EVALUATING THE ROBUSTNESS OF PRIVACY-SENSITIVE AUDIO FEATURES FOR SPEECH DETECTION IN PERSONAL AUDIO LOG SCENARIOS

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,gatica,herve.bourlard}@idiap.ch

ABSTRACT

Personal audio logs are often recorded in multiple environments. This poses challenges for robust front-end processing, including speech/nonspeech detection (SND). Motivated by this, we investigate the robustness of four different privacy-sensitive features for SND, namely energy, zero crossing rate, spectral flatness, and kurtosis. We study early and late fusion of these features in conjunction with modeling temporal context. These combinations are evaluated in mismatched conditions on a dataset of nearly 450 hours. While both combinations yield improvements over individual features, generally feature combinations perform better. Comparisons with a state-of-the-art spectral based and a privacy-sensitive feature set are also provided.

Index Terms— Privacy Sensitive Features, Speech/nonspeech detection

1. INTRODUCTION

Recording spontaneous conversations, also referred to as personal audio logs, to analyze face-to-face human interaction patterns is an emerging field [1, 2]. However, one of the biggest obstacles facing this field concerns privacy. For example, recording and storing raw audio could breach the privacy of people whose consent has not been explicitly obtained. A possible solution to this problem is to store task-specific features instead of raw audio, such that neither intelligible speech nor lexical content can be reconstructed [2]. These features are referred to as privacy-sensitive (or privacy-preserving) features [2].

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

In an earlier paper [6], we investigated the use of four different, privacy-sensitive features, obtained by temporal processing of the audio signal, for speech detection in a multiparty conversation scenario. These features are the classical features, energy (E), zero crossing rate (Z), spectral flatness (S), and kurtosis (K). We showed that modeling the temporal context explicitly yields improvements for all privacy-sensitive features, including the features from [4, 5]. We also showed that the performance of all the privacy-sensitive features modeled with context is close to that of state-of-the-art spectral-based features used in [3]. However, this was done in *matched conditions*.

This paper is motivated by the fact that real-life personal audio logs often contain audio recorded in various environments. Consequently, we propose to evaluate the robustness of these four features (S,E,Z and K) in *mismatched conditions*. We also benchmark the performance of other authors' privacy-sensitive [4, 5] and a set of state-of-the-art features [3] in mismatched conditions. An important handicap for this evaluation is the lack of standard datasets in the personal audio log domain, due to privacy concerns. To overcome this challenge, we use the scenario constructed in our earlier study [6].

Our study shows that explicitly modeling the temporal context is useful for SND in mismatched conditions as well. Furthermore, we show that combining features (referred to as "early integration") or combining classifiers built on the individual features (referred to as "late integration") yield improvements. Lastly, combinations of the four features with context modeling, or of the features described in [4, 5] can yield, in certain cases, performance comparable to the state-of-the-art spectral based SND features [3] in mismatched conditions. We emphasize that our goal here is not to design the best SND system, but to evaluate the robustness of the privacy-sensitive features in mismatched conditions, in order to assess such a design.

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

2. DEFINITION OF DATA AND ANNOTATIONS

We use the scenario that was constructed in our previous study [6]. In that study, personal audio logs collected by subjects wearing portable audio recorders was likened to a meeting room scenario captured using lapel microphones. It was remarked that the placement of the recorder is similar to that of a lapel microphone used in recording meeting room conversations [2]. In the contrast to the traditional meeting room applications where, given the lapel microphone signal, the interest generally lies in the speech segments of the wearer [3, 7], in conversation analysis, speech segments that are spoken by other speakers are also of interest.

The dataset and annotations were used from our setup [6]. It consists of "individual" lapel microphone recordings used in conjunction with the ground truths obtained by merging speech segments from individual lapel ground truths that are closer than a fixed time interval (100ms). Our experiments were performed on lapel microphone recordings from NIST [8], AMI [9], and ICSI [10] meeting room data. To summarize, the total data add up to 100 hours

of meeting speech spanned over 120 meetings. The actual amount of individual lapel recordings add upto 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. Finally, using the ground truth defined above, the overall ratio of nonspeech to speech was 1:4.2. To test the performance on mismatched conditions, the features were trained in turn on each of the 3 datasets and tested on the other two.

3. SND SYSTEM

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

The proposed and the reference privacy-sensitive features are briefly discussed.

3.1.1. Proposed features

We evaluate the robustness of the four features investigated in our earlier study [6]. These four short-term features are: energy (E), zero crossing rate (Z), spectral flatness measure (S), and kurtosis (K) as privacy-sensitive features for SND. To simplify notations, let us define $F(SEZK)$ as the system with the combination of all four features at "feature-level" and $F(EZK)$ as the system with the combination of energy, zero-crossing rate and kurtosis at "feature-level".

3.1.2. Reference privacy-sensitive features (AH)

Features proposed in [5, 4] for privacy-sensitive speech detection are the non-initial maximum of the normalized autocorrelation, the number of autocorrelation peaks and the relative spectral entropy. Let AH denote the system using these features.

3.2. Reference spectral-based features (MF-PLP)

The reference spectral-based features (that is, non privacy-sensitive) are taken from a state-of-the-art SND system [3]. The features consist of 12 mel-frequency PLP coefficients (computed using HTK) and first cepstral coefficient c_0, with their delta and acceleration coefficients, in addition to energy and kurtosis. In [3], 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. Let $MF-PLP$ denote the system using these features.

3.3. Classifier

In this paper, we used off-the-shelf trained multi-layer perceptron (MLP) nets for individual (S, E, Z, and K) and the joint features ($F(SEZK)$ and $F(EZK)$) from our earlier setup [6]. In that study, these were the best combination of joint features. The MLP was trained for speech/nonspeech classes based on the ground truth definition described in Section 2, using two output units, 200 hidden units and by minimizing the cross-entropy criterion. 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. The features are normalized to zero-mean and unit variance at the input of the MLP. All the features were augmented with delta and acceleration features. Further details can be obtained in [6].

3.4. Classifier combination

One of the objectives of classifier combination ([11, 12]) is to exploit the complementary information between the classifiers. Combination techniques typically combine either the decisions made by the individual classifiers or assign a weight to each classifier's evidence. These weights can be either estimated statically (on cross-validation data) or dynamically. In this paper, we consider two weight allocation strategies:

- Dynamic weighting using inverse entropy.
- Static weighting using equal weights/averaging.

3.4.1. Inverse entropy

Inverse entropy based classifier combination has been shown to be useful in automatic speech recognition studies [12]. In the discussion that follows, let $c \in \{s, n\}$ denote the speech/nonspeech classes and let x_t^k denote a feature vector at a time t for $k \in \{S, E, Z, K\}$. $P(c|x_t^k; \theta_k)$ denotes the posterior probability estimate obtained from the MLP classifier trained on a feature $k \in \{S, E, Z, K\}$, and θ_k denotes the MLP model for a feature k. Inverse entropy based combination assigns larger weights to classifiers that are more confident and smaller weights to classifiers that are less confident [12]. The confidence of the k^{th} classifier is measured in terms of the entropy (h_k) of its posterior probabilities. The weights for the k^{th} classifier are then estimated as:

$$w_k = \frac{\frac{1}{h_k}}{\sum_j \frac{1}{h_j}} \quad \forall k \in \{S, E, Z, K\} \tag{1}$$

The combined evidence using all the features X_t^k:

$$P(c = i|X_t^k) = \sum_{k \in \{S,E,Z,K\}} w_k \cdot P(c = i|x_t^k; \theta_k) \quad \forall i \in \{s, n\} \tag{2}$$

3.4.2. Averaging

In this technique [11], all the classifiers are assigned equal weights, i.e., $w_k = \frac{1}{N}$. The output evidence is combined using equation 2. As part of notation, we use $C_{mean}(SEZK)$ and $C_{ent}(SEZK)$ to denote the systems with combinations of classifiers built on individual features S, E, Z and K using equal weights and inverse entropy techniques.

3.5. Evaluation measure

For evaluation, we use the area under the receiver operating characteristics (ROC) curve as a metric to evaluate speech detection, as in [6, 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. *Effect of context on SEZK and EZK using feature combination (in percentage of area under ROC). 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. AH(x) refers to the reference privacy-sensitive features with a temporal context of x ms.*

	N	A	I	N→A	N→I	A→N	A→I	I→N	I→A
Context (ms)	Matched conditions			Mismatched conditions					
SEZK									
10	77.6	80.7	73.1	77.0	67.1	76.3	74.1	70.1	76.9
250	84.0	89.6	80.9	**83.8**	71.7	85.5	78.5	83.1	87.0
510	**84.0**	**91.5**	**81.5**	79.7	71.5	**86.7**	**80.6**	**83.6**	**87.2**
1010	83.8	91.1	80.6	82.7	**72.9**	86.3	79.4	82.7	86.2
EZK									
10	77.8	80.1	73.8	78.5	72.0	75.6	74.0	72.9	78.2
250	83.5	88.8	80.5	***82.3***	74.4	84.1	78.7	81.3	85.8
510	***84.1***	***90.8***	***81.8***	82.0	***75.5***	86.0	***80.3***	***82.5***	***86.7***
1010	83.5	90.6	81.3	80.9	73.8	***86.5***	79.7	81.7	85.6
Reference features									
Features	N	A	I	N→A	N→I	A→N	A→I	I→N	I→A
	Matched conditions			Mismatched conditions					
AH(10)	74.9	79.8	72.7	77.4	68.7	75.4	68.1	72.9	75.0
AH(510)	83.3	90.3	85.7	86.0	75.7	85.3	78.9	83.6	88.1
MF-PLP	83.0	91.3	90.3	84.9	73.5	86.5	84.8	84.3	88.4

Table 2. *Effect of context on SEZK using classifier combinations (in percentage of area under ROC). 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.*

	N	A	I	N→A	N→I	A→N	A→I	I→N	I→A
Context (ms)	Matched conditions			Mismatched conditions					
Averaging the posteriors									
10	77.7	79.1	71.6	76.4	68.5	78.1	71.5	76.6	78.5
250	84.8	87.6	78.1	83.8	74.3	85.2	75.6	83.0	84.7
510	85.7	89.2	80.1	84.6	75.0	86.4	76.4	**83.8**	**85.7**
1010	**85.9**	**90.1**	**80.4**	**85.0**	**75.6**	**86.9**	**77.7**	83.5	84.4
Weighting the posteriors using inverse entropy									
10	74.9	78.5	71.9	76.7	68.7	74.7	70.4	73.5	78.3
250	82.7	87.5	78.4	83.8	74.4	82.7	75.7	80.9	84.4
510	83.5	89.3	***80.1***	84.6	75.1	83.8	76.9	***82.1***	***85.4***
1010	***83.8***	***90.1***	79.8	***85.0***	***75.6***	***84.0***	***78.1***	81.8	84.7

4. RESULTS AND DISCUSSION

The results for the privacy-sensitive features and the spectral-based feature in mismatched conditions are reported in Tables 1, 2, and 3 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. We also report the results[1] in matched conditions.

In general, we observe a drop in performance for all features in mismatched conditions (Tables 1 and 2). The exception being, when the dataset used for training is NIST. A detailed analysis of the findings from the study are given below.

4.1. Effect of temporal context

In [6], we reported that when temporal context was used in matched conditions, the performance of the individual and the feature combinations improve. Tables 1 and 2 demonstrate that this is true for feature and classifier combinations in mismatched conditions as well. Also, tables 1, 2 show that a context of 500 ms provides a reasonable tradeoff between accuracy and latency for feature and classifier combinations. Among the individual features (Table 3), when a temporal context of at least 500 ms is provided, kurtosis is the best single feature in mismatched conditions as it was in matched conditions. Similarly, energy is the second best feature. As in matched condition studies, zero crossing rate fares worst on mismatched conditions as well. Furthermore, we note that when temporal context is modeled, all four features gain in performance. It can also be seen from Table 1 that modeling temporal context also improves the performance of AH features.

4.2. Feature and classifier combinations

Although not reported here, pairwise and three-way combinations of features generally led to an improvement in performance in mismatched conditions as well. Among the three-way feature combinations, $F(EZK)$ was again consistently the best on mismatched conditions. As was observed in matched conditions, it can also be seen that there is no consistent improvement from $F(EZK)$ to $F(SEZK)$. From Table 1, it can be observed that while testing on

[1]The performance figures reported here differ from [6] due to a corrected implementation of kurtosis.

Table 3. *Performance of individual features (in percentage of area under ROC) with a context of 500 ms, in matched and mismatched conditions. 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.*

Features	N	A	I	N→A	N→I	A→N	A→I	I→N	I→A
	Matched conditions			Mismatched conditions					
S	80.5	84.7	75.1	82.7	70.8	80.0	71.9	75.6	77.5
E	80.1	87.2	77.0	81.9	75.9	**82.3**	75.6	80.6	83.8
Z	78.8	81.5	69.5	72.8	55.4	79.3	64.4	64.0	65.0
K	**82.8**	**87.9**	**77.7**	**83.3**	**76.2**	81.4	**75.6**	**82.2**	**85.0**

AMI (by training on either NIST or ICSI) yielded better performance for $F(SEZK)$, testing on ICSI or NIST yielded better performance for $F(EZK)$.

Table 2 reports the comparison between the two classifier combination methods, $C_{mean}(SEZK)$ and $C_{ent}(SEZK)$. It can be observed that the two methods are very similar on matched conditions. On mismatched conditions, the two methods show an important difference: training on NIST, $C_{mean}(SEZK)$ is better while testing on NIST, $C_{ent}(SEZK)$ is better. This may be due to the fact that when the classifiers are trained on more data[2] and therefore yield more robust estimates of posteriors, the confidence-based "inverse-entropy" method performs better. Otherwise, averaging is better when the estimates are not so robust (when the training data is less).

Between feature and classifier combinations, it can be seen that on matched conditions, training on NIST shows classifier combination techniques to be better for SEZK, while feature combination technique is better for AMI and ICSI datasets. This could also be due to differences in amount of training data. With a larger amount of training data, MLP is able to exploit the cross-correlation between the features in feature combinations. On the other hand, with lesser data, classifier combination yields better results.

On mismatched conditions, when NIST is used as training data, classifier combination is better while when AMI or ICSI datasets are used for training, feature combinations are better.

4.3. Comparison between $F(SEZK)$ and AH

Table 1 shows that the comparison with the AH features shows mixed results. For example, training on NIST dataset, $F(SEZK)$ is better while testing on NIST, AH is better. Also, training on AMI, $F(EZK)$ is better than AH, while testing on ICSI, AH is better than $F(EZK)$.

However, the AH features are not significantly different from the $F(SEZK)$ features, except for the way the spectral entropy is estimated. In AH, it is estimated explicitly in the spectral domain while in the proposed features, it is done through the residual obtained from linear prediction. This could be the reason for the mixed results.

4.4. Comparison with $MF - PLP$

We now compare how the privacy-sensitive features perform against the reference spectral-based features (MF-PLP). In matched conditions, $F(SEZK)$ and AH perform similar to the reference features on NIST and AMI datasets. In mismatched conditions, we observe that MF-PLP features are better than both the privacy sensitive features in certain cases.

[2]NIST dataset has less training data than AMI and ICSI

5. CONCLUSIONS

In this paper, we evaluated the robustness of the four privacy-sensitive features, namely, energy, zero crossing rate, spectral flatness measure, and kurtosis in mismatched conditions. We believe that to be a necessary step, as in real-life, mismatched conditions might be pervasive. For SND, we showed that explicitly modeling the temporal context is useful in mismatched conditions as well. Feature and classifier combinations for the proposed features on matched and mismatched conditions were explored. Furthermore, we showed that combining features or combining classifiers built on the individual features yield improvements. In addition, we showed that in certain cases, the combinations of the four features with context modeling can yield performance comparable to the state-of-the-art spectral based features in mismatched conditions.

6. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation through the projects MULTImodal Interaction and MULTImedia Data Mining (MULTI2) and the National Centres of Competences in Research (NCCR) IM2.

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

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

[5] S. Basu, *Conversational scene analysis.*, Phd dissertation, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science, 2002.

[6] S. H. K. Parthasarathi, M. Magimai.-Doss, H. Bourlard, and D. Gatica-Perez, "Investigating Privacy-Sensitive Features for Speech Detection in Multiparty Conversations.," in *Proc. Interspeech*, 2009.

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

[9] 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.

[10] 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.

[11] 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.

[12] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR.," in *Proc. ICASSP*, 2003.