Automatic Detection of High Vocal Effort in Telephone Speech

Jouni Pohjalainen, Tuomo Raitio, Hannu Pulakka, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
firstname.lastname@aalto.fi

Abstract

A system is proposed for the automatic detection of high vocal effort in speech. The system is evaluated using both PCM-coded speech and AMR-coded telephone speech. In addition, the effect of far-end noise in the telephone conditions is studied using both matched-condition training and cases with additive noise mismatch. The proposed system is based on Bayesian classification of mel-frequency cepstral feature vectors. Concerning the MFCC feature extraction process, the substitution of a spectrum analysis method emphasizing the fine structure improves the results in the noisy cases.

Index Terms: vocal effort detection, speech analysis

1. Introduction

High vocal effort is used in speech production in order to increase the sound's distance of transmission or signal-to-background-noise ratio [1] [2] [3]. Possible triggering mechanisms for high vocal effort include the Lombard reflex (while speaking in a noisy environment) [2], changes in emotional expression of speech (for example from sadness or neutral to anger or excitement) and the need to communicate urgently or over a distance [3].

Raising the vocal effort causes certain systematic acoustical effects on speech. In particular, it increases both the fundamental frequency (F0) and the first formant (F1) [1] [2] [3] [4] of the produced speech signal. In addition, spectral energy of speech tends to shift to higher frequencies when vocal effort is raised. This is manifested, for example, as an increase in the spectral center of gravity or as a decrease in the spectral tilt [2] [4].

Frequency-domain effects caused by changing the vocal effort have implications for data-driven speech technology applications relying on short-time spectral features such as mel frequency cepstral coefficients (MFCCs). In particular, the performance of, e.g., automatic speech recognition (ASR) and speaker recognition systems will be affected by vocal effort mismatch between the training and recognition phase [5] [6] [7]. In order to avoid the performance degradation caused by this mismatch, a detection system is needed to aid the recognizer in choosing acoustic models that are most appropriate for the changed conditions [7].

Earlier studies have examined the identification of vocal effort in clean speech data with a high signal-to-noise ratio (SNR) by using MFCC features [5] [7] as well as simple spectral features such as the center of gravity [8]. There are, however, no previous studies on automatic detection of vocal effort from telephone speech (even though there are some earlier studies on the detection of emotions in telephone speech, e.g. [9]). A preliminary study on automatic detection of vocal effort from realistic telephone speech (i.e. speech that is bandlimited, encoded and corrupted with background noise) is thus called for. When operating in tandem with ASR, a vocal effort detection system could, besides improving the recognition performance, also alert a human call center attendant for increased vocal effort of the caller which could possibly indicate frustration. This application can be related to automatic audio-based surveillance in a noisy environment, in which shouts and screams are important target classes, e.g. [10] [11].

In the present study, robust detection of vocal effort from continuous speech is addressed by taking into account the effects of transmission channel (with focus on telephone speech), sampling rate and background noise. Different spectrum estimation methods are compared in order to take into account the spectral characteristics of normal and high-effort speech.

2. Detection system

2.1. Feature extraction

The input signal is first pre-emphasized with $H_p(z) = 1 - 0.97z^{-1}$ and then arranged into overlapping Hamming-windowed frames of 25 ms with a shift interval of 10 ms. An MFCC feature vector of 12 coefficients (excluding the zeroth one) is computed from each frame using the standard processing chain of 1) squared magnitude spectrum computation, 2) mel frequency filterbank, 3) logarithm and 4) discrete cosine transform [12]. The mel filterbank employed consists of 40 triangular filters spaced evenly on the mel scale. Inclusion of the delta coefficients has been investigated, but has not been found to improve detection performance.

The magnitude spectrum to be represented by the MFCC feature vector is typically obtained using discrete Fourier transform (DFT), implemented by fast Fourier transform (FFT) algorithms. However, DFT analysis is not particularly resistant to additive noise. Previous studies (e.g. [13], [14]) have shown improved performance in speaker verification and ASR in noisy environments when the FFT spectrum estimation was replaced in the MFCC computation chain by linear prediction (LP) and its noise-robust variants. Moreover, the LP-based spectrum estimation outperformed the conventional FFT-based feature extraction in detection of shouts in a generic noisy environment [11]. LP minimizes 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 with respect to the coefficients a_k, giving the infinite impulse response (IIR) filter $1/(1 - \sum_{k=1}^{p} a_k z^{-k})$ [15]. In order to model the envelope of the magnitude spectrum, the prediction order p is typically chosen to be the sampling frequency in kHz added by a small integer [16]. For example, $p = 20$ is a typical choice for a signal sampled at 16 kHz. However, in the present evaluation $p = 20$ will also be used for 8 kHz material giving a somewhat more detailed model. Despite this, LP models the spectral envelope, but not the fine structure, which is closely related to F0. Because vocal effort affects F0, F0 cues can be helpful in the detection of high vocal effort in various contexts. Thus, in

surveillance-oriented shout detection it has been found useful to multiply the LP envelope with a cepstrally separated fine structure, or excitation spectrum [11]. Specifically, the procedure can be described as follows:

1. Use LP analysis to obtain the magnitude spectrum envelope H_k.
2. Transform the signal into the cepstral domain [12] (using the processing chain 1) DFT magnitude spectrum 2) logarithm 3) inverse DFT), lifter this real cepstrum 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 transform the result back into a magnitude spectrum. When only the high-time part of the cepstrum is preserved, the resulting magnitude spectrum will mostly reflect the vocal tract excitation [16]. Denote the thusly processed excitation spectrum by G_k. Periodic excitation information up to 500 Hz (a frequency which the F0 of adults normally does not exceed in normal speech) is retained in the liftered excitation spectrum.
3. Compute the final squared magnitude spectrum by $S_k = (H_k G_k)^2$.

If the inverse filter of a LP model were to be applied to the spectrum given by step 3 above, the residual spectrum thereby obtained would be the cepstrally separated excitation spectrum. Because of this, the described spectrum analysis method is termed linear prediction with cepstral residual (LP-CR). Figure 1 illustrates the FFT, LP, excitation and LP-CR spectra for a vowel frame. It can be observed that while FFT is unable to show a clear formant structure due to background noise, LP-CR indicates emphasized formant peaks. A hypothetical reason for this behavior is that the LP analysis, which tends to place the formants at spectral energy maxima [16], will give a spectrum envelope model where the formants already correspond to prominent harmonics. Thus, those harmonics get further amplified in the final multiplication step of LP-CR.

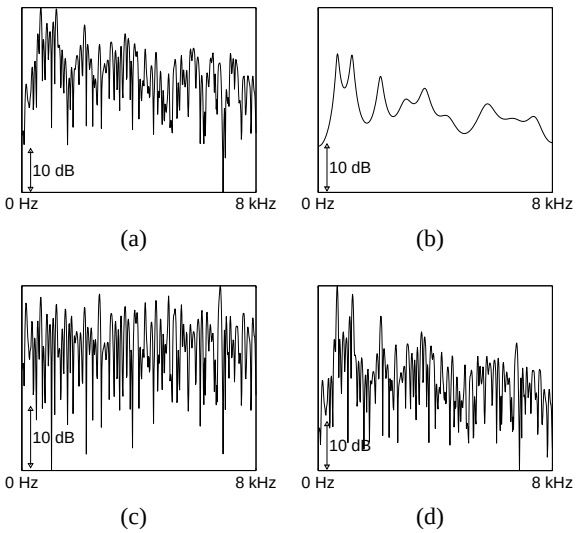

Figure 1: *Spectra computed from a noisy /a/ vowel frame spoken by a male speaker: a) magnitude spectrum given by FFT; b) magnitude spectrum envelope given by LP; c) excitation spectrum given by cepstral source-filter separation; d) LP-CR spectrum given by combining b) and c).*

In the present study, three spectrum analysis methods are compared: FFT, LP and LP-CR.

2.2. Frame selection

In both the training and classification phase, the feature vectors are analyzed in blocks of two seconds. Frame selection is used in order to focus the modeling and detection on the frames with the highest energy within the analysis block. By modeling and recognizing the locally most energetic frames within a short analysis window, the system focuses especially on the formant cues of high vocal effort [4]. The analysis block is shifted forward one second at a time. In the training phase, a frame is included in the training material if it is selected by the frame selection method in two successive, overlapping block positions.

Logarithmic energy is computed for each short-time frame, i.e. every 10 ms. For an analysis block of two seconds, this results in a sequence of 200 values, denoted by E_n. Frame selection classifies E_n into high and low values. In the present study, this is performed by an application of k-means clustering [12]. The mean values of two clusters are initialized with $\min(E_n)$ and $\max(E_n)$. After k-means iteration has converged, denote the obtained cluster assignment as $X_n = 1$, if E_n belongs to the cluster whose mean value was initialized with $\max(E_n)$, and $X_n = 0$ otherwise. The frames for which $X_n = 1$ will be selected for further processing.

2.3. Detection rule

The detection system models normal and high-effort speech with their own Gaussian mixture models (GMMs) for the purpose of binary classification according to the Bayes rule [17]. Each GMM has 8 components and a diagonal covariance structure [18]. They are trained using 10 iterations of expectation-maximization (EM) re-estimation for GMMs [18]. Before training, the component weights are initialized by uniform distributions, the variance parameters of each component by 0.1 times the global variances of the features, and the mean vectors of each component by the selection approach proposed by Katsavounidis et al. [19].

In the detection phase, after the high-energy frames inside a two-second analysis block (with a shift interval of one second) have been selected using the previously described unsupervised approach, the averaged log likelihoods of their corresponding feature vectors having been produced by each GMM are computed and denoted as L_{high} and L_{normal}. The logarithmic likelihood ratio decision statistic used in making the detection decision is

$$L = L_{\text{high}} - L_{\text{normal}} \quad (1)$$

3. Experimental evaluation

3.1. Original speech material

Speech data was collected from 11 male and 11 female speakers, all native speakers of Finnish. They read 24 sentences in Finnish, consisting of one to four words, first by using normal vocal effort and then by shouting. The speech signals were recorded with a condenser microphone in an anechoic chamber, where the speakers stood 0.7 m away from the microphone. The speakers were required to use a vocal effort increase high enough compared to their normal speech so that the voice could be accepted as shouting. To make sure that the speakers accomplished the task, operators monitored the recording in real time

from the outside and requested the speaker to repeat the shouting part if necessary. This was done using visual examination of the sound level, requiring it to reach at least 90 dB SPL level, and by listening with headphones.

The data was originally sampled at 96 kHz using a resolution of 24 bits and downsampled for the present evaluation to 16 kHz. Silences and pauses were removed by automatic voice activity detection. The total length of one speaker's normal speech material varied between 30 and 39 seconds, while the total length of one speaker's shouted speech material varied between 33 and 50 seconds. The speakers' SPL averaged over the most energetic 50% of 25-ms frames varied between 67 dB and 82 dB in normal speech and between 85 dB and 107 dB in shouted speech.

3.2. Preparation of the evaluation material

In order to examine the effect of both noise and channel on the detection performance, six different conditions, shown in Table 1, were examined.

PCM16 and PCM8 correspond to unprocessed pulse-code modulated (PCM) speech with sampling rate 16 kHz and 8 kHz, respectively.

For the four telephony conditions T1 to T4, 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.

Noise-corrupted speech signals sampled at 16 kHz were high-pass filtered with the mobile station input (MSIN) filter that approximates the input characteristics of a mobile terminal [20] and decimated to the sampling rate of 8 kHz. The speech level was normalized to 26 dB below overload point. Finally, the signals were processed with the adaptive multi-rate (AMR) codec [21], which is commonly used for speech coding in the GSM cellular system, at a bit rate of 12.2 kbps.

Table 1: *The different analysis conditions and their SNRs.*

	Additive noise condition (SNR)	Sampling rate	Simulated transmission channel
PCM16	none	16 kHz	none
PCM8	none	8 kHz	none
T1	car interior (30 dB)	8 kHz	telephone
T2	car interior (0 dB)	8 kHz	telephone
T3	factory (0 dB)	8 kHz	telephone
T4	speech babble (0 dB)	8 kHz	telephone

3.3. Evaluation methods

The experiments were performed as leave-one-out cross validation speakerwise, i.e., one speaker in turn was chosen as the test speaker and the other 21 speakers' material was used for training the models.

As a measure of performance of the detection task, the equal error rate (EER) was used. The EER is the value of both the miss rate and the false alarm rate using a decision threshold for the statistic given by Eq. 1 that makes these error rates equal to each other. In addition, detection-error-tradeoff (DET) curves were examined.

Statistical analysis between different analysis methods was performed using a statistical significance test appropriate for detection systems [22]. 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

Table 2 shows the results for matched-condition training, i.e., for each test condition (Table 1) the detection system was trained using material belonging to the same condition. The differences were statistically significant at the 95 % level between FFT and LP for the conditions PCM16, PCM8, T1, T3 and T4; between FFT and LP-CR for the conditions PCM8, T1 and T4; between LP and LP-CR for the conditions PCM16, PCM8, T1, T2, T3 and T4. Narrowing the bandwidth appears to degrade the results. Interestingly, however, the interaction of the telephone channel and background noise does not necessarily have a negative effect in comparison to clean PCM-coded data.

Table 2: *EER scores (%) for PCM and telephone speech with matched-condition training and different spectrum estimation approaches used in MFCC feature extraction.*

Spectrum estimation method	Test condition (Table 1)					
	PCM16	PCM8	T1	T2	T3	T4
FFT	3.3	5.1	4.0	3.2	4.4	5.5
LP	2.0	3.5	2.9	3.0	5.1	4.7
LP-CR	3.2	4.2	3.5	3.5	4.0	4.2

Table 3 shows the results for additive noise mismatch, i.e. varying background noise at the caller's location, while the detection system has been trained using material with the high-SNR car interior condition T1. The pairwise differences among the methods were all statistically significant except between LP and LP-CR in condition T2. The results show a performance advantage for LP-CR over FFT and LP.

Table 3: *EER scores (%) with the detector trained using T1 material and evaluated in mismatched telephone conditions.*

Spectrum estimation method	Test condition (Table 1)		
	T2	T3	T4
FFT	3.9	4.8	5.1
LP	3.2	5.9	5.9
LP-CR	3.5	4.0	4.4

Figures 2 and 3 show the DET curves corresponding to the three spectrum analysis methods in the case of factory noise corruption (T3) and matched-condition and mismatched training, respectively. It can be noticed that LP-CR yields the best detection performance over a wide range of operating points.

4. Conclusions

A system for the detection of high vocal effort was described and evaluated in various matched conditions regarding transmission channel, additive noise corruption and speech bandwidth. In addition, mismatched background noise conditions

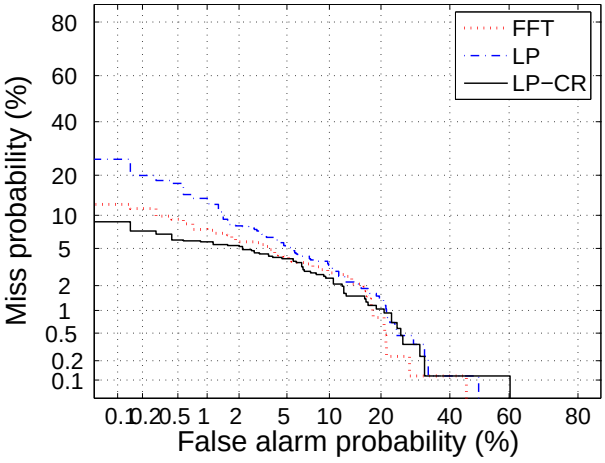

Figure 2: *Detection error tradeoff (DET) curves corresponding to different spectrum estimation methods for MFCC with training and evaluation in condition T3.*

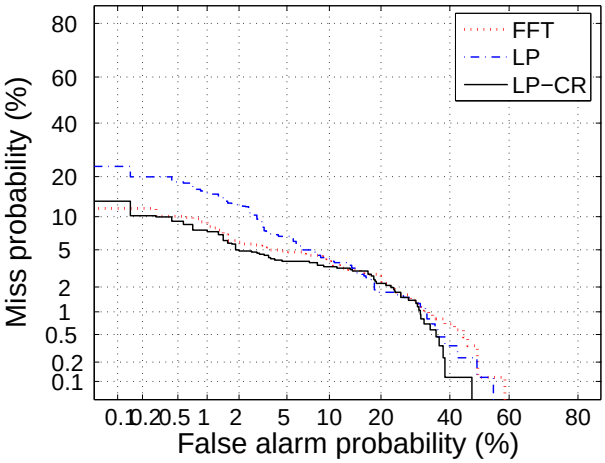

Figure 3: *Detection error tradeoff (DET) curves corresponding to different spectrum estimation methods for MFCC with training in condition T1 and evaluation in condition T3.*

occurring in combination with the telephone transmission channel were studied. In each case, the baseline system using conventional (FFT-based) MFCC features, unsupervised energy-based frame selection and Bayesian classification using GMMs provided reasonable performance. In most cases, FFT was outperformed by LP-CR as the spectrum estimation method in MFCC analysis. This behavior can be explained by the fact that LP-CR emphasizes the role of the spectral fine structure (i.e., F0 and its harmonics) which is known to be an important acoustic cue for speech with high vocal effort. Future research directions include the application of the proposed system as an aid to ASR systems in conditions where high vocal effort can be encountered. The performance of the LP-CR/MFCC features, which have shown good performance in paralinguistic tasks related to vocal effort, as a generic feature representation for ASR and other speech applications is another question of interest.

5. Acknowledgements

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

6. References

[1] Rostolland, D., “Phonetic structure of shouted voice”, Acustica, 51:80–89, 1982.

[2] Junqua, J.-C., “The Lombard Reflex and Its Role on Human Listeners and Automatic Speech Recognizers”, J. Acoust. Soc. Am., 93(1):510–524, 1993.

[3] Traunmüller, H. and Eriksson, A., “Acoustic effects of variation in vocal effort by men, women, and children”, J. Acoust. Soc. Am. 107(6):3438–3451, 2000.

[4] Liénard, J.-S. and Di Benedetto, M.-G., “Effect of Vocal Effort on Spectral Properties of Vowels”, J. Acoust. Soc. Am. 106(1):411–422, 1999.

[5] Zhang, C. and Hansen, J. H. L., “Analysis and Classification of Speech Mode: Whispered Through Shouted”, in Proc. Interspeech, Antwerp, Belgium, August 2007.

[6] Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S. S., Jameel, H., Richey, C. and Goodman, F., “Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification”, in Proc. Interspeech, Brisbane, Australia, September 2008.

[7] Zelinka, P. and Sigmund, M., “Automatic Vocal Effort Detection for Reliable Speech Recognition”, in Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010), Kittilä, Finland, September 2010.

[8] Harwardt, C., “Comparing the Impact of Raised Vocal Effort on Various Spectral Parameters”, in Proc. Interspeech, Florence, Italy, August 2011.

[9] Erden, M. and Arslan, L. M., “Automatic Detection of Anger in Human-Human Call Center Dialogs”, in Proc. Interspeech, Florence, Italy, August 2011.

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

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

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

[13] Saeidi, R., Pohjalainen, J., Kinnunen, T. and Alku, P., “Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification”, IEEE Signal Processing Letters, 17(6), 2010.

[14] Keronen, S., Pohjalainen, J., Alku, P. and Kurimo, M., “Noise robust feature extraction based on extended weighted linear prediction in LVCSR”, in Proc. Interspeech, Florence, Italy, August 2011.

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

[16] Rabiner, L. R. and Schafer, R. W., “Digital Processing of Speech Signals”, Prentice-Hall, 1978.

[17] Theodoridis, S. and Koutroumbas, K., “Pattern Recognition”, 2nd ed., Academic Press, 2003.

[18] 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.

[19] 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.

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

[21] 3GPP TS 26.090, Adaptive multi-rate (AMR) speech codec, transcoding functions, 3rd Generation Partnership Project, Sept. 2011, version 10.1.0.

[22] Bengio, S. and Mariéthoz, J., “A Statistical Significance Test for Person Authentication”, in Proc. ODYSSEY04, Toledo, Spain, June 2004.