 79.8 | 80.3 | 78.5 | 81.3 |
| 4010 | 77.1 | 76.1 | 78.3 | 79.5 | 76.4 | 80.0 | 77.6 | 79.7 |
| MF-PLP + Energy + Kurtosis (reference features): 83.0 | | | | | | | | |

Table 2: *Performance (in percentage of area under ROC) on AMI data for all studied features. The best performance by AH is highlighted in bold and italics while the best performances by EZK and SEZK are highlighted in bold.*

| context(ms) | E | Z | S | K | EZ | EZK | SEZK | AH |
|---|---|---|---|---|---|---|---|---|
| 10 | 77.5 | 62.2 | 72.2 | 52.2 | 77.4 | 79.4 | 79.7 | 79.8 |
| 250 | 85.4 | 79.7 | 82.5 | 85.6 | 88.3 | 89.4 | 90.1 | 89.7 |
| 510 | 87.2 | 81.5 | 84.7 | 87.9 | 90.4 | **91.1** | 91.5 | ***90.3*** |
| 1010 | 88.2 | 81.7 | 85.1 | 88.6 | 90.4 | 91.1 | **91.6** | 90.2 |
| 2010 | 88.4 | 81.7 | 84.9 | 88.5 | 89.3 | 90.6 | 91.0 | 89.8 |
| 4010 | 86.6 | 77.6 | 82.9 | 88.1 | 86.7 | 88.1 | 89.7 | 88.3 |
| MF-PLP + Energy + Kurtosis (reference features): 91.3 | | | | | | | | |

# 4. Results

The results for all the privacy-sensitive features and the spectral-based feature are reported in Tables 1, 2, and 3 for NIST, AMI, and ICSI meeting data, respectively. In the discussion that follows, EZ, EZK, and SEZK denote E + Z, E + Z + K, and S + E + Z + K, respectively. The findings from the study are summarized as follows.

## 4.1. Effect of temporal context

We first look at the 4 individual features: E,Z,S and K. We observe that when no temporal context is used, energy performs the best, and kurtosis fares the worst. On the other hand, when a temporal context of at least 1 sec is provided, kurtosis emerges as the best single feature. We note that when temporal context is modeled, all four features gain in performance. However, it can be observed that different features utilize different context lengths to attain their best performance. It can be seen that modeling temporal context also improves the performance of AH features, reaching a maximum at 500 ms context.

## 4.2. Combination of features

Combining the privacy-sensitive features results in an increase in performance. In most cases, it also leads to a decrease in the needed amount of context. For example, the feature EZ always yields performance equal to or better than either E or Z used individually. Further, we note that EZ achieves the best performance with a much shorter latency than either energy or zero crossing. This behavior of reduction in context was observed for other pairwise combinations as well. On the other hand, the addition of spectral flatness to EZK does not consistently improve the performance. Observe that while EZK is better than SEZK on NIST data, SEZK is better than EZK on AMI and ICSI data. This can be due to different temporal context lengths needed for different features. Finally, the combination study shows that the latency of the speech detection system, when modeling temporal context, can be reduced. For instance, one can observe from the results that a context of 500 ms or 1 sec is optimal, and that the difference between the performance of 500 ms context and 1 sec context is mostly negligible. Thus one can choose to operate at a lower latency.

## 4.3. Comparison between SEZK and AH

It can be observed that the performance of the AH features is not significantly different from the performance of the SEZK features. This can be explained in part by the fact that the two sets of features are similar in nature. For example, while zero crossing rate can be seen as being similar to the number of auto-correlation peaks, spectral flatness is similar to relative spectral entropy. The differences between the features arises from non-initial autocorrelation peak (provides an estimate of the amount of voicing) being different to energy and kurtosis.

## 4.4. Comparison with reference features

Lastly, we compare how the privacy-sensitive features perform against the reference spectral-based features. SEZK and AH perform similar to the reference features on NIST and AMI datasets. But on ICSI, we observe that the reference feature is significantly better than SEZK or AH features. We note that AMI was recorded in a small meeting room environment and consequently, the speakers were closer. On the other hand, ICSI meeting corpus was recorded in a larger meeting room with speakers being farther apart. This means that the signal-to-noise ratio (SNR) of the speech signal of a speaker who is farther from a lapel microphone is lower. Our hypothesis is that the MF-PLP features handle this case more effectively.

# 5. Discussion

Kurtosis emerges as the best single feature with context, but achieves a low performance with no context. However, [7] reports good performance for kurtosis with no context modeling. This could be due to a much larger short-term analysis window (160 ms) being used in [7] to estimate the fourth order moment, while we use a uniformly shorter analysis window of size 25 ms for all features. This suggests the use of multi-scale short-term analysis windows for different features. We want to investigate

Table 3: *Performance (in percentage of area under ROC) on ICSI data for all studied features. The best performance by AH is highlighted in bold and italics while the best performances by EZK and SEZK are highlighted in bold.*

| context(ms) | E | Z | S | K | EZ | EZK | SEZK | AH |
|---|---|---|---|---|---|---|---|---|
| 10 | 72.2 | 51.2 | 60.3 | 51.7 | 74.3 | 73.7 | 73.8 | 72.7 |
| 250 | 76.1 | 64.0 | 72.6 | 76.6 | 80.4 | 83.0 | 83.3 | 83.5 |
| 510 | 77.0 | 69.5 | 75.1 | 79.1 | 81.8 | **84.6** | **85.0** | ***85.7*** |
| 1010 | 76.7 | 69.0 | 74.1 | 81.4 | 81.8 | 84.4 | 83.6 | 81.8 |
| 2010 | 75.2 | 67.3 | 72.2 | 81.9 | 80.1 | 82.3 | 81.1 | 80.7 |
| 4010 | 74.6 | 67.3 | 69.9 | 81.8 | 77.3 | 79.6 | 79.1 | 83.0 |
| MF-PLP + Energy + Kurtosis (reference features): 90.3 | | | | | | | | |

this aspect in our future work.

On the other hand, providing large temporal context provided a way of regularizing kurtosis. As remarked in Section 4, the best performance of different single features suggests the use of different temporal context lengths. This could be one of the reasons why the performance of SEZK is not consistently higher than EZK. The issue of feature-dependent temporal context can be handled by first finding the optimal context for each feature and then utilizing that information to append the features accordingly when performing feature level combination or by combining the optimal classifier outputs. This is part of future work.

"Fundamentalness" was shown to be a promising feature in [7]. It was defined as having the maximum value when both amplitude and frequency modulation magnitudes are minimum. In addition, the long-term spectral-based feature proposed in [1] was shown to be effective for speech segmentation. We plan to include these features for comparison in our future studies.

Personal audio logs contain audio recorded in various places, and so features for speech detection need to be robust to varying environments. Future work will examine the robustness of all the privacy-sensitive features studied in this paper.

## 6. Conclusions

In this paper, we investigated four different privacy-sensitive features, namely energy, zero crossing rate, spectral flatness, and kurtosis, for speech detection in a multiparty conversation scenario that is closer to personal audio log scenario. We defined the datasets and annotations accordingly. Our studies showed that energy yields the best performance when no context was used, but using temporal context, kurtosis emerged as the most effective feature. Combination of features yielded a gain in performance. Combinations also revealed that a shorter temporal context could be sufficient. Finally, our study showed that privacy-sensitive features can achieve performance similar to the speech detection system using state-of-the-art non-privacy-sensitive spectral based features.

## 7. Acknowledgements

This work was supported by the Swiss National Science Foundation through the projects Micropower Integrated Face and Voice Detection (MIFAVO), MULTImodal Interaction and MULTImedia Data Mining (MULTI2), the National Centres of Competences in Research (NCCR) IM2, and Augmented Multiparty Interaction with Distance Access (AMIDA). The authors would like to thank John Dines for his help and support.

## 8. References

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

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

[3] S. Basu, "Conversational scene analysis." PhD Dissertation, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science, 2002.

[4] D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, "A Privacy-sensitive approach to modeling multi-person conversations." *Proc. IJCAI*, 2007.

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

[6] 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*, 1976.

[7] 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*, 2005.

[8] J. Makhoul, "Linear prediction: A tutorial review." *Proc. the IEEE*, pp. 561–580, 1975.

[9] J. S. Garofolo, C. D. Laprun, M. Michel, V. M. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus." in *Proc. LREC*, 2004.

[10] 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 *Proc. MLMI*, 2005.

[11] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style." in *Proc. ICSLP*, 2002.

[12] 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 *Proc. ICASSP*, 2003.

[13] S. H. K. Parthasarathi, P. Motlicek, and H. Hermansky, "Exploiting contextual information for speech/non-speech detection." in *Proc. TSD 2008*, 2008.