AUTOMATIC DETECTION OF ANGER IN TELEPHONE SPEECH WITH ROBUST AUTOREGRESSIVE MODULATION FILTERING

Jouni Pohjalainen and Paavo Alku

Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

ABSTRACT

A new system for automatic detection of angry speech is proposed. Using simulation of far-end-noise-corrupted telephone speech and the widely used Berlin database of emotional speech, autoregressive prediction of features across speech frames is shown to contribute significantly to both the clean speech performance and the robustness of the system. The autoregressive models are learned from the training data in order to capture long-term temporal dynamics of the features. Additionally, linear predictive spectrum analysis outperforms conventional Fourier spectrum analysis in terms of robustness in the computation of mel-frequency cepstral coefficients in the feature extraction stage.

Index Terms— emotion detection, speech analysis

1. INTRODUCTION

Automatic recognition of emotions from speech has many potential applications, e.g., in call centers. The area of emotion recognition has been actively studied in recent years [1] [2] [3]. Typically, the purpose of these systems is to recognize emotion classes such as anger, joy, sadness and surprise [1].

Among previous studies on vocal emotions, a few investigations have specifically focused on the detection of angry speech [4] [5] [6] [7]. These studies have used real-world data collected at call centers. Using a standard audio pattern recognition approach based on Gaussian mixture models (GMMs) and mel-frequency cepstral coefficient (MFCC) feature extraction supplemented with additional MFCC features to model long-term dynamics and prosody, recent systems have achieved class-averaged misclassification rates below 20 % [5] [7]. Support vector machines are a competing alternative to GMM classification [6] [7].

In contrast to the majority of previous work on the detection of anger, the present study focuses, firstly, on robustness with respect to changing acoustic conditions and realistic noise. In the past years, the robustness aspect has gradually gained momentum in the recognition of vocal emotions [1] [8] [9] [10]. Secondly, the present study uses acted emotional data from the popular Berlin database of emotional speech [11] in order to evaluate components of the detection system in a reproducible and comparable manner and also to place more emphasis on easily confusable emotion classes (although recognizing acted emotions is typically easier [1]). Because call centers are an important application domain for emotion recognition in general, and for anger detection in particular as it has a relation to customer satisfaction [12], the present study also keeps the main focus on telephone speech, however the original non-telephone data is also analyzed. The speech material is modified by simulating the GSM transmission channel. Together with far-end noise corruption using real-world noise types, this results in a controllable, yet realistic experiment. The main goal of this study is to evaluate, in terms of detection performance in various adverse conditions, a new method for modeling the temporal dynamics of features. The performance of linear predictive vs. Fourier spectrum analysis in feature extraction is also investigated.

2. DETECTION SYSTEM

2.1. Feature extraction

The signal is pre-emphasized with $H_p(z) = 1 - 0.97z^{-1}$ and arranged into overlapping Hamming-windowed frames of 25 ms with a shift interval of 10 ms. For each frame, 12 MFCCs (excluding the zeroth coefficient) are computed using the well-known processing chain: 1) obtain squared magnitude spectrum, 2) apply a mel filterbank to the squared magnitude spectrum, 3) take the logarithm of filtered band energies and 4) perform discrete cosine transform [13]. The mel filterbank consists of 40 triangular filters with center frequencies spaced evenly on the mel scale. The 12 MFCCs are complemented with logarithmic frame energy whose mean and variance have been normalized over the complete audio file. Finally, delta and double-delta coefficients of the MFCCs and log energy are appended, resulting in a 39-dimensional feature vector.

Typically, the magnitude spectrum represented by the MFCC feature vector is obtained using discrete Fourier transform (DFT), implemented by fast Fourier transform (FFT) algorithms. However, DFT analysis is not considered to be particularly resistant to additive noise. Previous studies on, e.g., speaker verification [14], automatic speech recognition

The work was supported by Academy of Finland projects 127345 and 135003 (LASTU programme).

[15] and vocal effort classification [16] have shown improved robustness when the DFT-based magnitude spectrum analysis has been replaced with linear predictive methods in the MFCC computation chain detailed above.

In linear prediction (LP), the coefficients of an all-pole filter $1/(1 - \sum_{k=1}^{p} a_k z^{-k})$ are obtained by minimizing the prediction error energy $\sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2$ of a short-time analysis frame consisting of speech samples s_n. The prediction order p is typically chosen as the sampling frequency in kHz incremented by a small integer [17]. For example, $p = 20$ is a typical choice for a signal sampled at 16 kHz. This way, LP models the envelope of the magnitude spectrum while excluding the fine structure. In the present evaluation, $p = 20$ will be used for both 8 kHz and 16 kHz material, giving a somewhat more detailed model in the narrowband case, yet not detailed enough to capture the spectral fine structure.

2.2. Frame processing

Prosody is known to be an effective cue in the recognition of vocal emotions. Approaches to modeling prosody in emotion recognition systems include, for example, features based on the modulation spectrum [3] [10], separate MFCC modeling for low frequencies down to 20 Hz [2] [5] and perhaps most commonly, long-term statistics and functionals of frame-based short-time features [1] [12]. In the present study, the approach chosen to modeling prosody is by means of intermediate frame processing which can be plugged in after the short-time feature extraction phase.

The short-time features are filtered across speech frames by autoregressive (AR) models learned from the training data. For each feature, least squares regression is used in the system training phase to fit an AR model so as to represent the time behavior of the feature within the target class (the target class in the present study being angry speech). Subsequently, after feature extraction in both training and detection, the AR models are used to predict the feature values based on the unprocessed features and the features are replaced by their predicted values, i.e., their predictable components. More precisely, if $x_{t,j}$ is the value of the jth feature in the tth frame, the original features $x_{t,j}$ are replaced by the predictions

$$y_{t,j} = c_j + \sum_{k=1}^{r} b_{j,k} x_{t-sk,j}, \qquad (1)$$

where c_j and $b_{j,k}$, $1 \leq k \leq r$ are the parameters of the rth-order AR model trained for the jth feature to represent the target class. c_j is a constant term and the $b_{j,k}$ are the AR predictor coefficients. s is an integer specifying the frame skip interval for the autoregression. If $s = 1$, the AR models consider each preceding frame in making the predictions, with $s = 2$ every second frame is considered, etc. Because the short-time frames generally overlap, and delta features capture some more information about the vicinity of the frame, it may not be necessary to have the autoregressive model see each frame and s can be chosen to be greater than 1; in this way, a longer time span can be covered with the same number of parameters, resulting in less complex models.

The frame filtering approach is related to the popular RASTA modulation filtering of speech feature vectors [18], which employs an IIR band-pass filter

$$H(z) = 0.1z^4 \frac{2 + z^1 - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}.$$

However, while RASTA broadly emphasizes modulation frequencies active in normal speech on the average, the present method learns a more particular modulation frequency representation of the target class in the detector training phase.

Most of the modulation spectrum energy in normal speech is concentrated around 4 Hz but there is significant energy also at the frequency range between 2 and 8 Hz [19]. If there is generally large energy in this frequency range, differences between various speech classes potentially also manifest themselves there. Short-time features and delta features can not accurately capture this information. The use of autoregressive modeling of the long-term dynamics of the features is thus motivated by the potential benefits obtainable by accurate modeling of this frequency range. Therefore, the length of the “history” of the autoregression, given by rs times the frame shift interval (here 10 ms), should be chosen to be long enough to capture these frequencies.

Some recent studies have used autoregressive models in emotion feature extraction [20] and classification [21]. The present approach is based on different principles and is an independent filtering method usable with various classifiers.

2.3. Detection rule

For the purpose of binary classification according to the Bayes rule [22], the detection system models non-angry and angry speech with their own Gaussian mixture models (GMMs), a typical approach to audio class detection [5] [6] [7] [16]. Each GMM has a diagonal covariance structure [23]. The GMMs are trained using 10 iterations of expectation-maximization (EM) re-estimation for GMMs [23]. Before training, the mean vectors of each component are initialized by applying a selection approach intended for the initialization of the cluster means in EM-style cluster-seeking algorithms [24]; the selection is applied to 12 MFCCs while the other 27 elements are initialized by averaging within the initial clusters. The component weights are initialized by uniform distributions and the variances by 0.1 times the global variances.

In the detection phase, for each utterance, the averaged log likelihoods of the processed feature vectors having been produced by each GMM are computed and denoted as L_{anger} and $L_{\text{non-anger}}$. With T denoting the decision threshold, the detection rule for the logarithmic likelihood ratio is

$$L = L_{\text{anger}} - L_{\text{non-anger}} > T. \qquad (2)$$

3. EXPERIMENTAL EVALUATION

3.1. Test material

The Berlin database of emotional speech, commonly referred to as EMO-DB, was used as the evaluation material [11]. The database has been widely used in emotion classification studies, e.g., [1] [3] [9] [10]. It consists of 535 utterances of German sentences spoken in seven different emotional styles by five male and five female actors. The emotion categories contained in this database are *anger*, *disgust*, *fear*, *joy*, *sadness*, *boredom* and *neutral*. The goal was to distinguish the *anger* class from the six other classes.

3.2. Preparation of the evaluation material

The emotional speech database was analyzed in two forms: original, unprocessed data sampled at 16 kHz and telephone channel data sampled at 8 kHz. For the telephony conditions, additive noise from the NOISEX-92 database was first added to the signal in order to simulate additive ambient noise at the location of a mobile station. Three noise types were used: *volvo* (inside a moving car), *factory1* (mechanical factory noise including frequent transient impulsive sounds) and *babble* (many people talking simultaneously). The noise corruption was performed at 16 kHz sampling rate with a controlled segmental signal-to-noise ratio (SNR), i.e., the average over 25 ms frames. The signal corrupted with car interior noise at SNR 30 dB simulated speaking in a relatively quiet location. The mismatched noisy telephone data contained seven conditions: car noise at SNR 0 dB and both factory noise and babble noise at SNRs 10, 0 and -10 dB.

In order to simulate the telephone transmission channel, noise-corrupted speech signals sampled at 16 kHz were first high-pass filtered with the mobile station input (MSIN) filter that approximates the input characteristics of a mobile terminal [25] and decimated to the sampling rate of 8 kHz. The speech level was subsequently normalized to 26 dB below overload point. Finally, the signals were processed with the adaptive multi-rate (AMR) codec [26], which is commonly used for speech coding in the GSM cellular system, at a bit rate of 12.2 kbps.

3.3. Evaluation procedure

Detection evaluation on the EMO-DB was conducted as 10-fold cross validation according to speaker: one speaker at a time was chosen as the test speaker and the utterances from the remaining nine speakers were used for training the detection system. As a speaker's speech is never recognized using models trained by his/her own speech, the possibility of the recognition scores being improved by learning particular emotional styles of individual speakers is avoided and the speaker dependency of the evaluation is thereby minimized. Not all published emotion classification studies with the EMO-DB use the speaker-independent approach; however, a clear difference between speaker-independent and random 10-fold cross validation approaches on this database, with the speaker-independent approach giving lower correct classification scores, has been reported in [3].

The equal error rate (EER) was used to evaluate the system performance in the detection task. The EER is the value of both the miss rate and the false alarm rate using a decision threshold in Eq. 2 that makes these error rates equal to each other. In addition, detection-error-tradeoff (DET) curves were examined. Statistical analysis between the different processing methods was performed using a statistical significance test developed for detection systems [27]. As all the detections use the same analysis block division and original speech material, the "dependent-case" version of this test was employed.

3.4. Results

The experiments begun by comparing two well-known spectrum analysis methods, FFT and LP, in mismatched telephony conditions in which the detector is trained with SNR=30 dB car interior noise. The results are shown in Table 1. The difference between FFT and LP was statistically significant at the 95 % level in the cases of SNR=0 dB car noise, SNR=10 dB factory noise and SNR=0 dB babble noise.

Table 1. *EER scores (%) for the detector using different spectrum analysis methods in MFCC feature extraction. The matched as well as seven mismatched background noise conditions have been evaluated after training the detector with telephone speech containing far-end car interior noise with SNR=30 dB. The number of components per GMM was 16.*

Noise	SNR (dB)	FFT	LP
Car	30	12.5	12.4
	0	14.8	10.7
Factory	10	15.0	10.9
	0	18.0	17.1
	-10	33.8	35.4
Babble	10	13.3	12.4
	0	14.1	11.8
	-10	17.2	16.5

As the second phase, the effect of AR feature filtering was evaluated. Both the order of the autoregression and the frame interval were varied in such a manner that five durations of the autoregression time span were covered: 80, 160, 240, 320 and 400 ms. This was done for two cases: 1) original, clean 16 kHz data without mismatch between training and detection and 2) one mismatched telephony condition in which the system has been trained in a car interior scenario with SNR 30 dB and evaluated in a factory noise scenario with SNR 0 dB. In addition, these two cases were evaluated without any frame processing and with RASTA processing. Table 2 shows the results. Considering both cases, two autoregressive methods stand out: order $r = 8$ and frame skip $s = 3$, giving a 240 ms time span, and $(r, s) = (8, 4)$, spanning 320 ms. With telephone data, both methods showed statistically significant improvement over every other evaluated method except each other and the AR method with $(r, s) = (16, 1)$. With the original, uncorrupted speech data, $(r, s) = (8, 4)$ showed statistically significant improvement over the other methods except for the AR (r,s) combinations $(4, 4)$, $(8, 3)$, $(6, 4)$ and $(10, 4)$.

Table 2. *EER scores (%) showing the effect of AR frame filtering and RASTA on detection performance in matched and mismatched test conditions with two different channels. The "original" case has been trained and evaluated on the original unprocessed 16 kHz data. The 8 kHz telephone case has been trained with car interior noise at SNR=30 dB and evaluated with factory noise at SNR=0 dB. The evaluations used FFT spectrum analysis and 16 components for each GMM.*

		Test condition	
AR order r	AR skip interval s	Original no mismatch	Telephone factory,SNR=0 dB
none		12.5	18.0
8	1	12.4	13.3
4	2	11.4	14.6
2	4	11.5	13.8
16	1	12.5	11.8
8	2	11.8	16.4
4	4	10.7	16.2
24	1	12.3	13.2
12	2	11.8	17.0
8	3	10.1	11.5
6	4	10.0	15.5
16	2	11.7	17.2
8	4	9.9	10.9
20	2	12.4	18.9
10	4	10.8	15.0
RASTA filtering		13.9	18.6

Finally, FFT- and LP-based MFCCs were again evaluated together, this time complemented with the proposed filtering method. Based on the previous results, the AR parameters were chosen as $(r, s) = (8, 4)$. With matched training on the original 16 kHz data and on telephone data with SNR 30 dB car noise, EER scores 7.7 % and 9.4 %, respectively, were obtained with LP-based MFCCs and 64 GMM components. Fig. 1 shows the corresponding DET curves. With perfect threshold selection, minimal half total error rate (HTER) [27] in both high-SNR cases was approximately 7 %. Table 3 shows the telephone results. Comparison with the results in Table 1 shows a statistically significant advantage for AR filtering in 5 (out of 8) cases with FFT and in 4 cases with LP. A larger number of GMM components led to significantly better telephone high-SNR performance but also decreased the performance significantly in a total of 6 out of 14 mismatched cases.

Table 3. *EER scores (%) for FFT- and LP-based MFCCs in combination with AR(8) feature filtering with a skip interval of 4 frames. The tests involve varying degrees of mismatch on telephone data. The detector has been trained using telephone speech with far-end car interior noise at SNR=30 dB.*

		16-comp. GMMs		64-comp. GMMs	
Noise	SNR (dB)	FFT	LP	FFT	LP
Car	30	11.8	10.9	9.4	9.4
	0	12.4	10.8	10.1	9.3
Factory	10	12.4	11.8	12.5	10.1
	0	10.9	11.4	17.2	13.9
	-10	22.6	25.1	29.7	21.8
Babble	10	12.4	10.9	11.4	10.9
	0	12.5	11.8	13.3	13.3
	-10	18.0	16.2	28.3	20.4

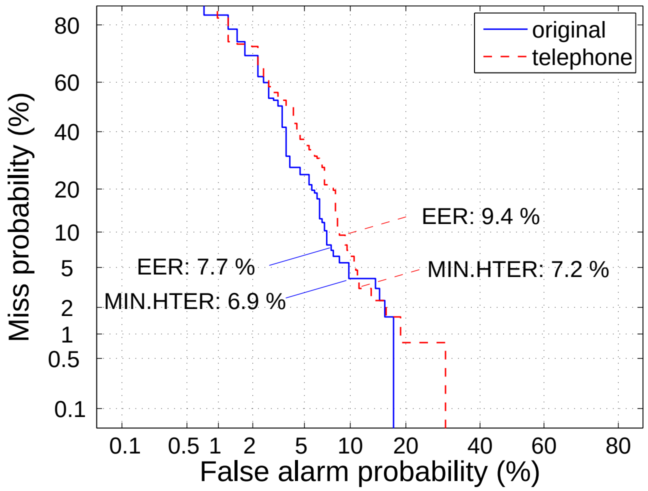

Fig. 1. DET curves for the detection system with original 16 kHz data and high-SNR telephone-channel data.

4. CONCLUSIONS

Detection of anger in speech was analyzed with a particular focus on realistic, adverse acoustic conditions involving the telephone channel. A new method for focusing on particular modulation frequencies of the features in classification was proposed and shown to lead to improved clean speech performance as well as to improved noise robustness. The method performs autoregressive (AR) prediction filtering across frames. The AR models have been learned to represent the dynamic behavior of the features within the target class of detection. In addition, linear predictive spectrum estimation for MFCC analysis showed a robustness advantage over the standard FFT, while increasing the number of GMM components had a detrimental effect on the noise performance in some cases. Future research directions include further studies of the proposed autoregressive feature processing approach in different paralinguistic classification tasks.

5. REFERENCES

[1] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. ICASSP*, Honolulu, Hawaii, Apr. 15–20 2007.

[2] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. Interspeech*, Pittsburgh, USA, Sept. 17–21 2006.

[3] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, pp. 768–785, 2011.

[4] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs," in *Proc. Interspeech*, Pittsburgh, USA, Sept. 17–21 2006.

[5] D. Neiberg and K. Elenius, "Automatic recognition of anger in spontaneous speech," in *Proc. Interspeech*, Brisbane, Australia, Sept. 22–26 2008.

[6] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, "Detecting real life anger," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 19–24 2009.

[7] M. Erden and L. M. Arslan, "Automatic detection of anger in human-human call center dialogs," in *Proc. Interspeech*, Florence, Italy, Aug. 28–31 2011.

[8] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Proc. ICASSP*, Toulouse, France, May 14–19 2006.

[9] A. Tawari and M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *Proc. Int. Conf. Pattern Recognition*, Istanbul, Turkey, Aug. 23–26 2010.

[10] T.-S. Chi, L.-Y. Yeh, and C.-C. Hsu, "Robust emotion recognition by spectro-temporal modulation statistic features," *J. Ambient Intell. Human Comput.*, vol. 3, pp. 47–60, 2012.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 4–8 2005.

[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - state-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, pp. 4–39, 2013.

[13] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[14] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Process. Lett.*, vol. 17, pp. 599–602, 2010.

[15] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo, "Noise robust feature extraction based on extended weighted linear prediction in LVCSR," in *Proc. Interspeech*, Florence, Italy, Aug. 28–31 2011.

[16] J. Pohjalainen, T. Raitio, H. Pulakka, and P. Alku, "Automatic detection of high vocal effort in telephone speech," in *Proc. Interspeech*, Portland, Oregon, USA, Sept. 9–13 2012.

[17] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," vol. 2, pp. 578–589, 1994.

[19] S. Greenberg, "On the origins of speech intelligibility in the real world," in *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997.

[20] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 3, pp. 116–125, 2012.

[21] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. ICASSP*, Honolulu, Hawaii, Apr. 15–20 2007.

[22] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, second edition, 2003.

[23] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Proc.*, vol. 3, pp. 72–83, 1995.

[24] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized Lloyd iteration," *IEEE Signal Process. Lett.*, vol. 1, pp. 144–146, 1994.

[25] Int. Telecommun. Union, *ITU-T G.191, Software tools for speech and audio coding standardization*, 2010.

[26] 3rd Generation Partnership Project, *3GPP TS 26.090, Adaptive multi-rate (AMR) speech codec, transcoding functions*, 2011, version 10.1.0.

[27] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. ODYSSEY04, The Speaker and Language Recognition Workshop*, Toledo, Spain, May 31–June 3 2004.