Detection of shouted speech in noise: Human and machine

Jouni Pohjalainen,[a] Tuomo Raitio, Santeri Yrttiaho, and Paavo Alku
Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13000, FI-00076 AALTO, Espoo, Finland

(Received 17 April 2012; revised 24 October 2012; accepted 16 February 2013)

High vocal effort has characteristic acoustic effects on speech. This study focuses on the utilization of this information by human listeners and a machine-based detection system in the task of detecting shouted speech in the presence of noise. Both female and male speakers read Finnish sentences using normal and shouted voice in controlled conditions, with the sound pressure level recorded. The speech material was artificially corrupted by noise and supplemented with pure noise. The human performance level was statistically evaluated by a listening test, where the subjects labeled noisy samples according to whether shouting was heard or not. A Bayesian detection system was constructed and statistically evaluated. Its performance was compared against that of human listeners, substituting different spectrum analysis methods in the feature extraction stage. Using features capable of taking into account the spectral fine structure (i.e., the fundamental frequency and its harmonics), the machine reached the detection level of humans even in the noisiest conditions. In the listening test, male listeners detected shouted speech significantly better than female listeners, especially with speakers making a smaller vocal effort increase for shouting.
 [http://dx.doi.org/10.1121/1.4794394]

PACS number(s): 43.71.Bp, 43.72.Dv [MAH] Pages: 2377–2389

I. INTRODUCTION

Shouting is used by speakers to produce a very loud acoustical signal in order to increase the sound's distance of transmission or its signal-to-noise ratio (SNR).[1,2] In a noisy environment filled with non-vocal sounds and normal speech, shouting is typically used to communicate something urgently. In addition, the use of high vocal effort in such an environment can be indicative of an alarming situation. Therefore, machine-based detection of shouted speech in adverse ambient noise conditions is a relevant research topic in audio-based surveillance.[3,4] Also, detection of high vocal effort can be applied in speech and speaker recognition in order to tackle a possible mismatch between training and testing conditions.[5,6] For all these technological applications, the performance of human listeners in shout detection serves as a natural point of comparison.

Detection of shouting by humans and machine in adverse noise conditions is compared in the present study. Since this topic calls for background knowledge from different areas of speech science and engineering, the introduction is divided into four subsections discussing separately (A) the spectral characteristics of shouting, (B) human perception of shouted speech, (C) its machine detection, and, finally, (D) the aims of the study.

A. Spectral characteristics of shouted speech

Several previous studies have observed that shouting cannot be regarded as normal speech produced with a very loud volume. Instead, many acoustical properties of the voice are altered when the vocal effort is increased from normal to shouting. In addition to the obvious effect of an increased sound pressure level (SPL) in shouts, also segmental durations and spectral features of speech differ between normal and shouted speech.[7] From the point of view of machine-based shout detection, the spectral characteristics are most important because, first, they can be easily implemented using frame-based feature vectors in a manner similar to that used in speech recognition[8] and speaker recognition.[6] Second, relying on spectral characteristics enables building shout detection systems that are scale invariant, i.e., the detection system does not utilize the SPL information of speech and is therefore independent of, for example, the microphone-to-speaker distance. Therefore, the acoustical properties of shouted speech are treated in the following from the point of view of their spectral characteristics only.

Rostolland[7] reported a large difference in the fundamental frequency (F0) between shouted and normal speech for both male and female speakers. The increase in F0 was especially noticeable for talkers of low pitch. Moreover, F0 differences among speakers in shouting were small compared to normal speech. In a subsequent follow-up study, largely increased values for the frequency of the first formant (F1) were reported for shouted French vowels[1] in comparison to those produced with normal effort. Liénard and Di Benedetto[9] analyzed vowel spectra within the vocal effort range typically used in everyday conversations. They found statistically significant increases in F0 and F1, but not in the second (F2) or third (F3) formant, in experiments where the distance between the speakers of the conversation was varied. Liénard and Di Benedetto[9] also analyzed the formant amplitudes, which showed a systematic increase for higher formants in shouted speech, reflecting a decrease of spectral tilt. Traunmüller and Eriksson[2] reported increased values in

[a]Author to whom correspondence should be addressed. Electronic mail: jouni.pohjalainen@aalto.fi

the frequency of both F0 and F1 when vocal effort was raised over a wide range from whispering to shouting. Spectral characteristics of Lombard speech were studied by Junqua.[10] His results also indicated that speech with high vocal effort is characterized by an increased F0 (more pronounced for male speakers), F1 (more pronounced for female speakers), and spectral center of gravity.

Schulman[11] found amplified articulatory movement patterns in loud speech relative to normal speech, in particular a generally lower jaw position. He explained these findings perceptually by relating them to the importance of maintaining the Bark distance between F1 and F0: since F0 increases in loud speech and shouting, the frequency of F1 must also shift up in order to maintain the correct phonetic identities. The shape of the glottal pulse is also heavily influenced by the vocal effort. Notably, the relative length of the glottal closing phase, the so-called closing quotient, decreases when speakers raise their vocal intensity.[12–14] In the frequency domain, this increased sharpening of the glottal pulse in the time domain results in the emphasis of the level of the higher frequencies, i.e., in a lower tilt of the speech spectrum. Ternström, Bohman, and Södersten[15] found a saturation point for the spectral tilt after which the 2–6 kHz band energy did not rise any more relative to the 0.1–1 kHz band energy. Simple spectral parameters such as the spectral center of gravity and the spectral tilt have been found to be effective features for the automatic discrimination between normal and loud speech of male speakers.[16]

Figure 1 illustrates the differences between the averaged spectra of normal and shouted speech from male and female speakers.

B. Human perception

Human perception of normal and shouted speech in noise was studied by Pickett.[17] His results indicate that the intelligibility of speech heard in white noise, with a constant SNR close to 0 dB, decreases rapidly when the vocal effort is raised towards shouting. At lower levels of the vocal effort, increasing vocal intensity causes a smaller degradation in the intelligibility.

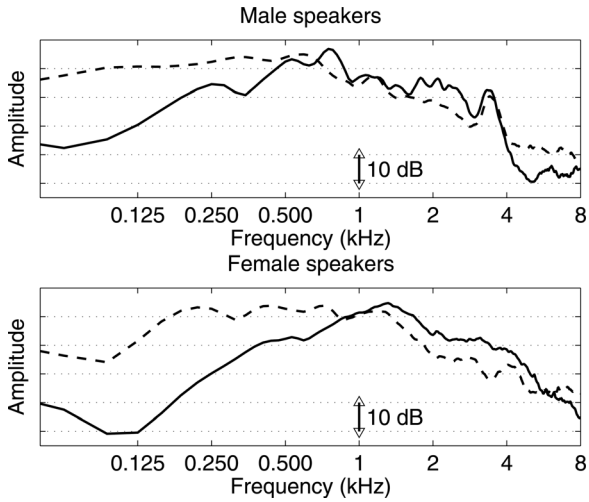

FIG. 1. Averaged spectra for normal (dashed line) and shouted (solid line) speech of 11 male (top) and 11 female (bottom) speakers.

Brandt, Ruder, and Shipp found human listeners to be capable of perceiving raised vocal effort separately from loudness[18] and suggested increased spectral bandwidth to be an important acoustic cue for the perception of raised vocal effort. Several other previous studies on loud or shouted speech indicate that listeners acutely perceive raised vocal effort and easily associate it with other features under study, e.g., loudness[19,20] and distance to speaker.[2,21] Interestingly, Allen[20] found that the judgment of loudness depends on both SPL cues and vocal effort cues in different proportions with different listeners. In the light of the investigation conducted by Glave and Rietveld,[22] the large effect of vocal effort on loudness appears not to be due to any speech-specific high-level perceptual processing, but can instead be explained by the short-term acoustic (spectral) characteristics discussed in Sec. I A. In particular, the characteristics of the glottal source have been found to greatly affect the loudness perception.[23]

Concerning the factors affecting the direct perception of the vocal effort, both the glottal source characteristics and F1 have been found to be important.[24] There are, however, no studies that have specifically addressed how accurately humans *detect* shouted or high-effort speech amidst competing background noise or how accurately they can discriminate between a normal and a deliberately high vocal effort of natural speech.

C. Machine detection

Automatic, machine-based detection of shouted speech in a noisy environment is a challenging research question in audio-based surveillance technology. The goal of this technology is to automatically detect sound events associated with potentially alarming situations in a specific acoustic environment. Systems have been developed for detection of, e.g., shouted speech in trains;[25] non-neutral speech and banging in elevators;[26] and screams, gunshots, and explosions in different urban and military environments.[3,4] An overview of audio event detection systems, their problem domains, and the techniques employed is provided by Ntalampiras *et al.*[3]

Machine-based audio detection systems typically consist of two major parts: the front-end and the back-end. The former transforms the input audio signal into a sequence of feature vectors, a process that is often performed by expressing the short-time magnitude spectrum of the input as mel frequency cepstral coefficients (MFCCs), while other solutions use other forms of cepstral coefficients or specialized features.[3] The back-end predominantly models the probability distributions of the MFCC vectors using Gaussian mixture models (GMMs) in a Bayesian classification framework.[3,25,26] Support vector machines are one competing approach to GMM-based classification.[25]

In machine detection studies, scream detection performance has been found to degrade steeply when the SNR is close to 0 dB.[3,4] In a realistic scenario, the environmental noise conditions are subject to change. In any given stationary conditions, the SNR is related to the distance between

the person shouting and the microphone, given that the SPL in free field conditions is inversely proportional to the square of the distance from the sound source.[27] Thus, to increase the usability and reliability of automatic detection, attention needs to be paid to the noise robustness of the method. This issue has been studied, e.g., by Pohjalainen *et al*.[28,29] leading to the development of a general-purpose, noise robust detection system for shouted speech. This system, which is further developed in the present paper, is based on MFCC feature extraction and GMM classification.

In addition to surveillance-oriented applications, increasing use of automatic speech recognition and speaker recognition systems in adverse environments may benefit from automatic detection of speech of high vocal effort.[5,6,30] Speaking in a noisy environment induces the Lombard effect, hence making the talker change his or her speaking style from normal to loud or very loud.[10] Changing the speaking style causes, in turn, a mismatch between the acoustical properties of the current speech signal and those represented by the previously trained statistical models, deteriorating the system performance. If, however, the system was provided with automatic detection of high-effort speech, the recognizer could switch between acoustical models trained with speech of different vocal effort levels and, consequently, the recognition performance would improve.[5] Regarding the effect of vocal effort variation on the performance of recognition applications, work with similar objectives has recently been conducted also with whispered speech.[30,31]

D. Aims of the study

In this study, an automatic machine-based shout detection system for acoustic environment monitoring is proposed, based on a feature representation that modifies the widely used MFCC vector by taking into account the most obvious acoustical consequence in the production of high-effort speech, the raising of the F0. The goal is to validate the machine-based system in several realistic noise conditions and to compare its performance to that obtained by human listeners. In addition, the study aims to find out how human detection of shouting in competing talking crowd noise differs between male and female speakers, quiet and loud shouters, and male and female listeners. Involving the Lombard effect, whose nature to a degree depends on the type of the noise,[32] is beyond the scope of this study. This choice was made deliberately in order to focus on high vocal effort alone and not on the effect that the background noise has on the production of speech. Specifically, the conditions simulated in this study are such that the ambient noise level at the talker's location is low or moderate and hence no Lombard effect is induced. However, the (fixed) position of the microphone may be at a long distance from the talker, giving rise to a low SNR.

II. MATERIAL

Speech data were collected from 11 males and 11 females. The subjects, all native speakers of Finnish, read 24 sentences in Finnish using both normal vocal effort and shouting. The speech signals were recorded with a condenser microphone (AKG CK92 omnidirectional capsule with SE300B power supply) in an anechoic chamber. The data were sampled at 96 kHz using a resolution of 24 bits. At the computer, the signals were downsampled to 16 kHz. Before each recording session, a calibration signal (1 kHz sine tone with SPL = 92.3 dB) was recorded. The calibration signal was later used to determine the SPL values of the recorded speech signals.

The speakers first produced the sentences using their normal vocal effort, after which the same sentences were repeated by shouting. Twelve of the selected sentences are in the imperative mood, consisting of one to four words. The semantic contents of these sentences were designed to represent vocal messages that people might use in potentially threatening situations such as "anna se kamera tänne" ("give me the camera"), "älkää liikkuko" ("don't move"), and "lopettakaa" ("stop it"). The other 12 sentences, each consisting of three words, are in the indicative mood and have a neutral, abstract information content. Because exactly the same textual material is used for normal and shouted speech, the shout detection cannot benefit from phonemic differences between the two speech classes. All the sentences are listed in Table I.

The speakers were instructed to use a very large vocal effort when shouting. A mere raised volume was not accepted as shouting. After giving the instructions and checking the position of the speaker relative to the microphone, the operators left the anechoic chamber, leaving the speaker alone in the chamber. The speaker stood at the distance of 0.7 m from the microphone. One operator monitored the recording from outside of the chamber, listening to the recorded samples using headphones and following the signal waveform in real time on the computer screen. The waveform was used to gauge the instantaneous SPL. If the waveform envelope level in shouting did not reach the level of the calibration tone (92.3 dB), or did not show enough amplification compared to the same talker's normal speech (according to informal visual judgment corresponding to a level difference of at least 10 dB), the talker was asked to repeat the shouting section. In addition, shouted speech was perceptually assessed by the operator. If he assessed a sample not to represent shouting, the talker was asked to repeat the shouting section until it was acceptable.

TABLE I. List of the Finnish sentences used in collecting the speech material.

Ottakaa tuo varas kiinni	Saara sukii laamaa
Anna se kamera tänne	Liinu tilaa viinaa
Et mene vielä minnekään	Paavi tavaa suuraa
Anna se takaisin	Taata tivaa taalaa
Tule pois sieltä	Siiri kuvaa jaalaa
Ei yhtään lähemmäs	Saana sahaa haapaa
Pysy siinä	Tuuli puhuu kiinaa
Älkää liikkuko	Piika vahaa tuubaa
Ole hiljaa	Taavi tekee siikaa
Lopettakaa	Tuula tukee Kuubaa
Juoskaa	Ruusu varoo laavaa
Ampukaa	Haamu lukee saagaa

The utterances were separated and concatenated by automatic voice activity detection, which is similar to the frame selection method to be described in Sec. III C. For the purpose of machine detection, the material was stored in 44 files with two files per speaker such that each speaker's normal and shouted speech material resided in separate files. The length of the normal speech files varied between 30 and 39 s, while the length of the shouted speech files varied between 33 and 50 s. For the listening test evaluation, the individual utterances were kept separate.

The averaged SPL levels of speech produced with normal vocal effort and shouting were computed separately for each talker. The overall SPL was determined using frame-based energy calculation together with the recorded calibration tone with a known SPL level at the recording location. Frames of 25 ms, taken every 10 ms, were used in this computation. The obtained SPL values were averaged for the most energetic 50% of the frames. This was done in order to decrease the dependency of the results on the source text and language, as the material is continuous speech instead of, e.g., sustained vowels. Including all the material in the computation of SPL would result in a more text-dependent value which would be influenced by, e.g., the proportion of voiced and unvoiced speech. The most energetic half of the signal typically consists of vowels and is thus less text-dependent.

Averaged SPL values for normal speech and shouted speech, as well as their differences in decibels, are listed for all speakers in Table II. From this table, the following observations on the recorded speech material can be made. The averaged shouting SPLs display rather large variation from one speaker to another. They vary over a 17 dB range both in the male and female speaker groups. The difference in decibels between a speaker's shouted speech and normal speech ranges from 15 to 33 dB for the male speakers and from 17 to 28 dB for the female speakers. Such SPL differences are in line with previous studies: for instance, Rostolland[7] reports C-weighted level differences between shouted and normal speech of 28 and 20 dB for male and female speakers, respectively.

In order to simulate shouting in noise, the speech material was artificially corrupted by two noise types from the NOISEX-92 database named *babble* and *factory1*.[33] The former comprises speech from multiple simultaneous talkers. The latter is mechanical noise recorded in a factory, including frequent transient impulsive sounds. The noise corruption was conducted to achieve the following SNR categories: 20, 10, 0, −10, and −20 dB.

TABLE II. Speaker-specific averaged SPL in decibels for normal and shouted speech and their difference. Each of the SPL values has been obtained by first integrating the signal energy in frames of 25 ms with a 10 ms sampling interval, relating the result to the reference signal to obtain the frame SPL value, and then averaging the frame SPL values over the most energetic 50% of the frames for the specific speaker and speaking condition.

		Speaker number										
		1	2	3	4	5	6	7	8	9	10	11
Male	Speech	73	69	78	71	82	71	74	74	71	74	76
	Shouting	106	99	107	96	106	93	94	93	90	91	90
	Difference	33	30	28	25	24	22	20	18	18	16	15
Female	Speech	72	76	70	67	77	76	71	73	78	70	67
	Shouting	100	102	96	90	100	98	92	93	97	88	85
	Difference	28	26	26	24	23	23	21	20	19	18	17

III. AUTOMATIC SHOUT DETECTION METHOD

A. Overview

An automatic detection system is proposed, based on recognizing the spectral distribution of the most energetic parts of an evaluated audio signal segment. The system consists of three processing stages, which are explained in detail in the following sections: (1) feature extraction, (2) frame selection, and (3) pattern classification. In particular, this study focuses on the short-time spectrum analysis part of the feature extraction stage. The role of spectral features in capturing information related to the vocal tract excitation is investigated in detail.

The feature extraction module converts a digital audio signal into a sequence of feature vectors, each representing the acoustic features of a short signal frame. The approach chosen is to model each short-time magnitude spectrum as mel frequency cepstral coefficients (MFCCs),[8] the computation of which is illustrated in Fig. 2. The squared magnitude spectrum can be obtained in different ways, as shown in Fig. 3. FFT gives a non-parametric spectrum estimate using the discrete Fourier transform, while the other branches in Fig. 3 employ parametric spectrum envelope modeling, as will be described in Sec. III B.

In the pattern classification module, the probability distributions of the MFCC vectors are modeled using Gaussian mixture models (GMMs) in the context of Bayesian classification.[34] The input segment is classified as either shouted speech or non-shouted. In general, the MFCC/GMM classification approach is popular in diverse speech and audio recognition applications, such as speaker recognition,[34] audio event detection,[3] and paralinguistic analysis of speech, e.g., the recognition of emotional state[35] or vocal effort class.[30]

Between the feature extraction and pattern classification stages, unsupervised energy-based frame selection is applied in order to focus the GMM modeling and recognition only on the most energetic frames, which presumably have the largest SNR (assuming a speech target signal is present). The classification rule is based on three separately trained GMMs: one for shouted speech, one for normal speech, and one for the expected type of ambient noise. A detection decision is made every second using an analysis block of two

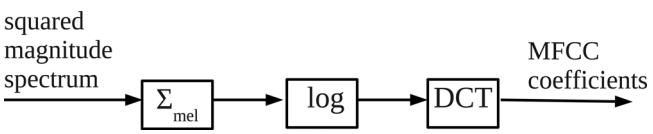

FIG. 2. Stages of obtaining MFCCs from the squared magnitude spectrum. The chain consists of three parts: computation of frequency band energies using filters with triangular passbands spaced evenly according to the mel scale, taking a logarithm of the band energies, and discrete cosine transform of the logarithmic energies.

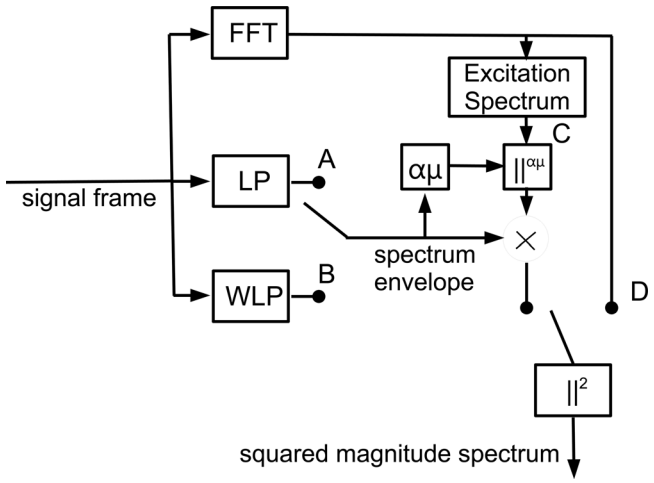

FIG. 3. Alternative paths for computing the squared magnitude spectrum, which is used as an input to the MFCC chain shown in Fig. 2.

seconds. The analysis block length has been chosen based on the considerations that it is long enough to typically cover energetic voiced segments in continuous speech, yet short enough so as not to distract the time resolution.

B. Acoustic feature extraction

The input to the system is sampled at 16 kHz and pre-emphasized by a first-order highpass filter $H_p(z) = 1 - 0.97z^{-1}$. The signal is processed in Hamming-windowed analysis frames of 25 ms with a 10-ms frame shift. For each frame, an MFCC feature vector is computed, as illustrated in Fig. 2, using the standard processing chain of squared magnitude spectrum computation, a filterbank of triangular filters spaced evenly on the mel frequency scale, logarithm, and discrete cosine transformation.[8] The MFCC vector is a representation of the short-time magnitude spectrum that also takes into account the nonuniform frequency resolution of human hearing. The MFCC analysis can thus be considered to coarsely mimic the processing that occurs on the basilar membrane in the inner ear.[36]

The magnitude spectrum which is represented by the MFCC features is typically obtained using discrete Fourier transform (DFT), implemented by a fast Fourier transform (FFT) algorithm (path D in Fig. 3). However, DFT analysis is not particularly resistant to additive noise. In earlier work dealing with noise robustness for automatic speech recognition (ASR) and speaker recognition, improvement in noise robustness has been achieved by replacing the FFT in the MFCC computation chain by linear predictive spectrum analysis,[37,38] such as conventional linear prediction (LP)[39] (path A in Fig. 3) and weighted linear prediction (WLP)[40] (path B in Fig. 3). 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 samples s_n with respect to the coefficients a_k, giving the infinite impulse response (IIR) filter $1/(1 - \sum_{k=1}^{p} a_k z^{-k})$. For the filter to depict the magnitude spectrum envelope (i.e., the formants), the prediction order p is typically chosen as slightly more than the sampling frequency in kHz,[41] for example, $p = 20$ would be a typical choice for a signal sampled at 16 kHz. For WLP, the corresponding error energy to be minimized is $\sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 W_n$, where the weighting function is chosen as the short-time energy $W_n = \sum_{i=1}^{p} s_{n-i}^2$. This weighting emphasizes the accurate modeling of the high-energy portions of the analysis frame that can be assumed to have a good SNR.

Other perceptually motivated feature representations, such as cepstral coefficients based on perceptual linear prediction,[42] perceptual MVDR (minimum variance distortionless response),[43] or perceptual MVDR-based cepstral coefficients[43] have been used in ASR in recent years, and they have shown improved recognition performance in noisy conditions. There are, however, no previous studies indicating that these methods can improve the detection of high vocal effort. Therefore, in order not to expand the experimental sections of this study too much, the perceptually motivated feature representations mentioned above were not involved in the current study. Instead, noise-robust feature extraction was addressed by utilizing only the two most widely used MFCC representations (i.e., FFT- and LP-based MFCCs) as references.

The change from normal to shouted speech has a distinct effect on the vocal tract excitation.[2,7,13] This effect manifests itself in the spectral fine structure of the produced acoustic speech pressure waves. In particular, the increased fluctuation speed of the vocal folds in the production of loud speech results in a more sparse spectral fine structure, characterized by an increased value of F0 and its harmonics. Therefore, an automatic system for the detection of shouted speech would most likely benefit from a feature representation capable of taking into account the change that occurs in the spectral fine structure of speech when vocal effort is altered from normal to shouting. Although using the linear predictive spectrum envelope in place of FFT in the MFCC computation chain may provide additional noise robustness, it does not preserve the spectral fine structure normally present in the FFT-based MFCC representation. To combine the benefits of both the conventional linear predictive analysis and the role of the spectral fine structure, an approach was adopted in which the linear predictive spectrum envelope is multiplied by the spectral fine structure obtained by cepstral analysis[28,29] (path C in Fig. 3). The present work uses a further modification of this approach, based on the observation that the cepstrally separated fine structure appears to be more resistant to heavy noise corruption than the linear predictive formants. The procedure, described in the flow diagram shown in Fig. 3, consists of the following steps:

(1) Use linear predictive analysis (either LP or WLP) to obtain the magnitude spectrum envelope H_k.
(2) Transform the signal into the cepstral domain[8] 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.[41] Denote the thus processed excitation spectrum by G_k. Periodic

excitation information up to 500 Hz is retained in the liftered excitation spectrum.

(3) Compute the spectral flatness[44] of the linear predictive spectrum envelope H_k as the ratio of the geometric and arithmetic mean of the spectrum,

$$\mu = \frac{\exp\left((1/N_q)\sum_{k=1}^{N_q}\log(H_k+\epsilon)\right)}{\epsilon + (1/N_q)\sum_{k=1}^{N_q} H_k},$$

where N_q denotes the DFT index corresponding the the Nyquist frequency and ϵ is a small constant added for numerical stability.

The spectral flatness measure μ assumes values between 0 and 1, with low values for highly shaped spectra and high values for flat spectra. The noisier the signal becomes, the more the speech formants are suppressed and the flatter the envelope spectrum will be.

(4) Compute the final squared magnitude spectrum $S_k = (H_k G_k^{\alpha\mu})^2$, where α is a parameter determining how much weight to assign to the spectral flatness weighted excitation spectrum. In this work, the experimentally determined value $\alpha = 3$ is used. The fine structure is thus emphasized more when the signal becomes noisier in order to rely on voiced speech harmonics instead of formants in the noisiest cases.

If the spectrum envelope is modeled by an LP all-pole filter and its inverse filter is applied to the spectrum model given by step 4 above, the residual spectrum will be the cepstrally separated excitation spectrum (with weighting). For this reason, this spectrum analysis method is referred to in this work as cepstral residual linear prediction (CRLP). Similarly, when WLP is used to model the spectrum envelope, the method is termed CRWLP. Figure 4 shows examples computed from voiced speech illustrating these spectrum analysis approaches and their relation to the conventional methods.

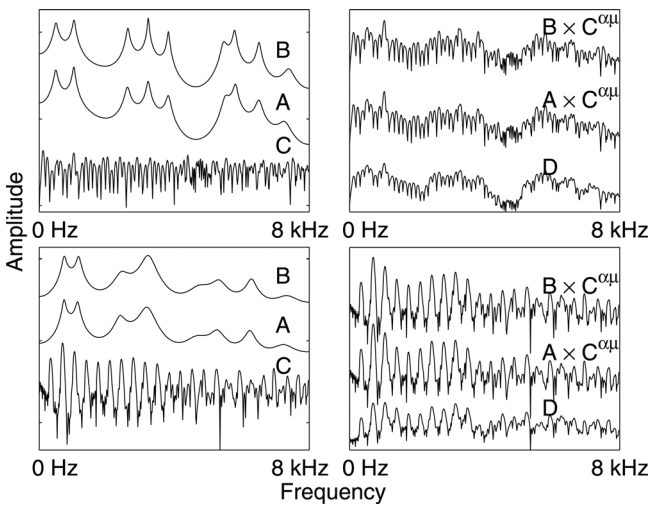

FIG. 4. Vowel [o] spoken normally (top) and with high vocal effort (bottom) by a male speaker. LP and WLP spectrum envelopes and the cepstrally liftered excitation spectrum (left) are used to construct alternative spectrum estimates (right) besides the FFT spectrum. The notation next to the curves corresponds to Fig. 3.

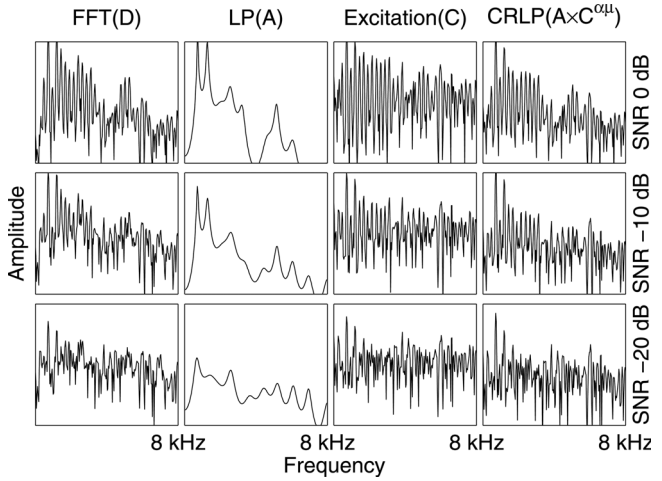

FIG. 5. Example spectra based on a shouted vowel frame by a male speaker. The rows correspond, from top to bottom, to SNR levels 0, −10, and −20 dB with factory noise corruption. The columns correspond to different types of spectra. The notation in parentheses corresponds to Fig. 3.

Figure 5 shows examples of spectra computed by the different feature extraction methods of Fig. 3 in moderate to heavily noisy conditions. The CRLP method is observed to preserve the spectral fine structure better than the FFT-based method when the amount of noise increases. In addition, Fig. 5 demonstrates that the formant cues weaken and the spectral tilt of speech decreases as the noise corruption increases, and this phenomenon is particularly apparent in the LP spectra. It is therefore hypothesized that even as the SNR decreases to such levels that the spectral envelope cues such as formants and spectral tilt vanish due to noise, a detection system using CRLP or CRWLP can still rely on cues present in the spectral fine structure in order to achieve better noise resistance. Thus, the rate at which the performance approaches chance level would be slowed down.

In applications such as ASR and speaker recognition, the feature representation is most often based on 12 MFCC coefficients, starting from index 1 and excluding the "zeroth" coefficient. These are possibly supplemented with the logarithmic energy of the analysis frame to give a 13-element vector. These coefficients are an auditory representation of the short-time magnitude spectrum envelope. They are usually concatenated with their first and second order "delta" coefficients[8] to depict the instantaneous time trajectory of each coefficient.

While the MFCC representation does not fully preserve the spectral fine structure, which is partially smoothed out by the mel filterbank, contributions due to the harmonics of F0 are still preserved in the higher-order MFCCs. Figure 6 shows the means and standard deviations of MFCCs for normal and shouted speech of the speaker population of this study. The lowest panel shows the difference of the mean vectors. There are noticeable differences in the distributions of the MFCCs at least until MFCC index 20. These considerations motivated the use of a longer-than-normal MFCC feature vector. Detection of shouted speech was evaluated by varying the length of the FFT-based MFCC feature vector. These experiments, conducted with MFCC lengths of 12, 18, 24, 30, and 36, were in accordance with previous studies,[28,29] indicating that the best performance is obtained by

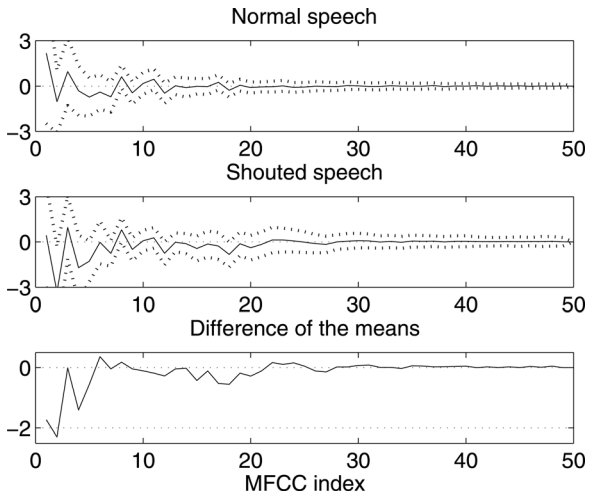

FIG. 6. Mean values (solid line) and standard deviation intervals (dotted line) of MFCCs averaged over normal and shouted speech from 11 male and 11 female speakers.

30 MFCCs, as shown in Table III (30 MFCCs being marginally better overall than 24 or 36 MFCCs). Table IV shows the results for MFCC lengths 12 and 30 with delta and double-delta coefficients. While the inclusion of delta coefficients boosted the noise robustness in comparison to 12 base MFCCs, the best overall detection performance was obtained with 30 MFCCs and no deltas, which was thus chosen as the form of the feature vector in subsequent tests.

C. Frame selection

Both in the training and detection phase of the present system, 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 SNR values within the analysis block. If the noise is assumed to be relatively stationary, frame energy is a good indicator of the SNR. Therefore, the modeling concentrates on the high-energy frames within each block. This is done in both the training phase and the detection phase. The analysis block is shifted forward one second at a time; in the training phase, the overlap between the frame selection decisions of two successive block positions is handled by averaging.

TABLE III. Equal error rates (%) for different numbers of MFCCs.

		Signal-to-noise ratio (dB)					
Type of noise	Number of MFCCs	20	10	0	−10	−20	−30
Factory	12	2.9	3.2	4.2	13.6	27.7	46.5
	18	2.8	2.7	3.5	12.7	20.7	44.7
	24	2.5	2.3	2.8	10.3	20.2	41.7
	30	2.7	2.4	2.9	10.1	17.8	45.7
	36	3.0	3.1	2.5	10.2	20.0	42.2
Babble	12	2.8	3.2	3.9	9.3	21.8	47.8
	18	3.2	2.9	4.1	8.5	19.8	48.1
	24	2.3	2.0	1.5	5.3	19.2	45.1
	30	2.7	2.2	2.4	5.6	16.7	43.0
	36	3.3	2.9	2.2	5.0	18.4	45.3

TABLE IV. Equal error rates (%) for 12 and 30 MFCCs concatenated with Δ and ΔΔ coefficients.

		Signal-to-noise ratio (dB)					
Type of noise	Number of MFCCs	20	10	0	−10	−20	−30
Factory	12	5.0	4.2	4.2	9.5	22.2	45.8
	30	3.4	3.4	4.4	13.7	25.1	43.2
Babble	12	5.2	4.8	4.2	6.7	20.2	44.5
	30	2.9	3.0	3.0	10.9	24.8	47.9

The 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 energy values (denoted by E_n). The purpose of the frame selection method is to classify this sequence into high and low values. In this study, this is performed by an application of k-means clustering.[8] The centers of two clusters are initialized with $\min(E_n)$ and $\max(E_n)$. After convergence of the k-means iteration, the cluster assignment is denoted 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 frames for which $X_n = 1$ are selected for further processing.

The system was evaluated both with and without frame selection. The k-means method was found to give better performance than no frame selection.

D. Detection rule

The detection system uses GMMs to model broad sound classes in binary classification according to the Bayes rule.[45] Each GMM has eight components and a diagonal covariance structure.[34] The GMMs are trained using ten iterations of EM (expectation-maximization) re-estimation for GMMs.[34] Before training, the component weights of the GMMs are initialized by a uniform distribution, the variance parameters of each component by 0.1 times the global variances of the features, and the mean parameters of each component by the heuristic selection approach proposed by Katsavounidis *et al.*[46]

Separate GMMs are trained for shouted speech, normal speech, and the expected noise type. 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 two-second analysis block (with a shift interval of 1 s at a time) have been selected using the unsupervised approach described in Sec. III C, the averaged logarithmic likelihoods of their corresponding feature vectors having been produced by each of the three GMMs are computed and denoted as L_{shout}, L_{speech}, and L_{noise}. The detection rule for shouted speech is

$$L = L_{\text{shout}} - \max(L_{\text{speech}}, L_{\text{noise}}) > T, \quad (1)$$

where T is the decision threshold.

In an earlier study, this detection rule was found to perform better than a direct two-way decision between shouting and non-shouting.[29] The decision threshold for the statistic given by Eq. (1) can be chosen in various ways, affecting the balance between missed detections and false detections.

IV. LISTENING TEST SETUP

The human performance in the detection of shouted speech was evaluated by a subjective listening test. Subjects were presented with samples through headphones and the task of the subject was to decide whether the sample represented shouting or not. The evaluation material, consisting of speech, shouting, and pure noise samples, was used to measure the human performance, but only with the babble noise condition in order to keep the listening test reasonable in size. Babble noise was chosen in order to focus especially on the discrimination between different types of speech: multitalker background, normal speech, and shouted speech. SNRs of 0, −10, −20, and −30 dB were evaluated.

In the test, subjects were seated in a quiet room with a graphical user interface in front, and samples were presented in random order through high-quality headphones (Sennheiser HD580). The subject could listen to each sample as many times as he or she desired before the decision. In order to prevent the loudness differences between samples of normal vocal effort and shouting from affecting the detection, the levels of the listening test samples were normalized according to ITU-T P.56.[47] Before the actual test, the subjects performed a practice session which consisted of ten samples not included in the test samples. During the practice session, the subject could adjust the volume of the headphones to a comfortable level, and during the test the volume was kept at the constant level chosen during the practice session.

The listening test material involved all the 24 sentences produced with normal vocal effort and shouting by all the 22 speakers. Each sentence was presented at four different SNRs. Thus, the total number of test sentences was $22 \times 24 \times 2 \times 4 = 4224$. In addition, a quarter of that number (1056) of pure babble noise samples were added to the test set. As a result, the total number of listening test samples was 5280.

Eight male and eight female Finnish listeners with no reported hearing problems took part in the listening test. The listeners were students or post-graduate university students. Since using all the data for every subject was impractical, in order to cover all the test material the samples were divided evenly among the listeners. Thus, each subject evaluated 330 test cases consisting of 264 speech/shout cases and 66 pure babble noise cases. The duration of the test per listener was approximately 30 min.

V. RESULTS

A. Overview

The sensitivity of the automatic detector and human listeners to detect shouting in speech samples (from 22 speakers) was investigated in conditions of variable SNR, noise type, and spectral estimation method. In addition, the speakers were categorized based on gender and on the level difference between spoken and shouted utterances, calculated as the SPL increase from normal speech to shouting (shown in Table II). The speakers were categorized into two classes of high shouters ($N_H = 11$) and low shouters ($N_L = 11$), where the decibel increase between shouting and normal speech was >22.5 dB and ≤22.5 dB, respectively.

The means of detection performance statistics across different conditions were compared with repeated measures analysis of variance (ANOVA). In the following, all statistically significant ANOVA effects pertaining to the spectral estimation methods are shown. The degrees of freedom (and, thus, p values) of the ANOVA effects were corrected with lower bound epsilon when appropriate. Pairwise *post hoc* comparisons between mean values were performed with Newman–Keuls tests.

B. Evaluation procedure for the machine system

The automatic detection 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 SNR, as well as a segment of noise equal in length to the speaker's normal speech material. Thus, the "non-shouting" part of the evaluation data had equal amounts of noisy normal speech and pure noise. The noise model of the detector was trained using two minutes of the noise material, while the remaining portion of the noise recording was used for testing.

C. Machine detection

The performance of machine detection was analyzed in different noise conditions and using different spectral estimation methods. The detection threshold for the likelihood given by Eq. (1) was adjusted in such a way that two empirical probabilities, the miss rate p_{miss} (the frequency of failing to detect a shouted speech sample) and the false alarm rate p_{fa} (the frequency of reporting shouted speech when it is not actually present), become equal. The corresponding error rate is known as equal error rate (EER) and is a widely used measure of performance in detection tasks.[3,6,38] Tables V and VI show the pooled-data EER results for factory noise and babble noise, respectively.

For each noise condition and spectral estimation method, the EER threshold was adjusted using pooled data and this threshold was used to obtain speaker-specific error rates of the form $0.5 \times p_{miss} + 0.5 \times p_{fa}$. Because of the method of threshold determination, these error rates will also be termed EERs, even though they do not consist of strictly equivalent miss and false alarm rates. The speaker-specific EER values were analyzed using ANOVA. The factors of the ANOVA consisted of SNR (−30, −20, −10, 0, 10, and 20 dB), noise type (factory noise, babble noise), and spectral

TABLE V. Equal error rates (%) for MFCC features using different spectrum analysis methods in factory noise.

	Signal-to-noise ratio (dB)					
Spectral estimation method	20	10	0	−10	−20	−30
FFT	2.7	2.4	2.9	10.1	17.8	45.7
LP	2.1	2.8	4.8	10.5	19.2	45.5
CRLP	2.9	3.8	4.5	5.2	14.3	44.5
CRWLP	2.9	3.3	3.7	5.6	14.3	44.8

TABLE VI. Equal error rates (%) for MFCC features using different spectrum analysis methods in babble noise.

	Signal-to-noise ratio (dB)					
Spectral estimation method	20	10	0	−10	−20	−30
FFT	2.7	2.2	2.4	5.6	16.7	43.0
LP	2.2	2.9	2.8	7.0	23.1	45.7
CRLP	2.8	2.3	2.8	5.8	13.6	40.2
CRWLP	2.9	2.7	3.4	5.8	13.9	40.8

estimation method (FFT, LP, CRLP, and CRWLP). The ANOVA also contained two categorical predictors, speaker gender and shouting level (low and high).

The EER of the automatic detector depended on the SNR $[F(1,18) = 394.02, p < 0.001]$, noise type $[F(1,18) = 295.1, p < 0.01]$, and on the spectral estimation method $[F(1,18) = 7.02, p < 0.05]$. The sensitivity of a given spectral estimation method to shouting also depended on the SNR $[F(1,18) = 6.51, p < 0.05]$, on the speaker gender $[F(1,18) = 7.18, p < 0.05]$, and on the combination of SNR and noise-type $[F(1,18) = 4.69, p < 0.05]$.

The effects of the spectral estimation method, SNR, and speaker gender on EER are shown in Figs. 7 and 8. The EER increased with decreasing SNR (p-values < 0.001) although the increases in EER for successive SNR values >0 dB were small (p values = not significant). Lower EER for shouting was observed in babble noise (12.1%) than in factory noise (13.1%). The differences in EER produced by distinct spectral estimation methods are compared separately for each SNR. For the three highest (≥0 dB) SNRs, the EER of different methods was roughly similar (p not significant). In conditions of SNR = −10 dB and SNR = −20 dB, the CRLP and CRWLP methods had smaller an EER than the other methods (p values < 0.05). In addition, LP had a larger EER than FFT when the SNR was −20 dB and a larger EER than CRLP and CRWLP when the SNR was −30 dB. However, as illustrated in Figs. 7 and 8, the above results depend somewhat on the noise type and speaker gender.

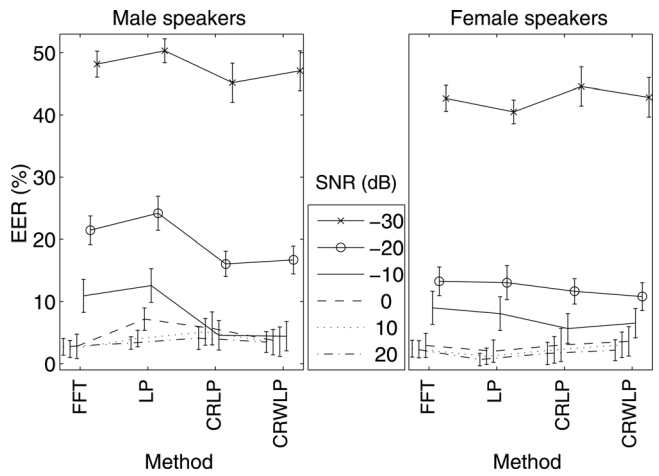

FIG. 7. Mean EER values in factory noise for the sensitivity of the automatic detector to shouting for factors spectral estimation method, SNR, and speaker gender. Error bars indicate standard errors of the mean.

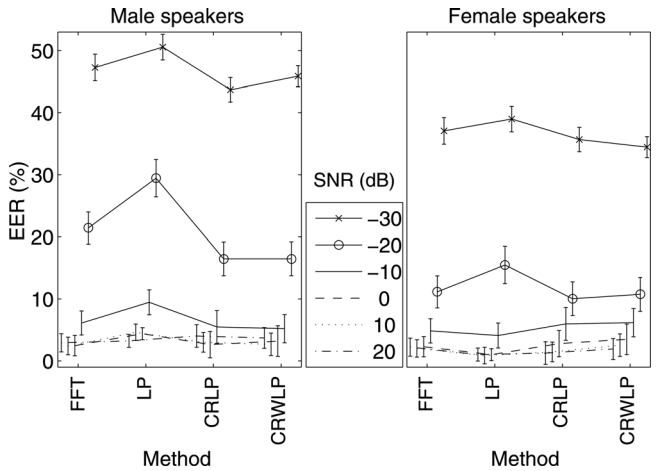

FIG. 8. Mean EER values in babble noise for the sensitivity of the automatic detector to shouting for factors spectral estimation method, SNR, and speaker gender. Error bars indicate standard errors of the mean.

The EER criterion corresponds to just one possible operating point of the detector. In order to illustrate the overall performance of the system using different spectrum analysis methods, Figs. 9 and 10 show the detection error tradeoff (DET) curves, a widely used visualization for the overall performance of a detection system,[48] for factory and babble noise, respectively, with SNRs of −20 and 10 dB.

D. Human vs machine

The behavioral data from human listeners was acquired from conditions of babble noise and four SNRs (−30, −20, −10, and 0 dB). Table VII shows the pooled miss rates and the false alarm rates for human listeners in the listening test. In comparison to the miss rates, the rate of false alarms made by humans is seen to stay remarkably low. Interestingly, male listeners appear to miss much fewer detections than the female listeners.

The "man vs machine" analyses extend the sensitivity analyses of automatic detectors by adding human male and female listeners as new "methods." In these analyses, the automatic detector was set to a detection threshold which

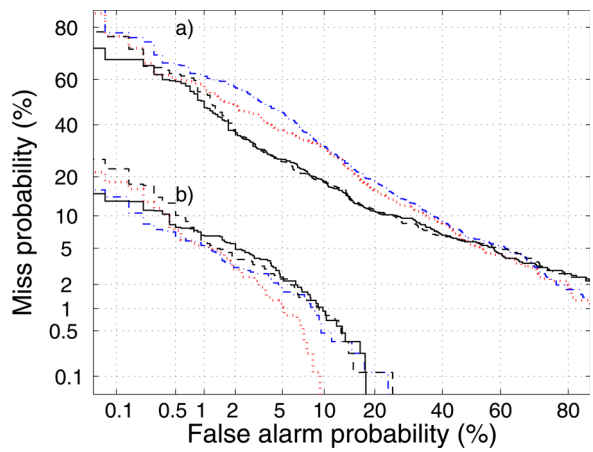

FIG. 9. (Color online) DET curves of the machine detection system for factory noise at SNR levels (a) −20 dB and (b) 10 dB.

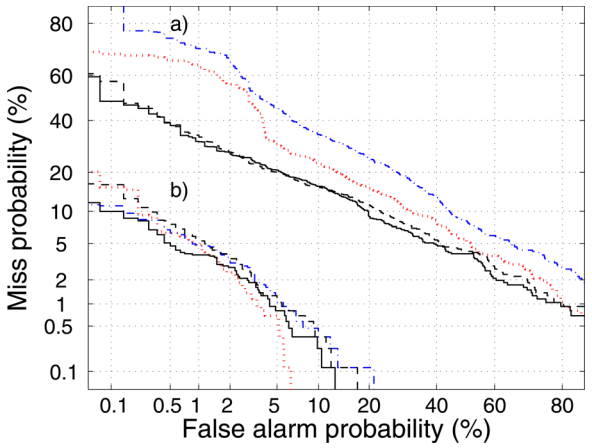

FIG. 10. (Color online) DET curves of the machine detection system for babble noise at SNR levels (a) −20 dB and (b) 10 dB.

yielded a pooled-data false alarm rate matching the total false alarm rate of the human listeners in the hardest SNR scenario −30 dB, i.e., 2.1% according to Table VII. This method of setting the detection threshold for Eq. (1), henceforth referred to as the limited false alarm (LFA) criterion, was used to obtain the speaker-specific error rates.

As a sensitivity index, the d' statistic is used.[49] The d' is calculated as $d' = Z[p_{\text{hit}}] - Z[p_{\text{fa}}]$, where Z is the inverse function of the standardized Gaussian cumulative distribution function and the hit rate $p_{\text{hit}} = 1 - p_{\text{miss}}$. The Z function can be evaluated for p values $(0, 1)$. If p_{hit} or p_{fa} was 0 or 1, a small number (10^{-6}) was added to or subtracted from the p value, respectively, to bring it in the $(0, 1)$ range.

The calculation of d' is based on transforming hits and false alarms into a sensitivity index and on the assumption of a particular underlying statistical model. The statistical model under which a similar d' can be obtained for a given detector operating with different criteria (e.g., EER or LFA) is that of normal distributions with equal variance of both hits and false alarms.[49] In cases of violations of the obtained data regarding the model assumptions, different d' may be obtained at different operating points of the detector. Therefore, the d' in the case of EER and LFA data were compared in the case of babble noise data which is used in the man vs machine comparisons. A statistically significant main effect of the criterion was found $[F(1, 18) = 7.90, p < 0.05]$ and the d' was somewhat higher in the case of the LFA criterion ($d' = 5.63$) than in the case of the EER ($d' = 5.50$) criterion. However, no statistically significant interaction effects of the criterion were found. That is, the contrasts in sensitivity between different methods appear to remain constant across different detection criteria.

The key results using the LFA criterion for the automatic detector are shown in Fig. 11. The FFT method was used as the primary reference for the performance of the human listeners. For SNR = 0 dB, both male and female listeners had lower sensitivity than the FFT (p values <0.01). For SNR = −10 dB, the sensitivity of male listeners was similar to that of the FFT, whereas female listeners had reduced sensitivity ($p < 0.001$). For SNR = −20 dB, both male and female listeners detected shouting better than the FFT method ($p < 0.001$). Finally, in the case of the lowest SNR (−30 dB), no consistent difference in sensitivity between the FFT and human listeners was observed.

In addition, the sensitivity to shouting of the proposed CR-based methods (CRLP and CRWLP) were contrasted to that of male and female human listeners. For SNR = 0 dB, both male and female listeners had lower sensitivity than CRLP and CRWLP (p values <0.01). For SNR = −10 dB, the sensitivity of male listeners was similar to that of the CR-based methods, whereas female listeners had reduced sensitivity ($p < 0.001$). For SNRs of −20 and −30 dB, no differences between CR-based methods and human listeners were found (p values not significant).

TABLE VII. Main results of the subjective listening test for shouted speech detection in babble noise by human listeners. The total, male listener and female listener false alarm rates were obtained by averaging the respective normal speech false alarm rates with the respective pure noise false alarm rates.

	Signal-to-noise ratio (dB)				
	0	−10	−20	−30	Pure noise
Miss % (total)	9.1	9.9	16.1	65.7	
Miss % (male shouting)	8.7	8.0	19.7	85.2	
Miss % (female shouting)	9.5	11.7	12.5	46.2	
Miss % (male listeners)	2.4	3.4	11.3	65.9	
Miss % (female listeners)	15.1	16.2	20.6	65.5	
False alarm % (total)	1.4	1.5	1.9	2.1	1.8
False alarm % (noisy speech only)	1.0	1.1	2.1	2.5	
False alarm % (male listeners)	1.2	0.9	1.1	2.0	1.1
False alarm % (female listeners)	1.6	2.0	2.8	2.2	2.5

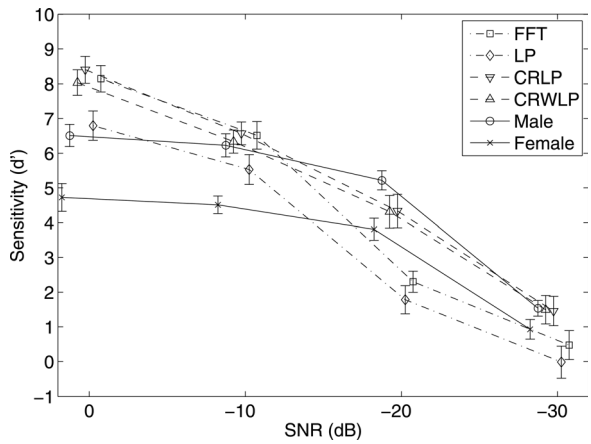

FIG. 11. Mean d' values for the sensitivity of the automatic detector and human listeners to shouting for factors analysis method (comprising different spectrum analysis methods as well as male and female listeners) and SNR. Error bars indicate standard errors of the mean.

E. Human data: Listener-wise analysis

The third part of the analyses was conducted using listener-specific d' scores. The factors of the ANOVA consisted of SNR (−30, −20, −10, and 0 dB), listener gender, and shouting level of the speaker. The effects of these variables are shown in Figs. 12 and 13.

The d' decreased with decreasing SNR $[F(1,18) = 123.50, p < 0.001]$ and was lowest for the −30 dB condition relative to the other SNRs (p values < 0.001). While the sensitivity to shouting was somewhat reduced in the −20 dB condition ($p < 0.07$) relative to the conditions with higher SNRs, no statistically significant differences were found between conditions with SNR ≥ -20 dB. The sensitivity to shouting was also higher in male ($d' = 5.00$) than in female ($d' = 3.66$) listeners $[F(1,18) = 86.736.61, p < 0.001]$. The greater sensitivity of male than that of female listeners was especially prominent in the condition of speech from low shouters $[F(1,18) = 18.50, p < 0.001]$, as shown in Fig. 13.

VI. DISCUSSION

Detection of shouted speech by human and machine in varying ambient noise conditions was studied. The main results are the following.

Detection by machine was based on Bayesian classification using auditorily motivated front-end processing. The performance of machine detection started to degrade in a statistically significant manner at segmental SNR of around −10 dB with both factory and babble noise. At a SNR of −30 dB, the detection performance approached chance level scores. The alternative spectrum analysis methods CRLP and CRWLP, which emphasize the spectral fine structure, improved upon the baseline FFT at SNR levels of −10 and −20 dB with both types of noise, confirming the authors' hypothesis. At the same time, LP analysis, which only depicts the spectrum envelope, showed the worst performance in noisy conditions. It thus appears that the role of the vocal tract excitation, manifested in the spectral fine structure, is important in the detection of shouting in noisy conditions.

Babble noise was used in the comparison between human and machine. The machine detector was tuned to an operating point where it exhibits a low rate of false alarms, corresponding to the human listeners. At the highest SNR condition included in the listening test (0 dB), the machine detector outperformed both male and female listeners. When noise was further increased, the performance of the listeners exceeded that of the baseline automatic system using FFT spectrum analysis. Both male and female listeners performed better than the FFT-based system at SNR = −20 dB. However, when FFT was substituted with either CRLP or CRWLP spectrum analysis, the machine achieved similar performance to male and female listeners also in the noisiest cases.

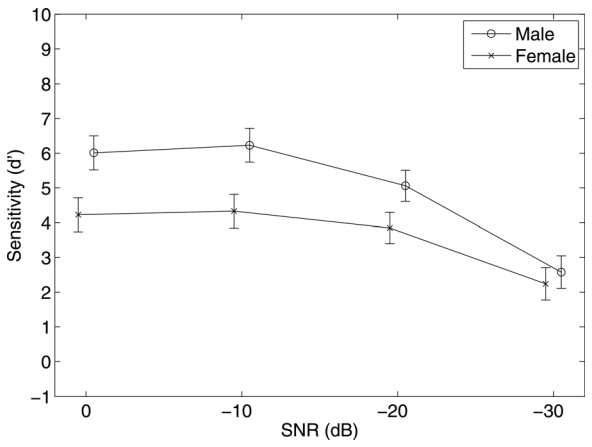

FIG. 12. Sensitivity of female and male listeners to shouting. Mean d' for factors SNR and listener gender are shown. Error bars indicate standard errors of the mean.

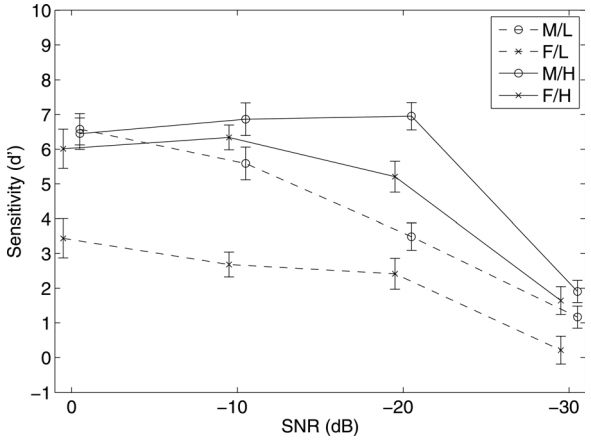

FIG. 13. Sensitivity (d') of female and male listeners to shouting. Mean d' for factors SNR, shouting class (of the speaker) and listener gender are shown. Error bars indicate standard errors of mean. M = male listeners, F = female listeners, L = low shouters, H = high shouters.

In the listening test evaluation, somewhat surprisingly, a clear sensitivity difference between male and female listeners was found in favor of the male listeners. This difference was especially prominent at the higher SNR levels, where arguably speech is not masked by noise to a degree sufficient to hide its vocal effort level. The sensitivity scores of the male and female listeners approached each other as the SNR was decreased. The difference in sensitivity appears to be primarily due to females missing the detection of more shouted samples than males at higher SNR levels. Moreover, the difference between the male and female listeners was found to be especially large in the case of low shouters, who do not raise their voice very much when shouting. The results thus suggest that male listeners are more sensitive than female listeners to even moderately raised vocal effort levels in people's speech, at least in the sense of labeling it as shouting when questioned. However, one must keep in mind that the idea of what kind of speech is considered shouting depends on many factors, including the norms of the society and the backgrounds of the individual listeners. The speakers and the listeners of this study were two groups of Finnish university students. The speakers were instructed to speak by shouting, while the listeners were asked whether they heard shouted speech. Thus, no disparity is believed to exist between the definitions used to produce the material and those used to analyze its perception. However, further studies are needed to determine whether the difference between sexes in the detection is innate or whether it depends on the culture and background of the listeners.

Even in equivalent SPL and SNR conditions, listeners detected high shouters easier than low shouters. The only case for which there was no difference was male listeners and a high enough SNR (0 dB), suggesting that in low-noise conditions males are equally capable of discriminating both high and low shouters apart from normal speech. However, increasing noise degraded this discrimination capability more for low shouters than for high shouters. Insofar as shouting detection sensitivity can be paralleled with the perception of speech loudness, the observation that higher vocal effort is generally easier to detect would appear to be supported by the connection between vocal effort and the loudness perception.[19,22] The effect of vocal effort on the detection performance was larger with female than male listeners. This, in turn, would corroborate and extend the earlier finding that different listeners place different proportional weights on vocal effort cues in the perception of the loudness of speech.[20] Recalling the importance of glottal excitation observed in machine detection and the fact that the glottal excitation appears to play an important role in the loudness effect of high vocal effort,[23] a hypothesis can be formulated: for listeners, a larger increase in the vocal effort may be easier to detect than a smaller one primarily due to the acoustical effect that the glottal excitation source has on the loudness characteristics of speech.

VII. CONCLUSIONS

This study analyzed the task of detecting deliberately high vocal effort, conceptualized as shouting, on a background of (machinery or multitalker) noise. Speech material was recorded, using the same textual content for normal and shouted speech, and artificially corrupted by noise with varying SNR. In addition, pure noise was used as test material.

In a subjective listening test conducted using multitalker noise, male listeners detected shouted speech better than female listeners. This difference was primarily due to male listeners missing the detection of much fewer shouted speech samples than females, while the rate of false detections was low for both male and female listeners. Shouting by speakers using a high SPL difference over their normal speech level was found to be more easily detected by the listeners, even though the SPL was equalized for all the listening test samples. The difference according to the shouting level was especially prominent with female listeners.

A machine system for the detection of shouted speech in ambient noise conditions was described and evaluated. The system consists of MFCC feature extraction, unsupervised frame selection based on a logarithmic frame energy, and Bayesian classification using GMMs. In the spectrum analysis for the MFCC computation, the best overall detection performance was obtained by the new CRLP and CRWLP methods. These methods use an all-pole spectrum envelope and emphasize the spectral fine structure in proportion to the estimated noisiness of the signal. The performance advantage of these methods over the baseline FFT method was statistically significant in the noisiest cases in which the system performance was not yet close to chance level. In noisy cases, FFT, in turn, was significantly better than LP which does not display the spectral fine structure. Because the spectral fine structure is closely connected with the vocal tract excitation and the F0 of voiced speech, the good performance obtained using CRLP or CRWLP in conjunction with the MFCC analysis highlights the usefulness of F0 cues in recognizing the vocal effort level within the range from normal to high vocal effort.

In the comparison between human and machine with moderate to high levels of multitalker noise, the basic machine system using the FFT spectrum analysis outperformed humans at moderate SNR levels but was outperformed by humans when the noise corruption was severe. Substitution of one of the proposed CRLP and CRWLP spectrum analysis methods, placing more weight on the vocal tract excitation cues, caused the machine to tie with humans in the noisiest cases while continuing to achieve better performance at the higher SNR levels.

ACKNOWLEDGMENTS

This work was supported by Academy of Finland (256961) and the EC FP7 project Simple4All (287678).

[1]D. Rostolland, "Phonetic structure of shouted voice," Acustica **51**, 80–89 (1982).
[2]H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," J. Acoust. Soc. Am. **107**, 3438–3451 (2000).
[3]S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," EURASIP J. Audio, Speech, Music Process. **2009**, 594103.
[4]G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2007)*, London, UK (2007), pp. 21–26.
[5]P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," Speech Commun. **54**, 732–742 (2012).
[6]E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proceedings of Interspeech 2008*, Brisbane, Australia (2008), pp. 609–612.
[7]D. Rostolland, "Acoustic features of shouted voice," Acustica **50**, 118–125 (1982).
[8]X. Huang and A. Acero and H.-W. Hon, *Spoken Language Processing* (Prentice Hall-PTR, Upper Saddle River, NJ, 2001), Chaps. 4, 6, and 9.
[9]J.-S. Liénard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**, 411–422 (1999).
[10]J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Am. **93**, 510–524 (1993).
[11]R. Schulman, "Articulatory dynamics of loud and normal speech," J. Acoust. Soc. Am. **85**, 295–312 (1989).
[12]R. B. Monsen and A. M. Engebretson, "Study of variations in the male and female glottal wave," J. Acoust. Soc. Am. **62**, 981–993 (1977).
[13]E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," J. Acoust. Soc. Am. **84**, 511–529 (1988).
[14]P. Alku, M. Airas, E. Björkner, and J. Sundberg, "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," J. Acoust. Soc. Am. **120**, 1052–1062 (2006).
[15]S. Ternström, M. Bohman, and M. Södersten, "Loud speech over noise: Some spectral attributes, with gender differences," J. Acoust. Soc. Am. **119**, 1648–1665 (2006).
[16]C. Harwardt, "Comparing the impact of raised vocal effort on various spectral parameters," in *Proceedings of Interspeech 2011*, Florence, Italy (2011), pp. 2941–2944.

[17]J. M. Pickett, "Effects of vocal force on the intelligibility of speech sounds," J. Acoust. Soc. Am. **28**, 902–905 (1956).
[18]J. F. Brandt, K. F. Ruder, and T. Shipp, Jr., "Vocal loudness and effort in continuous speech," J. Acoust. Soc. Am. **46**, 1543–1548 (1969).
[19]P. Ladefoged and N. P. McKinney, "Loudness, sound pressure, and subglottal pressure in speech," J. Acoust. Soc. Am. **35**, 454–460 (1963).
[20]G. Allen, "Acoustic level and vocal effort as cues for the loudness of speech," J. Acoust. Soc. Am. **49**, 1831–1841 (1971).
[21]A. Eriksson and H. Traunmüller, "Perception of vocal effort and distance from the speaker on the basis of vowel utterances," Attention, Percep. Psychophys. **64**, 131–139 (2002).
[22]R. D. Glave and A. C. M. Rietveld, "Is the effort dependence of speech loudness explicable on the basis of acoustical cues?," J. Acoust. Soc. Am. **58**, 875–879 (1975).
[23]G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," J. Acoust. Soc. Am. **126**, 2061–2071 (2009).
[24]A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "On the perception of emotions in speech: The role of voice quality," Logopedics Phoniatrics Vocol. **22**, 157–168 (1997).
[25]J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC'06)*, Toronto, Canada (2006), pp. 733–738.
[26]R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *Proceedings of IEEE WASPAA 2005*, New Paltz, USA (2005), pp. 158–161.
[27]I. Titze, *Principles of Voice Production*, 2nd ed. (Prentice-Hall, Englewood Cliffs, NJ, 1994), p. 220.
[28]J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *Proceedings of ICASSP 2011*, Prague, Czech Republic (2011), pp. 4968–4971.
[29]J. Pohjalainen, T. Raitio, and P. Alku, "Detection of shouted speech in the presence of ambient noise," in *Proceedings of Interspeech 2011*, Florence, Italy (2011), pp. 2621–2624.
[30]C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Proceedings of Interspeech 2007*, Antwerp, Belgium (2007), pp. 2289–2292.
[31]C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," IEEE Trans. Audio, Speech Lang. Process. **17**, 883–894 (2011).
[32]J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," IEEE Trans. Audio, Speech Lang. Process. **17**, 366–378 (2009).
[33]NOISEX-92 database, samples available online: http://spib.rice.edu/spib/select_noise.html (Last viewed 10/15/12) (1992).
[34]D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**, 72–83 (1995).
[35]D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. Interspeech 2006*, Pittsburgh, USA (2006), pp. 809–812.
[36]E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models* (Springer-Verlag, Berlin, 1990), Chap. 3.
[37]J. Pohjalainen, H. Kallasjoki, K. J. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis in noisy conditions," in *Proceedings of Interspeech 2009*, Brighton, UK (2009), pp. 1315–1318.
[38]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. **17**, 599–602 (2010).
[39]J. Makhoul, "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580 (1975).
[40]C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," Speech Commun. **12**, 69–81 (1993).
[41]L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Upper Saddle River, NJ, 1978), pp. 368–371, 419.
[42]H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**, 1738–1752 (1990).
[43]U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," Speech Commun. **50**, 142–152 (2008).
[44]J. D. Markel and A. H. Gray, *Linear Prediction of Speech, Vol. 12 of Communication and Cybernetics* (Springer-Verlag, Berlin, 1976), p. 141.
[45]S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 2nd ed. (Academic Press, New York, 2003), pp. 13–39.
[46]I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized Lloyd iteration," IEEE Signal Process. Lett. **1**, 144–146 (1994).
[47]ITU-T. Recommendation P.56: Objective measurement of active speech level (1993), pp. 3–5.
[48]A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech 1997*, Rhodes, Greece (1997), pp. 1895–1898.
[49]J. Zhang and S. T. Müller, "A note on ROC analysis and non-parametric estimate of sensitivity," Psychometrika **70**, 203–212 (2005).