SHOUT DETECTION IN NOISE

Jouni Pohjalainen[1]*, Paavo Alku*[1]*, Tomi Kinnunen*[2]

[1] Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
[2]University of Eastern Finland, School of Computing, Joensuu, Finland

ABSTRACT

For the task of detecting shouted speech in a noisy environment, this paper introduces a system based on mel frequency cepstral coefficient (MFCC) feature extraction, unsupervised frame dropping and Gaussian mixture model (GMM) classification. The evaluation material consists of phonemically identical speech and shouting as well as environmental noise of varying levels. The performance of the shout detection system is analyzed by varying the MFCC feature extraction with respect to 1) the feature vector length and 2) the spectrum estimation method. As for feature vector length, the best performance is obtained using 30 MFCC coefficients, which is more than what is conventionally used. In spectrum estimation, a scheme that combines a linear prediction spectrum envelope with spectral fine structure outperforms the conventional FFT.

Index Terms— shout detection

1. INTRODUCTION

Recently, several audio surveillance systems have been proposed to detect abnormal or potentially alarming sounds in specific acoustic environments. Examples include the detection of non-neutral speech and banging in elevator [1], the detection of shouts in train [2] and the detection of screams, gunshots and explosions in urban or military environments [3].

It can be argued that shouting is a quite generic acoustic indicator of a potentially hazardous situation in an environment typically characterized by normal speaking voices and non-vocal environmental sounds. Shouting in such an environment is typically associated with some degree of urgency. Hence, reliable detection of shouted speech in noisy environments is an essential research topic in the area of audio surveillance technology. This topic will be addressed in the present paper by proposing a system using which the performance of several techniques in shout detection can be compared.

Previous studies have examined the detection of shouted speech [2] [4] or screams [5] [6] [3] apart from environmental noise, often also including normal speech as test material [2] [4] [3]. Differently from previous approaches, the present study uses exactly the same textual material for both shouted speech and normal speech. It can be argued that this scenario is more challenging, because when the shouts and normal speech share the same phonemic content, phonemic differences between the two classes cannot aid the detection. In some previous studies, the robustness of scream detection with respect to decreasing signal-to-noise ratio (SNR) has been examined [5] [6] [3] and the performance has been found to degrade steeply when the SNR is close to 0 dB. This degradation has sometimes been tackled by training the shout/scream models with data that already contains the expected type and amount of noise corruption [2] [6], but this calls for a complete retraining whenever the noise environment changes and, as noted in [6], increases the number of false alarms. In practice, the distance between the microphone and the person shouting determines the SNR, and it is desirable that the performance is independent of whether the person shouting is close to or further away from the microphone. Clearly, there is a demand for techniques that improve the noise robustness of shout and scream detection. Towards this end, the present study trains the system on clean (not noisy) vocal data and investigates the degradation of performance as the SNR decreases. This is done using two different realistic noise types: factory noise and large crowd babble.

The proposed shout detection system is based on two well-known audio recognition techniques: feature extraction based on mel frequency cepstral coefficients (MFCC) and classification using Gaussian mixture models (GMMs), both of which have been popular in previous audio surveillance systems, e.g. [1] [2] [3]. These are complemented with several techniques to improve the robustness of shout detection in adverse noise conditions. In particular, the conventional Fourier-based spectrum estimation in the MFCC computation is replaced with new methods that combine linear predictive spectral envelope modeling with spectral fine structure, i.e., the fundamental frequency (F0) and its harmonics. Hence, the importance of F0 and its multiple integers can be increased in the MFCC feature extraction, a goal that is justified by the fact that shouting in speech communication correlates with the use of high pitch [7]. In addition, the number of MFCC coefficients is varied with different spectral modeling techniques in order to better capture shout-discriminating characteristics. The proposed system utilizes an unsupervised time series segmentation method for energy-based frame dropping prior to GMM training and detection.

2. SHOUT DETECTION SYSTEM

2.1. MFCC feature extraction

The input to the system is sampled at 16 kHz and pre-emphasized with $H_p(z) = 1 - 0.97z^{-1}$. The signal is processed in Hamming-windowed frames of 25 ms with a 10 ms interval between two frames. Fig. 1 shows the complete chain of MFCC computation used in the present work.

The feature extraction uses MFCCs as a representation for the short-time magnitude spectrum [8]. Different methods are evaluated for the estimation of the magnitude spectrum. The fast Fourier transform (FFT) is the conventional spectrum estimation method for the MFCC computation. Recently, the present authors have investigated the use of different forms of linear predictive models in the MFCC feature extraction for automatic speech recognition and speaker verification in adverse conditions. In particular, weighted linear prediction (WLP) and its variants have led to improved robustness in these applications, e.g. [9] [10] [11]. The explanation of LP and WLP is

The work was supported by Academy of Finland projects 127345 and 132129.

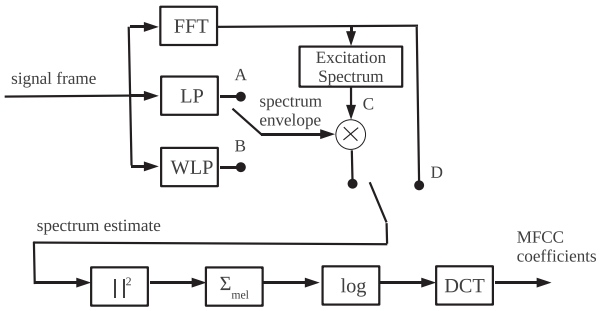

Fig. 1. Diagram of the MFCC feature extraction.

deferred to Section 2.2.

FFT spectrum estimation preserves the spectral fine structure, while LP and WLP (with the typical choice of prediction order $p = 20$ for 16 kHz sampling rate, as used in this work) only depict the spectral envelope. Because shouting clearly has an effect on the vocal tract excitation (e.g., increased F0) and the spectral fine structure is related to the vocal tract excitation, using the all-pole spectrum envelope alone may be sub-optimal for the detection of shouts. To include the spectral fine structure, the all-pole spectral envelope is multiplied by the excitation spectrum obtained by liftering the FFT-based cepstrum in the cepstral domain and transforming the liftered cepstrum back into the spectral domain. The lifter used for this purpose forces to zero the cepstral coefficients corresponding to lags less than $(F_s/500) + 1$, where F_s denotes the sampling rate in Hz. This means that periodic vocal tract excitation information up to 500 Hz is retained in the liftered excitation spectrum. Figs. 1 and 2 illustrate the idea.

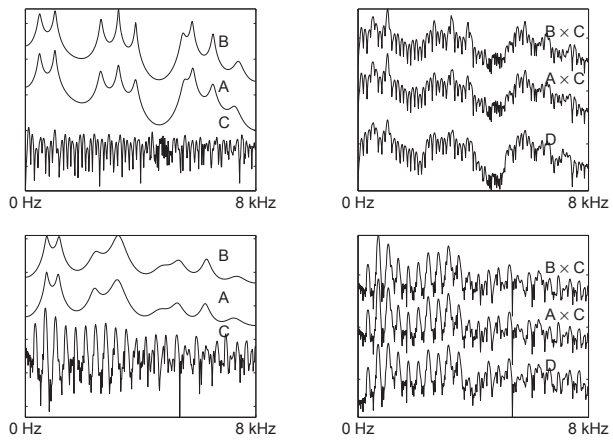

Fig. 2. The vowel /o/ spoken normally (top) and shouted (bottom) by a male speaker. LP and WLP spectrum envelopes and the cepstrally liftered excitation spectrum (left) are used to construct spectrum estimate alternatives (right) to the FFT spectrum. The notations A, B, C and D correspond to Fig. 1.

Once a spectrum estimate is obtained using some of the above methods, it is squared and passed through a conventional mel filterbank with 40 mel filters. After converting the mel spectrum into a logarithmic form, it is transformed by the discrete cosine transform (DCT) to yield the desired number of MFCC coefficients (the so-called "zeroth" MFCC coefficient, which is related to the energy of the analysis frame, is not used). The longer the MFCC vector, the more of the detail left in the mel spectrum it preserves. By examining the mean and variance of the MFCC coefficients in speech and shouting, it was observed that these two classes differ mainly in MFCC coefficients with index below 30. Therefore, in order to analyze the effect of the MFCC vector length, the values 8, 12, 20 and 30 were chosen for the experimental evaluation. The effect of concatenating the basic 12 MFCC coefficients with their delta and double-delta coefficients to depict their instantaneous trajectories was also investigated, but this was not found to improve detection performance.

2.2. Linear predictive spectrum envelope estimation

Linear predictive (LP) speech spectrum modeling [12] assumes that each speech sample is predictable as a linear combination of p previous samples, $\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}$, where s_n are the samples of the speech signal in a given short-term frame. The $\{a_k\}$ are the predictor coefficients, which in the frequency domain form an all-pole filter $H(z) = 1/(1 - \sum_{k=1}^{p} a_k z^{-k})$. The number of predictor coefficients p is the *order* of linear prediction. Conventional LP analysis minimizes the energy of the prediction error signal $E_{\mathrm{LP}} = \sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2$ by setting the partial derivatives of E_{LP} with respect to each coefficient a_k to zero. This results in the normal equations [12] $\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-j} = \sum_n s_n s_{n-j}$, $1 \le j \le p$. The range of summation of n is chosen in this work to correspond to the *autocorrelation method*, in which the energy is minimized over a theoretically infinite interval, but s_n is considered to be zero outside the analysis window [12].

WLP, originally introduced in [13], is a generalization of LP that introduces a temporal weighting of the squared prediction error in model coefficient optimization. Specifically, in WLP, the predictor coefficients $\{b_k\}$ are solved by minimizing the energy $E_{\mathrm{WLP}} = \sum_n (s_n - \sum_{k=1}^{p} b_k s_{n-k})^2 W_n$, where W_n is the weighting function chosen as the short-time energy of the immediate signal history: $W_n = \sum_{i=1}^{p} s_{n-i}^2$. This kind of weighting can be used to emphasize the importance of the prediction error in the high-energy regions assumed to be less affected by noise (large local signal-to-noise ratio), and de-emphasize the importance of modeling the noisier low-energy regions. The WLP model is obtained by solving the normal equations $\sum_{k=1}^{p} b_k \sum_n W_n s_{n-k} s_{n-i} = \sum_n W_n s_n s_{n-i}$ for all $1 \le i \le p$.

2.3. Unsupervised segmentation for energy-based frame selection

For long-time processing of the features, frame dropping is implemented using an unsupervised time series segmentation method. This step is performed in both the training and testing phases. Within each long-time analysis block of two seconds, the logarithmic frame energy is computed for each short-time analysis frame, i.e., every 10 ms. This sequence of 200 energy values, denoted by E_n, is segmented into "low state" and "high state" using unsupervised training of a univariate Gaussian density ergodic hidden Markov model (HMM) with two states. The HMM is parametrized by the parameter set $\lambda = \{(a_{ij}), \pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$, where (a_{ij}) denotes the 2×2 state transition probability matrix, π_1 and π_2 are the initial state probabilities, μ_1 and μ_2 are the state density means and σ_1^2 and σ_2^2 are the state density variances. The most important aspect of this method is the initialization of the HMM parameters.

While the probability parameters (a_{ij}), π_1 and π_2 are initialized conventionally using uniform distributions [14], the state means are initialized as $\mu_1 = \max(E_n)$ and $\mu_2 = \min(E_n)$. For the variance parameters, the initialization $\sigma_1^2 = \sigma_2^2 = R \cdot \mathrm{var}(E_n)$ with $R = 0.1$ has been found to yield satisfactory results. From these initial values, the HMM parameters are estimated in a typical fashion using an implementation of the EM re-estimation principle [14].

During each iteration, the EM algorithm estimates the probability distribution $(\gamma_n(1), \gamma_n(2))$ for being in state 1 or state 2 at time instant n. To convert these values to a segmentation in terms of the two states, we simply take $X_n = 1$ if $\gamma_n(1) \geq \gamma_n(2)$ and $X_n = 0$ otherwise, for all n. The EM re-estimation is deemed to have converged and the iteration is stopped once this segmentation X_n does not change between two successive EM iterations. Since state 1 was initialized with the maximal value and state 2 with the minimal value, we can safely assume that state 1 will be the "high state" for the frame energy time series. The additional benefit of using an HMM for this purpose instead of simple unsupervised threshold determination is that the HMM smoothes the segmentation in time. All successive processing for the block is done using only the high-state frames for which $X_n = 1$. Modeling and recognizing only these high-energy frames is justifiable because these frames presumably have the best local SNR. Fig. 3 shows the evolution of the state segmentation X_n with EM iterations for a noisy (SNR 0 dB factory noise) two-second speech segment.

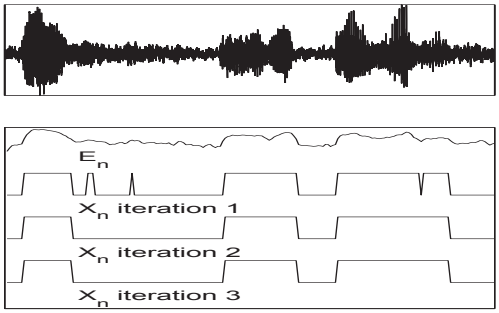

Fig. 3. Evolution of HMM segmentation with EM iterations for the log energy sequence of a noisy speech segment.

2.4. GMM classification

The classifier uses a specialized GMM to model each of the three primary classes: the noise environment, normal speech (both male and female) and shouting (both male and female). These GMMs $\lambda_{\mathrm{environment}}$, $\lambda_{\mathrm{speech}}$ and λ_{shout} have 8 mixture components with a diagonal covariance structure [15]. For initializing the component mean vectors of the GMMs, the simple heuristic approach proposed in [16] is used. The variances of each variable in each component are initialized by 0.1 times the variable's global variance over the training data. The mixture weights are initialized with uniform distributions. Each GMM is trained by running four EM iterations [15]. For the number of mixture components in each model, both 8 and 16 were evaluated and the former was selected because it provided slightly better performance.

When the GMMs are used in detection, the audio signal is processed in blocks of two seconds, with a block shift of one second. After the high-energy frames have been selected using the HMM frame dropping detailed in Section 2.3, the averaged log likelihoods of them having been produced by each of the three GMMs are computed and denoted as L_{shout}, L_{speech} and $L_{\mathrm{environment}}$. The shouting detection is considered as a binary classification problem and treated according to the Bayes rule. The logarithmic likelihood ratio decision statistic is defined as $L = L_{\mathrm{shout}} - L_{\mathrm{nonshout}}$. The shout score is the shout GMM likelihood and the non-shout score is obtained as the maximum of the speech and noise GMM likelihoods as $L_{\mathrm{nonshout}} = \max(L_{\mathrm{speech}}, L_{\mathrm{environment}})$. For each detection block, the statistic L can be recorded and used in evaluating the system performance with variable detection threshold.

3. EXPERIMENTAL EVALUATION

3.1. Test material and setup

To represent different types of acoustic environments, two types of noise from the NOISEX-92 database were used. The *factory1* noise contains machine noise with frequent transient impulsive sounds. The *babble* noise contains many people talking simultaneously in a cafeteria-like environment.

The speech and shouting was recorded with high quality equipment in an anechoic chamber. The data consists of 11 male and 11 female speakers, each speaking 24 Finnish sentences, both in a normal fashion and by shouting. The shouting was controlled both by listening and by monitoring the sound pressure level. A mere raised voice was not accepted as shouting. Twelve of the sentences are sentences in the imperative mood, consisting of one to four Finnish words, with a message that could plausibly be uttered in a potentially threatening situation, such as "anna se kamera tänne" ("give me the camera"), "älkää liikkuko" ("don't move") and "lopettakaa" ("stop it"). The other 12 sentences consist of three Finnish words, are in the indicative mood and have a neutral, abstract information content.

The experiments were carried out as leave-one-out cross validation. Each speaker in turn was selected as the test speaker, and data from the other 21 speakers was used to train the speech and shout models. The test material for each speaker consisted of that speaker's 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 combined speech and shout 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 measure to assess the performance is the equal error rate (EER), a common metric to assess the quality of a two-class detector. The EER corresponds to the decision threshold for which the miss and false alarm rates are equal.

3.2. Results

Tables 1 and 2 show the shout detection results for the NOISEX-92 *factory1* and *babble* noises, respectively. In the case of 12 MFCCs, results are also shown by using only the LP or WLP spectrum envelope without the excitation spectrum. The usefulness of including the excitation spectrum is easily observed. Several different types of features give good performance at low to moderate noise levels. However, at SNR levels -10 dB and -20 dB, at which the system performance is degrading rapidly, the most resistant features are 30 MFCCs obtained using LP or WLP envelope combined with the excitation spectrum.

Table 1. *Shout detection EER scores (%) for* factory1 *noise. Excitation spectrum is denoted by "ex.".*

Spectrum estimation	# MFCCs	Signal-to-noise ratio (dB) 20	10	0	-10	-20	-30
FFT	8	2.3	3.2	3.9	13.9	27.7	49.4
FFT	12	2.4	2.5	3.2	12.2	28.1	50.2
FFT	20	2.5	2.3	3.3	11.5	21.3	48.2
FFT	30	2.5	2.7	2.9	10.1	20.2	46.0
LP + ex.	8	4.0	3.9	4.6	7.6	22.4	46.6
LP + ex.	12	2.7	2.3	3.3	6.6	21.2	46.4
LP	12	3.9	4.3	5.6	10.3	22.0	45.4
LP + ex.	20	4.7	4.7	4.8	10.2	21.6	50.9
LP + ex.	30	2.9	3.1	3.0	6.1	17.0	45.9
WLP + ex.	8	4.0	3.4	4.5	8.4	22.4	46.1
WLP + ex.	12	3.3	2.8	3.4	7.6	19.9	44.9
WLP	12	4.0	4.3	6.8	12.1	22.4	45.6
WLP + ex.	20	3.9	3.6	3.4	8.8	18.7	48.6
WLP + ex.	30	2.6	2.5	3.3	6.8	18.4	46.6

Table 2. *Shout detection EER scores (%) for* babble *noise. Excitation spectrum is denoted by "ex.".*

Spectrum estimation	# MFCCs	Signal-to-noise ratio (dB) 20	10	0	-10	-20	-30
FFT	8	2.2	2.8	3.9	11.0	24.2	45.6
FFT	12	2.6	2.9	3.2	9.5	22.8	46.0
FFT	20	2.2	2.5	3.0	7.1	21.7	46.4
FFT	30	2.6	2.3	2.1	5.1	19.8	45.0
LP + ex.	8	4.0	4.0	4.2	6.1	19.4	44.1
LP + ex.	12	2.8	2.7	2.9	4.7	17.0	43.8
LP	12	3.5	4.3	4.6	7.7	22.9	45.6
LP + ex.	20	4.7	4.6	4.5	6.7	16.0	41.9
LP + ex.	30	3.2	2.9	3.5	4.6	15.6	42.9
WLP + ex.	8	3.3	3.6	3.6	6.0	20.4	44.0
WLP + ex.	12	3.4	3.0	3.3	6.5	19.4	45.1
WLP	12	4.3	4.9	4.9	9.9	23.4	46.6
WLP + ex.	20	3.9	3.4	3.5	5.2	16.9	44.5
WLP + ex.	30	2.2	2.4	2.2	4.7	15.2	43.8

4. CONCLUSIONS

This study introduced a system for shout detection, including new methods for feature extraction and energy-based segmentation. The emphasis was on noise robustness with respect to realistic environmental noises. With all evaluated MFCC features, the system showed reasonably good performance with increasing noise until at least 0 dB SNR level. The largest differences between different features were observed in the noisier cases. The overall best results were obtained with feature vectors consisting of 30 MFCCs. A possible explanation is that they preserve more spectral fine structure than shorter MFCC vectors and thus contain more information about vocal tract excitation. Concerning spectrum estimation in MFCC computation, the best noise performance was, instead of the conventional FFT, shown by the composite spectra obtained by multiplying an LP or WLP spectrum envelope with the excitation spectrum.

5. REFERENCES

[1] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *Proc. IEEE WASPAA*, New Paltz, USA, October 2005.

[2] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *Proc. IEEE Intelligent Transportation Systems Conf.*, Toronto, Canada, September 2006, pp. 733–738.

[3] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP J. on Audio, Speech, and Music Processing*, 2009.

[4] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 813–816.

[5] M. Vacher, D. Istrate, L. Besacier, J.-F. Serignat, and E. Castelli, "Sound detection and classification for medical telesurvey," in *Proc. Int. Conf. Biomedical Engineering*, Innsbruck, Austria, February 2004, pp. 395–399.

[6] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Int. Conf. Advanced Video and Signal based Surveillance*, London, UK, September 2007.

[7] P. Alku, J. Vintturi, and E. Vilkman, "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal, and loud phonation," *Speech Communication*, vol. 38, no. 3-4, pp. 321–334, 2002.

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

[9] J. Pohjalainen, H. Kallasjoki, K.J. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis in noisy conditions," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 1315–1318.

[10] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[11] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 1477–1480.

[12] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[13] C. Ma, Y. Kamp, and L.F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 2, pp. 69–81, 1993.

[14] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[15] 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, no. 1, pp. 72–83, 1995.

[16] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized Lloyd iteration," *IEEE Signal Processing Letters*, vol. 1, no. 10, pp. 144–146, Oct. 1994.