Detection of Shouted Speech in the Presence of Ambient Noise

Jouni Pohjalainen, Tuomo Raitio, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
jpohjala@acoustics.hut.fi, tuomo.raitio@tkk.fi, paavo.alku@aalto.fi

Abstract

This study focuses on the detection of shouted speech in realistic noisy conditions. An automatic system based on modified mel frequency cepstral coefficient (MFCC) feature extraction and Gaussian mixture model (GMM) classification is developed. The performance of the automatic system is compared against human perception measured by a listening test. At moderate noise levels, the automatic system outperforms humans. In severe conditions, classification by humans is clearly better.

Index Terms: shout detection

1. Introduction

The acoustic characteristics of shouted speech differ noticeably from normal speech. As systems for, e.g., automatic speech recognition or speaker recognition are increasingly used in adverse environments, they can be expected to also work with shouted speech. In such cases, it becomes necessary to automatically detect shouting in a realistic acoustic environment [1].

The field of automatic audio event detection has recently attracted research interest, e.g., [2] [3] [4] [5]. In these applications, the purpose is to automatically detect sounds related to alarming situations in a specific acoustic environment. Arguably the two most common classes of sounds that these systems aim to detect are shouted speech and explosive sounds [5]. Shouting in an environment normally occupied by non-vocal ambient noise and normal speech can be viewed as a generic indicator of a potentially hazardous situation. Therefore, reliable detection of shouted speech is an essential research topic in the emerging field of audio-based surveillance.

Previous studies have examined the detection of shouted speech [3] [6] and screams [4] [5] on the background of environmental noise. While specialized acoustic features have also been evaluated (e.g., [4] [5]), many studies have used the popular mel frequency cepstral coefficient (MFCC) features (e.g., [1] [2] [3] [5]). Most of the published systems use Gaussian mixture model (GMM) classification (literature overview in [5]). In studies analyzing the effects of typical noise, scream detection performance has been found to degrade steeply when the signal-to-noise ratio (SNR) is close to 0 dB [4] [5].

This study further develops and analyzes the system proposed in [6], based on specialized MFCC feature extraction and GMM classification. The emphasis is on noise robustness with realistic noise types and varying SNR. It is desired that a system trained with high SNR can work reliably across a range of SNR, i.e., regardless of whether the person shouting is close to or far away from the microphone. Differently from most other studies which also consider normal speech, the same textual material is used for normal and shouted speech, eliminating any possible help from phonemic differences. Importantly, a subjective listening test has been conducted and used to evaluate and compare human and computer performance in the detection task.

2. Description of the detection system

2.1. Feature extraction

The input signal is sampled at 16 kHz and pre-emphasized with FIR filter $H_p(z) = 1 - 0.97z^{-1}$. It is then arranged into overlapping Hamming-windowed frames of 25 ms with a shift interval of 10 ms. An MFCC feature vector is computed from each frame using the standard processing chain of squared magnitude spectrum computation (typically using FFT), triangular mel frequency filterbank, logarithm and discrete cosine transform [7].

In an earlier study [6], it was found useful to apply an alternative spectrum estimation method, in place of FFT, when computing the MFCC coefficients. In this method, the magnitude spectrum is computed in three steps. First, the magnitude spectrum envelope is computed using linear prediction (LP) analysis [8] with linear prediction order 20 and the autocorrelation method of LP. Second, the spectral fine structure, comprising F0 and its harmonics, is obtained by eliminating the spectrum envelope from the FFT magnitude spectrum using cepstral source-filter separation. Specifically, the signal is transformed into cepstral domain [7], liftered by suppressing to zero the cepstral coefficients corresponding to lags less than $(F_s/500) + 1$, where F_s is the sampling rate in Hz, and transformed back to the spectral domain. This way, periodic excitation information up to 500 Hz is retained in the liftered excitation spectrum. As the final step, the squared magnitude spectra from LP analysis (the spectrum envelope) and cepstral analysis (the fine structure/excitation spectrum) are multiplied with each other. The resulting magnitude spectrum is termed linear predictive-cepstral residual (LP-CR) spectrum.

Figure 1 shows the FFT spectrum, the LP envelope, the excitation spectrum, and the LP-CR spectrum for one noisy speech vowel frame. It can be observed that while in FFT analysis the formants are drowning under the noise floor, LP-CR analysis tries to preserve harmonics associated with the major formants. A possible explanation is that the LP envelope amplifies the locations of the harmonics in the excitation spectrum.

2.2. Frame selection

The features are analyzed in blocks of two seconds and the modeling concentrates on the high-energy frames within each block. This is done in both the training phase and the detection phase. A frame selection method is used to automatically select the high-energy frames inside each two-second analysis block. First, the logarithmic energy is computed for each short-time analysis frame, i.e., every 10 ms. This results in a sequence of 200 energy values, denoted by E_n. The purpose is to automatically classify the sequence into high and low values. Two alternative methods, explained below, can be used.

The simpler frame selection method is an application of k-means clustering [7], in which the centers of two clusters are

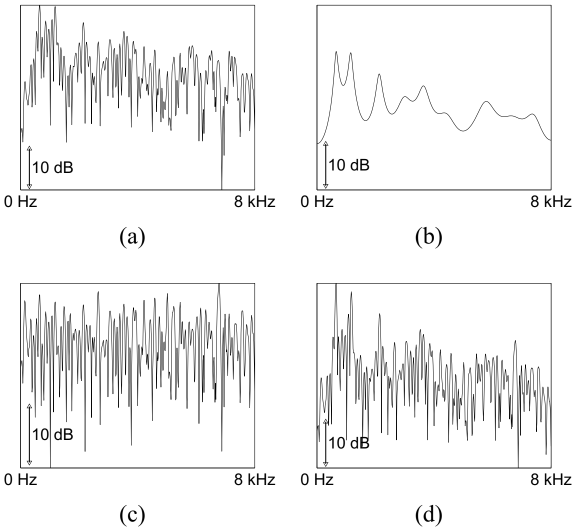

Figure 1: *Spectra computed from a noisy /a/ vowel frame spoken by a male speaker: a) magnitude spectrum given by FFT; b) magnitude spectrum envelope given by LP; c) excitation spectrum given by cepstral source-filter separation; d) LP-CR spectrum given by combining b) and c).*

initialized with $\min(E_n)$ and $\max(E_n)$. After the k-means iteration converges, denote the cluster assignment as $X_n = 1$, if E_n belongs to the cluster whose center was initialized with $\max(E_n)$, and $X_n = 0$ otherwise. The method selects for further processing the frames for which $X_n = 1$.

Another frame selection method is based on unsupervised training of an ergodic hidden Markov model (HMM) [7] with two states and a univariate Gaussian density characterizing the observations in each state. The initial state probabilities and state transition probabilities of the HMM are initialized with uniform distributions. The state specific variance parameters of the Gaussian densities are both initialized with 0.1 times the global variance of the sequence E_n. The state specific mean parameters of the Gaussian densities are initialized, similarly to the k-means method, with $\max(E_n)$ and $\min(E_n)$. The HMM parameters are then estimated using an implementation of EM re-estimation principle [7]. After convergence, denote the inferred states of the HMM as $X_n = 1$ if E_n was produced by the state whose mean value was initialized with $\max(E_n)$, and $X_n = 0$ otherwise. The frames for which $X_n = 1$ are selected.

2.3. Detection

The detection system uses Gaussian mixture models (GMMs) to model broad sound classes. Each GMM has 8 components and a diagonal covariance structure. The GMMs are trained using 10 iterations of EM re-estimation for GMMs [9]. Before training, the GMMs are initialized by uniform mixture weights, the variance parameters of each component 0.1 times the global variances of the features, and the mean parameters of each component given by the heuristic approach proposed in [10].

Depending on the chosen classification rule, different GMMs have to be trained. In this work, separate GMMs are trained for (1) shouted speech, (2) normal speech, (3) ambient noise and (4) non-shouting, which consists of normal speech and ambient noise together. The training data for shouted speech and normal speech is clean, i.e., not corrupted by noise. In the detection phase, after the high-energy frames inside a 2 s analysis block (with 1 s shift interval) have been selected using one of the frame selection methods described in Section 2.2, the averaged log likelihoods of their corresponding feature vectors having been produced by the different GMMs are computed and denoted as L_{shout}, L_{speech}, L_{noise} and L_{nonshout}.

In this work, three classification rules are examined. First, the detection is considered a direct binary classification problem and treated according to the Bayes rule. In this method, denoted as “D2”, the logarithmic likelihood ratio decision statistic is defined as $L_{\text{shout}} - L_{\text{nonshout}}$. In another direct classification approach, denoted as “D3” because it uses three GMMs instead of two, the nonshout score is replaced by a maximum of the speech and noise scores, so that the decision statistic becomes $L_{\text{shout}} - \max(L_{\text{speech}}, L_{\text{noise}})$.

In many proposed systems for audio event detection, hierarchical classification structures are preferred. Thus, the direct classification rules are compared against a hierarchical classification tree rule in which the first stage separates shouting and noise from normal speech according to the statistic $\max(L_{\text{shout}}, L_{\text{noise}}) - L_{\text{speech}}$ and the second stage separates shouting from other sounds according to the statistic $L_{\text{shout}} - \max(L_{\text{speech}}, L_{\text{noise}})$ (which is the same as the decision rule D3). This classification rule is denoted as “H3”.

3. Experimental evaluation

3.1. Test material and test setup

Two types of noise from the NOISEX-92 database [11] were used to simulate two different acoustic environments. The *factory1* noise is mechanical noise from a factory including frequent transient impulsive sounds. The *babble* noise contains many people talking simultaneously in a canteen.

The speech material was recorded with high quality equipment in an anechoic chamber. Speakers (11 males and 11 females) produced 24 Finnish sentences using normal speech and shouting. The sufficient level of shouting was controlled by listening and by monitoring the sound pressure level. Twelve of the sentences are in the imperative mood, consisting of one to four words, with a message that could plausibly be uttered in a potentially threatening situation, such as “anna se kamera tänne” (“give me the camera”), “pysy siinä” (“stay there”) and “juoskaa” (“run”). The other 12 sentences, consisting of three words, are in the indicative mood and have a neutral, abstract information content.

The experiments were carried out as leave-one-out cross validation. One speaker in turn was selected as the test speaker while the other 21 speakers’ material was used to train the models. The test material for each speaker consisted of his or her speech and shout material, both corrupted by noise with a given segmental, or frame-averaged, SNR, as well as a segment of noise equal in length to the speaker’s normal speech material. The noise model was trained using two minutes of the noise material, while the remaining portion of the noise recording was used for testing. The primary performance measure was the equal error rate (EER), a common metric to assess the quality of a two-class detector (used, e.g., in [5]). The EER corresponds to the decision threshold for which the miss and false alarm rates are equal. Another performance measure used in this study is constrained false alarm error rate, in which an upper limit is imposed on the false alarm rate and the equally-weighted total error rate is found using the threshold corresponding to this. For a hierarchical detector, the minimum scores of EER and the constrained false alarm measure are found by evaluating the score

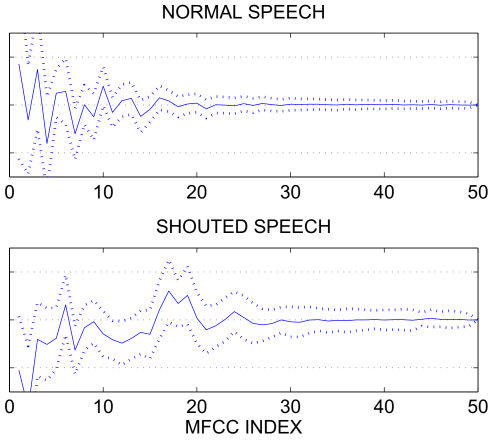

Figure 2: *Mean values (solid line) and standard deviation intervals of MFCCs averaged over normal and shouted speech from one male speaker.*

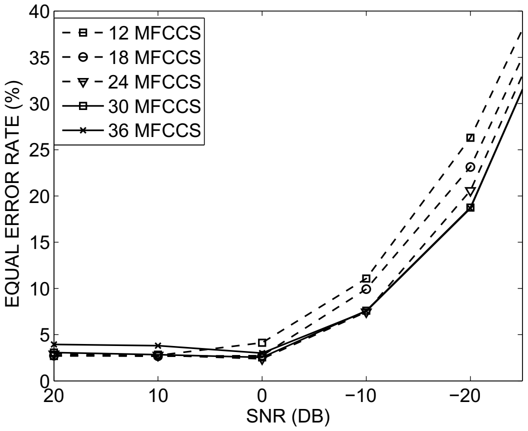

Figure 3: *Effect of the number of MFCC coefficients on the performance (measured by EER) of the shout detection system with different levels of noise. The MFCCs were computed using FFT spectrum analysis.*

resulting from the second stage separately for every possible first stage decision threshold value. Unless otherwise noted, the D2 decision rule is used and the presented results are combined by averaging the scores from factory noise and babble noise.

3.2. Effect of the form of the MFCC feature vector

In applications such as automatic speech recognition, 12 MFCCs concatenated with their first and second order delta coefficients are typically used to form the feature vector [7]. With this length of the MFCC vector, however, the spectral fine structure cannot be captured. In contrast, by using a longer MFCC vector, the spectral fine structure can at least partially be captured even though it is partially smoothed out by the mel filterbank. In shouting, F0 increases compared to normal speech. Consequently, the spectral fine structure becomes sparse and sound energy is increasingly located at the few harmonics. Therefore, the role of the spectral fine structure is manifested in shouting and it is justified to be taken into account in the feature vector by extending the length of the MFCC vector.

Figure 2 shows the mean and standard deviation of the MFCC vector for both normal and shouted speech of one male speaker. Evidently, large differences between the two classes of speech manifest themselves also in the MFCC coefficients with index 13 to 30, which are usually not included in an MFCC feature vector. Figure 3 shows the effect of the number of MFCC coefficients on the baseline detection system with FFT-based MFCCs, k-means frame selection and D2 classification. Based on this experiment, the number of MFCC coefficients was fixed at 30 for the rest of the experiments.

The inclusion of the first and second order delta coefficients of the MFCCs in the feature vector was investigated. While in the case of 12 MFCCs the performance was slightly improved when the delta coefficients were included, with 30 MFCCs inclusion of the delta coefficients could not be deemed to improve the performance. The performance of 30 MFCCs without deltas was better than the performance of 12 MFCCs with deltas. Thus, delta coefficients were not included in further evaluations.

3.3. Effect of frame selection and spectrum analysis

The effect of the frame selection method (Section 2.2) in both training and testing phase was investigated. The best combination obtained was to use the HMM-based frame selection in the training phase and the k-means frame selection in the detection phase. The good performance of this choice could be explained by the fact that the HMM segmentation, which otherwise performs beneficial temporal smoothing of the frame segmentation, can more easily be distracted by noise, while the k-means method always selects the high energy frames in which the SNR is largest. For this combination, the MFCC features based on LP-CR spectrum analysis performed better than the conventional FFT-based MFCC features, as shown in Table 1. With k-means frame selection in training and none in detection, for example, 30 LP-CR-based MFCCs gave scores 5.4 (SNR=20 dB), 4.2, 4.9, 7.6, 19.1 and 43.3 (SNR=-30 dB). Further tests used LP-CR features and HMM/k-means frame selection.

Table 1: *EER scores (%) for two MFCC spectrum analysis methods using HMM/k-means frame selection.*

Spectrum analysis	Signal-to-noise ratio (dB)					
	20	10	0	-10	-20	-30
FFT	4.0	3.6	2.9	7.4	18.2	43.8
LP-CR	3.0	2.7	3.3	6.3	16.4	44.5

3.4. Effect of the classification rule

Three different classification rules were introduced in Section 2.3. Their results for the two different types of noise are tabulated separately in Table 2. There appears to be no difference between the hierarchical rule H3 and the direct rule D3, but they are slightly better than the simplest direct rule D2.

Table 2: *EER scores (%) for three classification rules, defined in Section 2.3, with two types of noise and varying SNR.*

Noise type	Rule	Signal-to-noise ratio (dB)					
		20	10	0	-10	-20	-30
Factory	D2	2.8	2.6	3.2	8.1	17.4	45.5
	D3	2.4	2.4	3.6	6.8	15.8	45.4
	H3	2.4	2.4	3.6	6.5	15.8	45.4
Babble	D2	3.2	2.9	3.3	4.6	15.4	43.5
	D3	2.9	2.7	3.0	4.4	15.2	42.8
	H3	2.9	2.7	3.0	4.3	15.2	42.8

3.5. Human performance

A subjective listening test was performed in order to compare the shouting detection accuracy between computer and humans. The same speech/shouting material and samples with pure noise were included in the test. Babble noise conditions with signal-to-noise ratios of 0, -10, -20, and -30 dB were evaluated. In the listening test, samples were presented to the listener through headphones and the task of the listener was to decide, whether a sample contained shouting or not. The listener could listen to each test sample as many times as desired before decision.

The test was performed in a quiet room with a graphical user interface. High-quality headphones (Sennheiser HD580) were used for listening. The loudness of listening test samples was normalized according to ITU-T P.56. Before the test, test subjects performed a practice session with 10 samples, similar to the test samples. During the practice session, the listeners were allowed to adjust the volume of the headphones to a comfortable level. During the test, the volume was kept constant.

The listening test material consisted of 22 speakers both speaking and shouting 24 sentences. Each sentence was presented with four SNRs. Thus, the total number of sentences is $22 \times 2 \times 24 \times 4 = 4224$. In addition, a quarter of that number (1056) of pure babble noise samples were added. As a result, a total of 5280 listening test samples were created.

The test material was divided evenly among 16 naive Finnish listeners, of which 8 were male and 8 female, thus corresponding to the average population. Each listener evaluated a total of 330 test cases, consisting of 264 speech/shout cases and 66 pure babble noise cases. The duration of the test per listener was about 30 minutes.

The listening test error score was obtained by calculating, for each SNR, $0.5P^h_{\text{miss}} + 0.5P^h_{\text{fa}}$, where the miss rate estimate P^h_{miss} is based on the shouting samples and the false alarm rate estimate P^h_{fa} is based on an equal number of speech samples and babble noise samples. Table 3 shows these scores.

Table 3: *Shout detection error scores given by the listening test.*

	Signal-to-noise ratio (dB)			
	0	-10	-20	-30
Total score %	5.1	5.7	9.0	33.8
Miss rate %	8.7	10.0	16.0	65.6
False alarm rate %	1.4	1.5	1.9	2.1

Figure 4 illustrates the total error and compares it against the scores of the automatic system (D3 rule) using two methods of operating point selection: the EER criterion, where the threshold makes the miss rate P^c_{miss} and the false alarm rate P^c_{fa} equal, and a criterion in which P^c_{fa} is set to be no higher than the maximum P^h_{fa} of the human listeners (2.1 % at SNR -30 dB). In the latter case, the error score is obtained as $0.5P^c_{\text{miss}} + 0.5P^c_{\text{fa}}$, where the decision threshold corresponds to the chosen P^c_{fa}.

4. Conclusions

A system for automatic detection of shouted speech was described. Experiments showed that MFCC vectors consisting of up to 30 coefficients improve the performance over conventional, shorter MFCC vectors. In spectrum analysis for MFCC computation, linear predictive envelope multiplied by cepstrally separated excitation spectrum gave better results than the conventional FFT. Unsupervised frame selection was observed to be beneficial. Inclusion of delta coefficients or using a hierarchical classification approach were not found to improve the performance, but using separate models for shouting, normal speech and noise slightly outperformed direct binary detection.

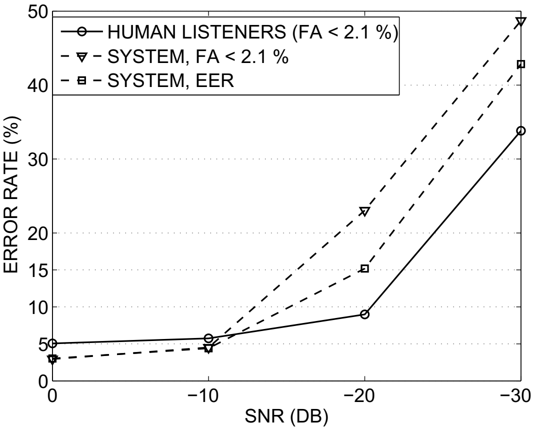

Figure 4: *The total error rate of human listeners compared with two error measures of the automatic detection system, corresponding to different operating points: low false alarm rate, similar to the human listeners, and equal error rate. Babble noise was used. Equal numbers of shouting and non-shouting samples are implicitly assumed with each error measure.*

The automatic system outperformed human listeners at babble noise SNR levels -10 dB and 0 dB, while human listeners showed better detection in the very noisy conditions. Future work should further improve the robustness in high noise conditions so as to catch up with the human level of performance.

5. Acknowledgements

This work was supported by Academy of Finland (127345).

6. References

[1] Nanjo, H., Mikami, H., Kawano, H. and Nishiura, T., "A Fundamental Study of Shouted Speech for Acoustic-Based Security System", in Proc. Interspeech, Brighton, UK, Sep 2009.

[2] Radhakrishnan, R., Divakaran, A. and Smaragdis, P., "Audio Analysis for Surveillance Applications", in Proc. IEEE WASPAA, New Paltz, USA, Oct 2005.

[3] Rouas, J.-L., Louradour, J. and Ambellouis, S., "Audio Events Detection in Public Transport Vehicle", in Proc. IEEE Intelligent Transportation Systems Conf., Toronto, Canada, Sep 2006.

[4] Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F. and Sarti, A., "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems", in Proc. IEEE Int. Conf. Advanced Video and Signal based Surveillance, London, UK, Sep 2007.

[5] Ntalampiras, S., Potamitis, I. and Fakotakis, N., "An Adaptive Framework for Acoustic Monitoring of Potential Hazards", EURASIP J. on Audio, Speech, and Music Processing, 51(5):401–411, 2009.

[6] Pohjalainen, J., Alku, P. and Kinnunen, T., "Shout Detection in Noise", in Proc. ICASSP, Prague, Czech Republic, May 2011.

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

[8] Makhoul, J., "Linear prediction: a tutorial review", Proceedings of the IEEE, 63(4):561–580, 1975.

[9] Reynolds, D. A. and Rose, R. C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech and Audio Proc., 3(1):72–83, 1995.

[10] Katsavounidis, I., Kuo, C.-C. J. and Zhang, Z., "A New Initialization Technique for Generalized Lloyd Iteration", IEEE Signal Processing Letters, 1(10):144–146, 1994.

[11] NOISEX-92 database, samples available online: http://spib.rice.edu/spib/select_noise.html.