INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Phoneme state posteriorgram features for speech based automatic classification of speakers in cold and healthy condition

Akshay Kalkunte Suresh[1], *Srinivasa Raghavan K M*[2], *Prasanta Kumar Ghosh*[2]

[1]Telecommunication Engineering, PES Institute of Technology, Bangalore 560085, India
[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, Karnataka, India
akshaykalkunte@gmail.com, srinivasaraghavankm@gmail.com, prasantg@ee.iisc.ernet.in

Abstract

We consider the problem of automatically detecting if a speaker is suffering from common cold from his/her speech. When a speaker has symptoms of cold, his/her voice quality changes compared to the normal one. We hypothesize that such a change in voice quality could be reflected in lower likelihoods from a model built using normal speech. In order to capture this, we compute a 120-dimensional posteriorgram feature in each frame using Gaussian mixture model from 120 states of 40 three-states phonetic hidden Markov models trained on approximately 16.4 hours of normal English speech. Finally, a fixed 5160-dimensional phoneme state posteriorgram (PSP) feature vector for each utterance is obtained by computing statistics from the posteriorgram feature trajectory. Experiments on the 2017-Cold sub-challenge data show that when the decisions from bag-of-audio-words (BoAW) and end-to-end (e2e) are combined with those from PSP features with unweighted majority rule, the UAR on the development set becomes 69% which is 2.9% (absolute) better than the best of the UARs obtained by the baseline schemes. When the decisions from ComParE, BoAW and PSP features are combined with simple majority rule, it results in a UAR of 68.52% on the test set.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Speech signal encodes complex physiological information which could be used for developing voice-based non-invasive clinical solutions. Commonly occurring upper respiratory tract infection, which is also referred to as common cold, conspicuously impacts characteristics of speech production. This change in voice characteristics, in turn, could degrade the performance of the automatic speech recognition (ASR) systems used in health-care applications and customer service industry as the ASR models are typically built with speech from people not necessarily affected by cold. Automatically identifying whether a voice sample is from a speaker affected by cold or not could help in selection of appropriate models to improve accuracies in such applications. In this work, we consider the task of classifying the speaker in cold and healthy condition from his/her voice.

There have been several works investigating the behavior of cold speech in speaker recognition. Studies by Tull et al. [1, 2, 3] reveal differences in formant patterns, nasality parameters and melcepstral coefficients between normal and cold speech. Shan et al. [4] observed variations in the energy levels at lower and higher frequency bands and concluded that using mel-frequency cepstral coefficient (MFCC) computed on cold speech processed with pre-emphasis filter improves speaker recognition performance. Recently, studies have been made to classify speech affected by cough [5], by using support vector machine (SVM), Bayesian and neural network trained on two sets of features that include time and frequency domain approaches. These studies indicate discernable variations in acoustics of speech affected by cold. However, cold classification problem remains a challenge when test voice sample is available from any unknown speaker in natural recording scenario. This requires analysis and modeling of cold speech characteristics in a speaker-independent manner.

P. Rose [6] pointed out that the cold is often accompanied by nasal cavity's inflammation and swelling, which changes the volume and shape of nasal cavity and furthermore affects the nasal modulation of sound source excitation signal and causes the speaker's voice to change. Thus, the condition could affect both glottal source as well as the nasal cavity characteristics that, in turn, could affect the quality of speech. Following the work by Tull et al. [7], we, in this work, hypothesize that characteristics of different sounds (phonemes) are affected to different degrees by cold condition of the speaker. We propose to capture the differences in acoustics of various phonemes due to the cold condition using a phonetic state posteriorgram approach. Using posteriorgram, we aim to normalize the variability due to speaker and recording conditions and emphasize the effect due to cold condition.

Posteriorgrams have been used for a number of applications in the past. For large vocabulary continuous speech recognition (LVCSR) tasks, Ma et al. [8] show that combining conventional cepstral features with class posterior probability features obtained from multi layer perceptron (MLP), results in a lower word error rate (WER). Williams et al. [9] show that phone posteriors are effective in differentiating clean speech from other acoustic segments like music and perform similar to features derived specifically for that purpose. Aradilla et al. [10] show that for template based ASR approaches, posterior based features outperform MFCC features. Zhang et al. [11] show that for unsupervised speech pattern recognition, a Gaussian posteriorgram based approach removes speaker variability. In detecting mispronunciation on a word level, Lee et al. [12] show that posteriorgram features obtained from a Deep Belief Network (DBN) perform better compared to MFCCs and unsupervised Gaussian posteriorgram features.

We compute the posteriorgram using an acoustic model built on a large number of speakers in healthy condition speaking in different recording environments. Thus, given the acoustics of cold speech, we assume that the likelihoods of the cold speech features against this model would be smaller compared to those of healthy speech features. In order to capture the difference in the distribution of the likelihoods for cold and healthy conditions for each phonetic state, we compute a set of statistics from the likelihoods across frames of each utterance. With 120 phonetic states, this results in 5160 phonetic state posteri-

http://dx.doi.org/10.21437/Interspeech.2017-1550

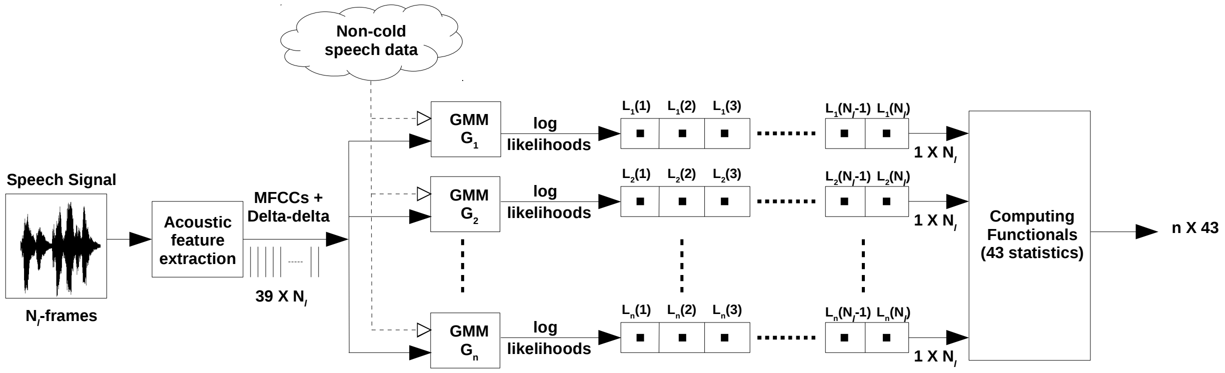
Non-cold speech data
Speech Signal
N_l-frames
Acoustic feature extraction
MFCCs + Delta-delta
39 X N_l
GMM G_1
GMM G_2
GMM G_n
log likelihoods
log likelihoods
log likelihoods
$L_1(1)$ $L_1(2)$ $L_1(3)$ $L_1(N_l\text{-}1)$ $L_1(N_l)$
$L_2(1)$ $L_2(2)$ $L_2(3)$ $L_2(N_l\text{-}1)$ $L_2(N_l)$
$L_n(1)$ $L_n(2)$ $L_n(3)$ $L_n(N_l\text{-}1)$ $L_n(N_l)$
1 X N_l
1 X N_l
1 X N_l
Computing Functionals (43 statistics)
n X 43

Figure 1: *Steps in computing phoneme state posteriorgram (PSP) features*

orgram (PSP) features. Cold classification experiments on the 2017 Cold-sub challenge [13] Upper Respiratory Tract Infection Corpus (URTIC) show that the proposed PSP features result in an unweighted average recall (UAR) of 64% and 61.09% on the development and test sets. When the PSP features are used for a majority rule based late fusion with the baseline features, namely ComParE and BoAW, we obtain a UAR of 65.3% on development set and 68.52% on test set. We also consider late fusion on the decisions from PSP features, BoAW and end-to-end (e2e) models. This results in a UAR of 69% on the development set, which is 2.9% better than the best of the UARs obtained by different combinations of baseline schemes [13] although this combination yields a UAR of 66.7% on the unknown test set.

2. Phoneme state posteriorgram features

The steps in computing PSP features for a given utterance are summarized in Figure 1. 39-dim Mel frequency cepstral coefficients (MFCCs) with their first and second derivatives are used as the acoustic features for computing utterance level PSP features. At first, we train a phonetic hidden Markov model (HMM) with three states using samples from a large set of speakers in healthy condition. The distribution of the acoustic features in each state is modeled by a 256 component Gaussian mixture model (GMM). To train the HMM in this work, we consider speech from two large corpora, namely TIMIT [14] and Boston University Radio News (BN) [15]. TIMIT corpus consists of recordings by 630 speakers in eight dialects of American English, reading phonetically rich sentences in a quiet environment. It contains recordings from 438 males and 192 females. The duration of the entire TIMIT corpus 5.4 hours. The BN corpus is made up of speech recorded from radio announcers both during broadcasts and laboratory recordings. They are in American English, from three male and four female announcers. The laboratory recordings are done in both radio and non-radio styles. The duration of the entire BN corpus is 11 hours. TIMIT corpus is comprised only of read speech while the BN corpus also includes spontaneous speech. Combining these two datasets provides more than 16 hours of speech and helps in improving the acoustic richness with respect to speakers, recording environments and speaking styles.

Given the speech samples along with their transcriptions, the parameters of the GMM-HMM system is learnt using standard HMM training in ASR. In this work we consider 40 phonemes as listed in the work by [16]. Considering three HMM states per phoneme, we obtain $n = 3 \times 4 = 120$ phonetic HMM states. Lets denote the GMMs for these states by G_1, G_2, $\cdots$, G_n. The parameters for the i-th GMM (G_i) is given by $\lambda^i = \{w_j^i,\, \mu_j^i,\, \Sigma_j^i,\, j = 1 : 256\}$, where w_j^i is the weight for the j-th component; μ_j^i and Σ_j^i are the mean vector and diagonal covariance matrix for the j-th component. Given 39-dim acoustic feature vector $\underline{x}_k$, the log likelihood using G_1, G_2, $\cdots$, G_n are computed as follows:

$$L_i(k) = P(\underline{x}_k|G_i) = \log\left[\sum_{j=1}^{256} w_j^i \mathcal{N}(\underline{x}_k;\, \mu_j^i, \Sigma_j^i)\right] \qquad 1 \le i \le n, \tag{1}$$

where $\mathcal{N}$ denotes a multi-dimensional Gaussian density function. Suppose there are N_l frames in the l-th utterance. Computing the log likelihood using eqn. (1) results in a $n \times N_l$ matrix L^l whose (i, k)-th element is given by $L_i^l(k)$ defined in eqn. (1). The matrix L provides the log likelihood of each phonetic state in every frame.

In order to examine how the likelihood of cold speech differs from that of the non-cold speech, we compute the phonetic state likelihood matrix L^l for l-th utterance from the URTIC dataset [13]. Suppose there are C cold utterances and N_C healthy subjects' utterances. We compute the average likelihood $\bar{L}_i^{C,l}$ for i-th phonetic state across all frames in the l-th $(1 \le l \le C)$ cold utterance as follows:

$$\bar{L}_i^{C,l} = \frac{1}{N_l}\sum_{k=1}^{N_l} L_i^l(k) \tag{2}$$

Similarly, for the l-th $(1 \le l \le N_C)$ non-cold utterance we obtain:

$$\bar{L}_i^{NC,l} = \frac{1}{N_l}\sum_{k=1}^{N_l} L_i^l(k) \tag{3}$$

We compute the mean of the L_i^l across all cold and non-cold utterances separately in the URTIC dataset [13] to obtain an average value of likelihood for each phonetic state. This provides us the the average phonetic state specific likelihoods for cold and non-cold utterances separately as follows:

$$L_i^C = \frac{1}{C}\sum_{l=1}^{C} L_i^{C,l} \qquad L_i^{NC} = \frac{1}{N_C}\sum_{l=1}^{N_C} L_i^{NC,l} \tag{4}$$

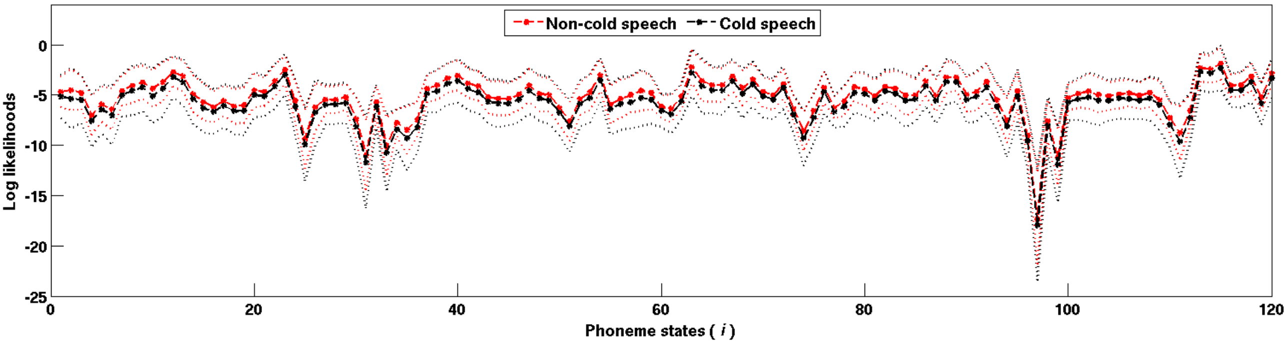

Figure 2: *Comparison of average log likelihoods of cold and non-cold speech with +/- standard deviations indicated with fine dotted lines*

Figure. 2 shows the variations of L_i^C and L_i^{NC} with different phoneme states (i). Although there is a large overlap, on average, it is clear that L_i^C is lower than L_i^{NC} for all phoneme states. This indicates that the cold speech features, on average, result in lower likelihoods against the GMMs of each phoneme state compared to the non-cold speech features. We find that the phonemes with highest ten differences in the likelihoods L_i^C and L_i^{NC} are AA, EH, V, DH, IY, AX, JH, W, T, NG. It is interesting to note that the nasal sound NG appears in the top ten most discriminating phonemes particularly due to the change in the nasal cavity due to cold.

In order to compute the PSP feature from $L_i^l(k), 1 \leq k \leq N_l$, we calculate 43 functionals over all the frames of an utterance, for each of the 120 phoneme states following the work in [17]. The 43 functionals describe the distribution of the likelihoods for each phoneme state. We hypothesize that the shape and nature of the distribution, captured through these functionals, would discriminate the cold and non-cold classes well. We, thus, obtain a 120×43 matrix of functionals for each utterance, which we vectorize to obtain a 5160-dim PSP feature vector.

3. Experiments and Results

3.1. Dataset

URTIC corpus considered for Interspeech 2017 Cold Sub-Challenge, is a German speech corpus provided by the Institute of Safety Technology, University of Wuppertal, Germany. The recordings consist of 630 subjects including 382 males and 248 females, with ages ranging from 12 to 84 years. The speech signal is quantized at 16bit and downsampled from 44.1kHz to 16kHz. The total duration of the corpus is 45 hours where each audio chunk has a duration of 3 to 10 seconds. The reference for cold condition is binarised for global illness severity based on German version of the Wisconsin Upper Respiratory Symptom Survey (WURSS-24) [14]. URTIC corpus contains 9505 train, 9596 development and 9551 test utterances. The train and development sets are accompanied by cold (C)/non-cold (NC) labels while the test set is devoid of labels. Each utterance is divided into frames with a window size of 25ms shifted by 10ms.

3.2. Results and Discussions

We report results obtained using the proposed 5160-dim PSP features, end-to-end (e2e) model and discuss the effect of feature selection and decision fusion. Decision fusion of PSP, BoAW and e2e models results in the highest UAR of 69% on the development set which is 2.9% better than the best baseline model. Decision fusion of PSP, BoAW and ComParE models results in the highest UAR of 68.52% on the test set.

Table 1: *UAR% on Development and Test sets*

Scores for cold speech classification (UAR%)		
Model	**Dev**	**Test**
2017 InterSpeech Cold Sub-Challenge baseline results		
ComParE functionals	64.00	70.20
ComParE BoAW	64.20	67.30
PSP and fusion (PSP+ComParE+BoAW)		
PSP	64.00	61.09
fusion(unweighted maj.)	65.30	**68.52**
fusion(weighted maj.)	66.70	65.09
fusion (PSP+BoAW+e2e)		
fusion(unweighted maj.)	**69.00**	66.70

3.2.1. *PSP features*

The 5160-dim PSP feature vector is used to train an SVM classifier. We standardize each feature separately to have zero mean and unit standard deviation and use Weka [18], a data mining software to implement SVM classifier. We train the SVM classifier with a complexity constant of 10^{-5} and an epsilon intensive loss of 0.1. The classification performance is reported using unweighted average recall (UAR) in percentage, which is a metric invariant to class imbalance. The PSP features, as seen in Table 1, result in a UAR of 64% on development set and 61.09% on the test set with SVM classifier using linear kernel.

3.2.2. *Feature Selection*

Table 2: *UAR% on Development set for PSP and feature selection criteria*

Scores for cold speech classification (UAR%)	
Features	**Dev**
PSP	64.00
PSP(top 10 phonemes)	63.60
PSP(top 500 fisher discrimination features)	63.50

We analyze the effect of feature selection on the PSP features in two different ways. The first one involves generating a 473-dim PSP feature vector by considering 11 phone states corresponding to the 10 phones (AA, EH[1],V, DH, IY, AX, JH, W,T, NG) that result in the top 10 largest difference in the likelihoods between cold and non cold samples. As seen in Table 2, we obtain a UAR of 63.6% on the development set with this selected 473-dim PSP feature vector.

The second type of feature selection considers the Fisher discrimination value $F_d = (m_c - m_{nc})^2/(v_c + v_{nc})$ for each of the 5160 PSP features ($1 \leq d \leq 5160$) between cold and non-cold utterances , where (m_c, v_c) and (m_{nc}, v_{nc}) are the mean and variances of the feature values across all cold and non-cold utterances respectively. We consider only features with the top 500 Fisher discrimination values for this purpose and generate a 500-dim PSP feature vector for each utterance. We observe a UAR of 63.5% with this reduced feature vector on the development set. Both the feature selection criteria applied to the PSP features result in a UAR score very close to the original 5160-dim PSP feature vector indicating that few features are critical for discriminating cold and non-cold classes.

We further analyze feature selection on the ComParE features. For this purpose, we divide the 6373 ComParE features into 27 unique categories as listed in Table 3. We obtain each category by considering together the features related to that unique category. As an example, the 'JitterDDP' category is composed of 39 functionals computed on JitterDDP smoothened and JitterDDP smoothened delta-delta features. We analyze the 27 feature categories independently to find the class that performs the best for the classification task. We combine the training and development sets for each of the 27 categories and perform 6-fold cross validation. We train SVM models holding out one fold as the test set each time and average the UARs. We also show the UARs obtained using the 27 feature categories in Table 3. The class pcm_fftMag_mfcc performs the best with a UAR of 73.1%. However, the rest of the classes perform uniformly and worse than pcm_fftMag_mfcc.

Table 3: *27 unique feature categories from ComParE features*

Feature Category	Feature count	UAR%
audspec_lengthL1norm	100	58.80
audspecRasta_lengthL1norm	100	58.30
audSpec_Rfilt	2600	68.20
F0final	83	58.35
jitterDDP	78	56.55
jitterLocal	78	56.85
logHNR	78	55.10
pcm_fftMag_fband1000-4000	100	61.75
pcm_fftMag_fband250-650	100	60.70
pcm_fftMag_mfcc	**1400**	**73.10**
pcm_fftMag_psySharpness	100	58.50
pcm_fftMag_spectralCentroid	100	58.50
pcm_fftMag_spectralEntropy	100	58.15
pcm_fftMag_spectralFlux	100	59.30
pcm_fftMag_spectralHarmonicity	100	62.20
pcm_fftMag_spectralKurtosis	100	58.30
pcm_fftMag_spectralRollOff25.0	100	58.15
pcm_fftMag_spectralRollOff50.0	100	58.60
pcm_fftMag_spectralRollOff75.0	100	59.75
pcm_fftMag_spectralRollOff90.0	100	60.50
pcm_fftMag_spectralSkewness	100	57.60
pcm_fftMag_spectralSlope	100	61.60
pcm_fftMag_spectralVariance	100	58.70
pcm_RMSenergy	100	62.10
pcm_zcr	100	58.60
shimmerLocal	78	58.05
voicingFinalUnclipped	78	56.25

3.2.3. *End-to-End model*

We train the baseline end-to-end (e2e) model [19] with 8 convolutional and maxpooling layers followed by 2 Long Short-Term Memory (LSTM) layers on the raw audio files which are divided into 40ms samples. We hypothesize that the e2e classification approach could learn unique time-frequency representations using the convolutional and LSTM layers with the potential to observe new representations in the data. With this model, we obtain a UAR of 66.5% on the development set.

3.2.4. *Decision Fusion*

We perform decision fusion from the SVM model trained on PSP features and the SVM models trained from the baseline schemes, namely ComParE, BoAW, using the unweighted majority rule, where the models built on the three feature sets are equally weighted and the majority decision is considered as the final classification output. This rule results in a UAR of 65.3% on the development set and a UAR of 68.52% on the test set as shown in Table 1. We also perform decision fusion from the SVM model trained on PSP features, the SVM model trained on BoAW and the e2e model using the unweighted majority rule. This results in a UAR of 69% on the development set and a UAR of 66.7% on the test set.

We also perform a weighted majority decision for decision fusion. Accordingly, the decisions from the 27 feature categories are equally weighted with a score of 1. The BoAW and the PSP features are each weighted with a score of 27. We perform decision fusion on these 29 weighted decisions and determine the class labels. We hypothesize that such a weighted fusion would provide a decision by combining each feature category from the ComParE features. With this rule, we obtain a UAR of 66.7% on the development set which is 0.6% higher than the best baseline model. On the test set, we observe a UAR of 65.09%. The UAR values are provided in Table 1.

3.2.5. *Effect of corpus on computing PSP features*

We compute PSP features using the HMMs trained on TIMIT speech corpus, BN speech corpus and a combination of TIMIT and BN speech corpora, in order to examine the effect of the corpus chosen. The UAR on the development set using PSP features are found to be 65.1%, 60.5% and 64.0% when TIMIT, BN and TIMIT+BN are used as the corpus for training the GMM-HMM respectively. The poor UAR using BN could be due to noisy recordings present in BN unlike those in TIMIT.

4. Conclusions and future work

We have proposed phoneme state posteriorgram based features to capture the acoustic variability due to speaking in cold condition compared to speaking in healthy condition. We obtain a UAR on the development set comparable to that using the baseline scheme. When combined with the baseline scheme with a weighted decision fusion approach, we obtain 2.9% (absolute) improvement in the UAR on the development set.

In the present work, the HMMs were trained on American English speech corpora, while the language of the URTIC is German. As future work, we would like to examine the benefit of the proposed PSP features computed using HMMs trained on a German speech corpus. It will be interesting to use a deep neural network (DNN) model to classify the cold and non-cold utterances, which could perform well as the training data is large. We also intend to compute the PSP features from a HMM trained with a deep neural network, which could provide a PSP feature vector that could result in better discrimination between cold and normal speech.

[1] 2 phoneme states correspond to the same phoneme 'EH' resulting in 11 states each with 43 functionals resulting in $43 \times 11 = 473$ features.

5. References

[1] R. G. Tull, *Acoustic analysis of cold-speech: Implications for speaker recognition technology and the common cold*. PhD Thesis, Northwestern University, 1999.

[2] R. G. Tull and J. C. Rutledge, "Analysis of cold-affected speech for inclusion in speaker recognition systems." *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2549–2574, 1996.

[3] R. G. Tull, J. C. Rutledge, and C. R. Larson, "Cepstral analysis of cold-speech for speaker recognition: A second look," *The Journal of the Acoustical Society of America*, vol. 100, p. 2760, 1996.

[4] Y. Shan and Q. Zhu, "Speaker identification under the changed sound environment," in *International Conference on Audio, Language and Image Processing (ICALIP)*. IEEE, 2014, pp. 362–366.

[5] B. Ferdousi, S. F. Ahsanullah, K. Abdullah-Al-Mamun, and M. N. Huda, "Cough detection using speech analysis," in *18th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2015, pp. 60–64.

[6] P. Rose, *Forensic speaker identification*. CRC Press, 2003.

[7] R. Tull, "Investigating the common cold to improve speech technology," *Media lay paper, Am. Institute of Physics*, 1996.

[8] C. Ma, H.-K. J. Kuo, H. Soltau, X. Cui, U. Chaudhari, L. Mangu, and C.-H. Lee, "A comparative study on system combination schemes for LVCSR," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4394–4397.

[9] G. Williams and D. P. Ellis, "Speech/music discrimination based on posterior probability features." in *Eurospeech*, vol. 99, 1999, pp. 687–690.

[10] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," IDIAP, Tech. Rep., 2006.

[11] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4366–4369.

[12] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8227–8231.

[13] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," *in Proc. INTERSPEECH 2017*.

[14] C. Lopes and F. Perdigao, "Phone recognition on the timit database," *Speech Technologies/Book*, vol. 1, pp. 285–302, 2011.

[15] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[16] K.-F. Lee and H.-W. Hon, "Speaker-independent phoneme recognition using hidden Markov models," *The Journal of the Acoustical Society of America*, vol. 84, no. S1, pp. S62–S62, 1988.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[19] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5200–5204.