Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection

Zheng-Hua Tan, *Senior Member, IEEE*, and Børge Lindberg, *Member, IEEE*

Abstract—**Frame based speech processing inherently assumes a stationary behavior of speech signals in a short period of time. Over a long time, the characteristics of the signals can change significantly and frames are not equally important, underscoring the need for frame selection. In this paper, we present a low-complexity and effective frame selection approach based on *a posteriori* signal-to-noise ratio (SNR) weighted energy distance: The use of an energy distance, instead of e.g. a standard cepstral distance, makes the approach computationally efficient and enables fine granularity search, and the use of *a posteriori* SNR weighting emphasizes the reliable regions in noisy speech signals. It is experimentally found that the approach is able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like vowels and no frames to silence, even for very low SNR signals. The resulting variable frame rate analysis method is applied to three speech processing tasks that are essential to natural interaction with intelligent environments. First, it is used for improving speech recognition performance in noisy environments. Secondly, the method is used for scalable source coding schemes in distributed speech recognition where the target bit rate is met by adjusting the frame rate. Thirdly, it is applied to voice activity detection. Very encouraging results are obtained for all three speech processing tasks.**

Index Terms—**Distributed speech recognition, frame selection, noise-robust speech recognition, variable frame rate, voice activity detection**

Z.-H. Tan and B. Lindberg are with Aalborg University, DK-9220, Aalborg, Denmark.

Corresponding author and contact details: Zheng-Hua Tan, phone: +45 9930 8686; fax: +45 9815 1583; e-mail: zt@es.aau.dk.

Source code in Matlab for noise-robust voice activity detection is available on the following page http://kom.aau.dk/~zt/online/rVAD/

I. INTRODUCTION

T HE DESIRE is strong for natural interaction with our environments pervaded with devices, automobiles, robots and smart houses filled with technology and intelligence. Naturalness implies freedom from constraint, formality or awkwardness.

Speech is apparently the most natural means for human beings to communicate and speech interaction with intelligent environments is enabled by automatic speech recognition (ASR). With recent advances, the state-of-the-art ASR technology is maturing for many applications, especially in controlled conditions. When placed in less controlled conditions, such as intelligent environments, several extra dimensions are to be considered.

First, devices in intelligent environments are generally characterized as having restricted resources and being interconnected. Consequently, the complexity of an algorithm becomes a decisive factor for its deployment. ASR systems are therefore optimized towards low-resource implementations or alternatively, a distributed architecture is adopted by distributed speech recognition (DSR) to make use of powerful servers [1].

Secondly, acoustic noises are ubiquitous in intelligent environments and dramatically degrade the ASR performance. Although sophisticated robustness algorithms have been developed, they may not be economically viable as compared to the cost associated with the performance degradation caused by the noises [2].

Thirdly, there are often no buttons within reach to support a push-and-talk mode or it is too cumbersome to use in such environments. Voice activity detection (VAD) is therefore necessary for natural speech interaction. As VAD is required to run constantly, low complexity is favored especially for low-resource devices.

This paper addresses these challenges by investigating the process of frame selection. Being a continuous time series, a speech signal is generally analyzed at short intervals, typically on the order of 10–30 ms, and with a frame length of 20–30 ms, corresponding to a few pitch-periods, to reflect the physiological constraints of speech production. This yields a series of frames each representing 20–30 ms of speech [3]. As a convention, a fixed frame rate (FFR) is deployed disregarding the signal is non-speech or speech, or is a steady region or a rapidly changing event. In most speech applications, however, there is no sense in selecting any frames for the non-speech parts. In DSR, to save bandwidth and computation and to be robust against transmission errors (by allocating more bandwidth for fast changing regions), no data should be sent to the remote server during silence and less data (frames) should be sent in steady regions than in fast changing ones.

Further, speech sounds like plosives or speech attributes like transitions can last a very short period of time, making an FFR analysis insufficient to provide a fine representation for these events, as experimentally verified in [4]. On the other hand, sounds like vowels can last a relatively long period without significant changes in characteristics and over-sampling them may generate

unnecessary frames that can increase the computational load and, even worse, increase the number of insertion errors in ASR in noisy environments [5].

Although FFR analysis assumes that speech signals exhibit quasi-stationary behavior in a short interval, no evidence supports that a fixed-rate processing is applied in the human auditory system. Clearly, the FFR analysis is not optimal [6].

As a different approach, variable frame rate (VFR) analysis aims at selecting frames according to the signal characteristics. This is realized by first extracting speech feature vectors (frames) at an FFR and then determining which frames to retain. The decision on frame selection relies on some distance measures and thresholds [4], [5], [7] − [9].

A Euclidean distance between the last retained feature vector and the current vector is calculated as the distance measure in [7]. The current frame is discarded if the measure is smaller than a predefined threshold. This approach uses only two frames for frame selection. To make use of neighboring frames of the current frame, the norm of the first derivative cepstrum vector is calculated as the distance measure in [5]. Again, a threshold based decision criterion is applied. Due to the reduced number of frames, VFR analysis saves the computation time of ASR decoding, which was one of the incentives of deploying VFR in early days. Recent research in VFR moves towards finding optimal representation of a speech signal to improve performance in e.g. noise robustness, and this requires searching frames in steps smaller than the standard 10 ms while the average frame rate largely remains the same [4], [9], [10].

The speed of singing voice changes rapidly and voices are often significantly prolonged, which can deteriorate the performance of an ordinary ASR system. A method for prolonged sound detection and elimination in singing voice recognition is presented in [8]. When the information change measure of a predefined number of successive frames is below a threshold, the group of frames is identified as a prolonged sound and consequently the following frames are omitted as long as the measure remains below the threshold. This has shown to significantly increase lyrics recognition accuracy.

VFR has demonstrated its capability in dealing with additive noise as well. In [4], Zhu and Alwan proposed an effective VFR method that uses a 25 ms frame length and a 2.5 ms frame shift for calculating Mel-frequency cepstral coefficients (MFCCs) and conducts frame selection based on an energy weighted cepstral distance. The method has shown good performance in ASR noise robustness. In [9], an entropy measure instead of a cepstral distance is used, resulting in an improvement in recognition performance as well as a higher complexity. To provide a fine resolution for rapidly changing events, these methods examine speech signals at much shorter intervals (i.e. 2.5 ms) than the normal frame shift of 10 ms. The procedure of extracting features such as MFCCs and entropy in a short interval and then discarding these, or the majority thereof, is computationally inefficient. It also limits the possibility of pursuing an even finer search granularity.

On the other hand, note that the first-order difference in frame-to-frame energy provides greater discrimination than the components of MFCCs other than c_0 [11]. Evidently, energy based search is much more computationally efficient and can

potentially enable a determination of frame shift without pre-computing cepstral feature vectors at a high and fixed rate.

In addition, speech segments are accounted in ASR not only on their characteristics, but also on their reliability. The latter is important in particular for ASR in noisy environments and is pursued in missing data theory [12] and weighted Viterbi decoding [13] methods where low signal-to-noise ratio (SNR) features are either neutralized or less weighted in the ASR decoding process. Research in [14] shows that splicing frames has the same effect as weighted Viterbi decoding under a certain assumption. It is therefore expected that VFR can benefit from SNR information e.g. selecting less frames for low SNR parts of a speech signal.

Inspired by these observations, the paper presents a low-complexity VFR method based on the measurement of *a posteriori* SNR weighted energy. To sum up, the motivations are multifold: 1) the difference in frame-to-frame energy provides a great discrimination for speech signals, 2) speech segments, besides their characteristics, are accounted also on their reliability e.g. measured by SNR, 3) the *a posteriori* SNR for noise-only segments will be theoretically equal to 0 dB, being ideal for VAD, and 4) both energy and a posteriori SNR are easy to estimate, resulting in a low complexity.

VFR has a broad spectrum of applications, ranging from computational reduction in the early days, through improved acoustic modeling and noise robustness, to prolonged speech recognition in singing voice or in spontaneous speech. In addition to noise robust ASR, the present work applies the proposed VFR method to two new applications: source coding in DSR and VAD.

The paper is organized as follows. The *a posteriori* SNR weighted energy based VFR method is presented in Section II. Frame selection and distance measurement experiments are conducted in Section III. Sections IV, V and VI apply the VFR method to noisy speech recognition for robustness, to DSR for source coding and to VAD for accuracy, respectively. Section VII concludes the paper.

II. *A POSTERIORI* SNR WEIGHTED ENERGY BASED VFR

The *a posteriori* SNR weighted energy based VFR method, as detailed in this section, conducts frame selection on the basis of an accumulative *a posteriori* SNR weighted energy distance. Since the involved calculations have a low complexity, a 1 ms frame shift and a 25 ms frame length are used to provide a fine granularity search.

A. A Posteriori SNR versus A Priori SNR

A posteriori SNR is defined as the logarithmic ratio of the energy of noisy speech to the energy of noise:

$$SNR_{post}(t) = \log \frac{E(t)}{E_{noise}(t)} \tag{1}$$

where $E(t)$ is the energy of noisy speech of frame t, and $E_{noise}(t)$ is the energy of noise of frame t.

In contrast, *a priori* SNR is the logarithmic ratio of the energy of clean speech to the energy of noise

$$SNR_{prio}(t) = \log \frac{E_{speech}(t)}{E_{noise}(t)} \tag{2}$$

where $E_{speech}(t)$ is the energy of clean speech of frame t.

Calculating *a posteriori* SNR is much more straightforward than calculating *a priori* SNR as the latter requires estimating the energy of clean speech which is a challenging task in itself.

B. The VFR Method

The frame selection is conducted through the following steps:

1. Compute the *a posteriori* SNR weighted energy distance of two consecutive frames as

$$D(t) = |\log E(t) - \log E(t-1)| \cdot SNR_{post}(t) \tag{3}$$

where $\log E(t)$ is the logarithmic energy of frame t, and $SNR_{post}(t)$ is the *a posteriori* SNR value of frame t.

2. Compute the threshold T for frame selection as

$$T(t) = \overline{D(t)} \cdot f(\log E_{noise}(t)) \tag{4}$$

where $\overline{D(t)}$ is the average weighted distance over a certain period (in this work, it is calculated over one utterance for simplicity; in practice, $\overline{D(t)}$ calculated over preceding segments can be used and it is then updated on a frame-by-frame basis). The function $f(\log E_{noise}(t))$ is a sigmoid function of $\log E_{noise}(t)$ to allow a smaller threshold and thus a higher frame rate for clean speech. The sigmoid function is defined as $f(\log E_{noise}(t)) = \alpha + \frac{\beta}{1 + e^{-2(\log E_{noise}(t)-13)}}$ where $\alpha = 9.0$ and $\beta = 2.5$ (unless stated otherwise). The constant of 13 is chosen so that the turning point of the sigmoid function is at an *a posteriori* SNR value of between 15 and 20 dB.

3. Update the accumulative distance: $A(t) += D(t)$ on a frame-by-frame basis and compare it against the threshold $T(t)$: If $A(t) > T(t)$, the current frame is selected and $A(t)$ is reset to zero; otherwise, the current frame is discarded. The search continues, that is, go back to step 1.

C. Discussion

In this work, $E_{noise}(t)$ for calculating $SNR_{post}(t)$ in (1) and for calculating $T(t)$ in (4) are both simply estimated by averaging the first 10 frames of an utterance. The first 10 frames correspond to 34 ms speech signal, as the frame shift is 1 ms, and they are considered noise only.

The parameters of the sigmoid function in (4) are set as $\alpha = 9.0$ and $\beta = 2.5$ to generate approximate 100 Hz frame rate in the final output. The 100 Hz frame rate is chosen due to the required match between the front-end frame rate and the number of states of the back-end hidden Markov models (HMMs), and a mismatch can result in a significant degradation in recognition accuracy as found in [15].

A number of variations of the algorithm will be presented and evaluated in Section IV and they include the use of a frame shift of 2.5 ms instead of 1 ms and the abandoning of the sigmoid function in (4).

The complexity of the method is relatively low since only the logarithmic energy and the *a posteriori* SNR value are calculated for each frame. The use of *a posteriori* SNR, rather than *a priori* SNR, avoids the problem of assigning zero or negative weights to frames with $SNR_{prio}(t) \leq 0dB$ and subsequently discarding them due to their non-positive weights. As such, the *a posteriori* SNR weight for noise-only frames will be theoretically equal to 0 dB, making it ideal for VAD; in practice, however, negative *a posteriori* SNR values may still appear and are then set to zero to prevent negative weights.

III. DATABASE AND FRAME SELECTION EXPERIMENTS

A. Database

Experiments in this paper are conducted on the Aurora 2 database [16], which is the TI digits database artificially distorted by adding noise and using a simulated channel distortion. Whole word models are created for all digits using the HTK recognizer [17]. Each of the whole word digit models has 16 HMM states with three Gaussian mixtures per state. The silence model has three HMM states with six Gaussian mixtures per state. A one state short pause model is tied to the second state of the silence model.

The word models used in the experiments are trained on clean speech data. The three test sets include clean speech and noisy speech corrupted by different types of noise with SNR values ranging from 0 to 20 dB. The four noise types in Test Set A are "subway", "babble", "car" and "exhibition" while the four types of noise in Test Set B are "restaurant", "station", "airport" and "street". Test Set C includes convolutional noise. The speech features are 12 MFCC coefficients, logarithmic energy as well as their corresponding velocity and acceleration components.
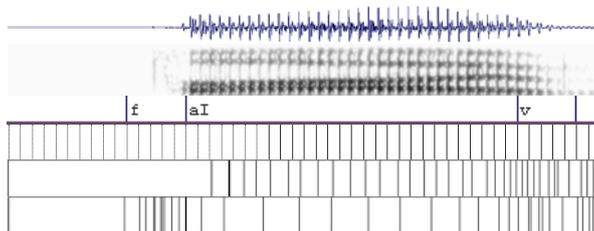
B. Frame Selection

A thorough comparison of several VFR methods was conducted in [18] and the energy weighted cepstral distance based VFR in [4] was found to outperform the others for both frame selection and speech recognition accuracy and therefore chosen as a baseline in this work. None of the compared methods, however, showed improvement over an FFR analysis in speech

recognition accuracy on a general database.

Figure 1(a) illustrates a comparison between the proposed method and the method in [4] in terms of frame selection for the clean speech of the English digit "five". The six panels in Fig. 1(a), sequentially, illustrate the waveform, the wideband spectrogram, the phoneme annotation generated by a tri-phone based HTK recognizer, the frames produced by a 100 Hz FFR analysis, the frames selected by the referenced method and the frames selected by the proposed one. Figure 1(b) shows the same comparison for 0 dB speech. To achieve maximum fairness against the reference method, in this work, the parameters for the referenced method are experimentally optimized for the Aurora 2 database.

Figure 1(a) shows that the proposed VFR assigns a higher frame rate to fast changing events such as consonants, lower frame rate to steady regions like vowels and no frames to silence, which exactly represents the objective of applying VFR analysis. In contrast, the referenced method also performs well but with one weakness namely eliminating the first part of speech following a silence in the clean speech signal. A close analysis of the referenced method reveals that its energy weight is calculated as the difference between the logarithmic energy of the frame under consideration and the mean of logarithmic energy over the whole utterance. Consequently, for clean speech, due to the significant difference in energy between silence and speech regions, the weights will be negative for a silence region and this results in negative distance values as exemplified in Fig. 2(a). The negative distance values will accumulate and thus influence the frame selection for the speech segment right after the silence region, e.g. in Fig 1.(a) where there is no frame selected for the short consonant 'f'. This is likely to be the reason why it performs well for low SNR speech, but shows no improvement on a general database.

Figure 1(b) shows that the proposed VFR method realizes an implicit VAD very well even for a 0 dB signal as there is only one frame output for the silence part, while the referenced method results in almost evenly distributed frames.
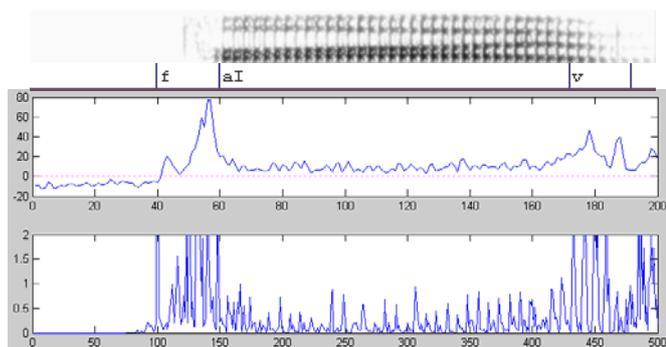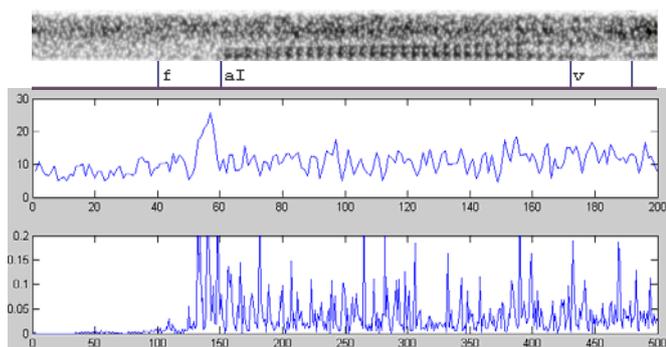

(a)
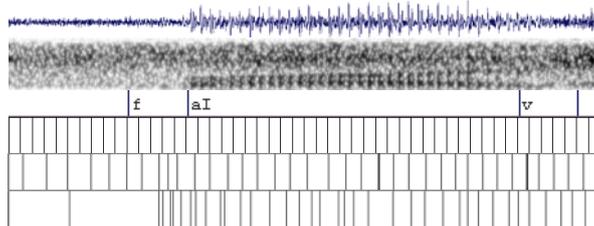
(b)

Fig. 1. Frame selection for the English digit "five": (a) For clean speech: waveform (the 1st panel), spectrogram (the 2nd panel), phoneme annotation (the 3rd panel), the frames produced by a 100 Hz FFR analysis (the 4th panel), the frames selected by the referenced method [4] (the 5th panel), and the frames selected by the proposed method (the 6th panel); (b) for 0 dB speech with the same order of panels as in (a).

Figure 2(a) illustrates the energy weighted Euclidean MFCC distance used by the referenced method and the proposed *a posteriori* SNR weighted energy distance for the clean speech of the English digit "five". Figure 2(b) shows the same comparison for 0 dB speech. The results verify that due to the weighting of the *a posteriori* SNR, the distance $D(t)$ as given in (3) is close to zero in the silence region for both clean and noisy speech.

(a)

(b)

Fig. 2. Distance measurement for the English digit "five": (a) For clean speech: spectrogram (the 1st panel), phoneme annotation (the 2nd panel), the energy weighted Euclidean MFCC distance used by the referenced method (the 3rd panel), and the proposed *a posteriori* SNR weighted energy distance (the 4th panel); (b) for 0 dB speech with the same order of panels as in (a).

In the following sections, the VFR method is applied to noise robust ASR and two new applications namely DSR and VAD.

IV. NOISE ROBUST SPEECH RECOGNITION

Poor robustness is considered the primary barrier to the widespread adoption of ASR technology. Noise robustness is therefore a key metric in measuring ASR performance, especially in intelligent environments with large acoustic variations.

In general, robustness methods aim at reducing the mismatches between the training and test speech signals through feature-domain or model-domain methods. Feature based methods include feature enhancement, distribution normalization and noise robust feature extraction. Feature enhancement attempts to remove the noise from the signal, such as in spectral subtraction (SS) [19] and in Vector Taylor Series (VTS) [20]. Distribution normalization reduces the distribution mismatches between training and test speech, for example in cepstral mean subtraction (CMS) [21] and in cepstral mean and variance normalization (CMVN) [22]. Noise robust features include improved MFCCs e.g. root-cepstrum [23].

Note that noise robustness techniques are generally applied either in the feature domain, in the model domain or in both of them. On the other hand, VFR analysis works in the time domain in the sense that it determines which time frame to retain and more importantly, it has shown good performance in noise robustness. This attribute gives it a great potential to be combined with other methods, such as complementary spectral- and cepstral-domain enhancement methods, to achieve a truly cumulative improvement.

A. VFR Combined with Spectral- and Cepstral-Domain Methods

VFR analysis relies on some distance measures for frame selection. These measures, however, can be largely affected by noises that corrupt the speech signal. If the noisy speech signal is first de-noised by a speech enhancement method as e.g. SS and thereafter analyzed by the VFR method, it is expected that applying the speech enhancement method will both enhance the speech signal and improve the frame selection.

The speech enhancement method applied in this work is the minimum statistics noise estimation (MSNE) [24] based SS. MSNE assumes that speech cannot occupy a frequency bin all the time and thus treats the minimum value of each frequency bin in the power spectral density domain within a long-enough window as the noise estimate of the current frame. This method gets rid of the VAD and is capable of tracking noise changes even within speech segments.

The joint time- and spectral-domain method is further combined with the method which consists of Mean subtraction, Variance normalization and Auto-regression moving-average based filtering (MVA) in the cepstral domain [25]. Here, the MVA processing is applied to the static MFCC features only.

B. Experimental Results

ASR experiments were conducted on the Aurora 2 database introduced in Subsection III.A. The word error rate (WER) results for a number of methods are presented in Table I. In the table, Cep-VFR refers to the energy weighted cepstral distance based VFR with parameters optimized for this task. The Cep-VFR method unfortunately does not give an acceptable performance for clean speech. The reason is that the potential negative distance results in no frames output for the first part of speech right after the silence which is often a short-duration consonant, as exemplified in Figs. 1(a) and 2(a).

TABLE I

Percent WER across the Methods for Test Set A. The Results for Cep-VFR + VAD Are Cited from [9], the Results for LogE-VFR Are Cited from [10] and the Results for MVA Are Cited from [25].

Methods	0 ~ 20 dB (Average)	Clean
FFR baseline	38.7	1.0
Cep-VFR	29.5	3.5
Cep-VFR + VAD	30.0	1.4
LogE-VFR	31.4	1.1
SNR-LogE-VFR	28.7	1.4
MSNE-SS	33.7	1.5
MSNE-SS +SNR-LogE-VFR	21.6	1.3
MVA	24.8	1.0
MSNE-SS + MVA+ SNR-LogE-VFR	19.0	1.4

The low performance of Cep-VFR on clean speech can be improved by combining it with a VAD to remove the silence from the speech signal. The results for the Cep-VFR method combined with VAD (Cep-VFR+VAD) presented in Table I are cited from [9] and they show that Cep-VFR+VAD gives a good performance for both clean and noisy speech.

An energy based VFR (LogE-VFR) is presented in [10] that uses a delta logarithmic energy as the criterion for determining the size of the frame shift on the basis of a sample-by-sample search. The results of LogE-VFR, cited from [10] and included in Table I, show LogE-VFR obtains a performance on clean speech comparable to that of the FFR baseline as well as a good performance on noisy speech although worse than both Cep-VFR and Cep-VFR+VAD.

Finally the proposed method (SNR-LogE-VFR) demonstrates the best performance for noisy speech and a good performance for clean speech. As compared with Cep-VFR+VAD, it has a substantially lower complexity and no support from VAD (need for a rough estimation of $E_{noise}(t)$ but no explicit need for a VAD).

Table I also shows the results for the MSNE based spectral subtraction (MSNE-SS) and its combination with the *a posteriori* SNR weighted energy based VFR. It is observed that the combination of the proposed SNR-LogE-VFR and MSNE-SS achieves a 17.1% absolute WER reduction on noisy speech as compared with the FFR baseline. Interestingly, the improvement of the combined method is greater than the summation of the gains obtained by applying the two methods individually (10.0% for

SNR-LogE-VFR and 5% for MSNE-SS) – it is often the opposite way when combining two methods. This justifies the dual contributions of speech enhancement when combined with the VFR method, i.e. improving frame selection and enhancing speech.

The last part of Table I gives the results for MVA and for the combination of SNR-LogE-VFR, MSNE-SS and MVA. The performance for the MVA is cited from [25]. The results show that the combination with MVA further improves the performance and suggest that the VFR method is orthogonal to other methods. The method is expected to benefit from combination with other advanced methods as well, especially model based noise-robustness methods.

C. Parameters for Frame Selection Threshold

This subsection investigates the effect of the settings of parameters for calculating frame selection threshold, i.e., α and β in the function $f(\log E_{noise}(t)) = \alpha + \dfrac{\beta}{1 + e^{-2(\log E_{noise}(t)-13)}}$ in (4). The default setting is $\alpha = 9.0$ and $\beta = 2.5$. In the experiments here, the two parameters are varied across a range, which are shown together with their corresponding WER results in Table II.

TABLE II

PERCENT WER ACROSS DIFFERENT PARAMETER SETTINGS FOR TEST SET A.

Methods	$0 \sim 20$ dB (Average)	Clean
($\alpha_0 = 9.0$, $\beta_0 = 2.5$)	28.7	1.4
($\alpha_0 + 0.5$, β_0)	28.8	1.4
($\alpha_0 - 0.5$, β_0)	28.9	1.4
(α_0 , $\beta_0 + 0.5$)	28.7	1.4
(α_0 , $\beta_0 - 0.5$)	28.9	1.4
($\alpha_0 - 1$, $\beta_0 + 1$)	28.5	1.4
($\alpha_0 - 2$, $\beta_0 + 2$)	28.2	1.5
($\alpha_0 + 1$, $\beta_0 - 1$)	29.4	1.5
($\alpha_0 + 2.5$, $\beta_0 - 2.5 = 0$)	30.9	1.7

In the first set of experiments, either α or β is increased or is decreased by 0.5 as compared with the default settings. Table II shows the resulting changes in WER are negligible. In the second set of experiments, if α is increased, β is then decreased by the same amount, vice versa. Decreasing α while increasing β means further reducing the frame rate for noisy speech. Table II shows that increasing β from 2.5 to 4.5 gives even better overall performance. On the other hand, decreasing β while increasing α means the difference in frame rate between clean and noisy speech is reduced as compared with the default setting, in which moderate performance degradation is observed (note that the changes in parameters are quite large). With the

setting ($\alpha_0 + 2.5$, $\beta_0 - 2.5 = 0$), the threshold is equal to $T(t) = \overline{D(t)} \cdot f(\log E_{noise}(t)) = \overline{D(t)}$, corresponding to no use of the sigmoid function. This gives the same frame rate for both clean and noisy speech. Overall, the results indicate that varying the parameters around the default setting only has a minor influence on the resulting WER.

The default frame shift value is set to 1 ms in Section II. A different frame shift value can be used which requires different α and β as well due to the requirement of approximate 100 Hz frame rate. When the frame shift is set to 2.5 ms, α and β are set as 2.6 and 1.2, respectively. With these settings, the average WER for 0~20 dB noise speech is 28.5% and the WER for clean speech is 1.6%. It is observed that WER for clean speech increases slightly as compared to the case of 1 ms frame shift. With varying values of α and β , trends similar to that of Table II are observed.

D. Analysis of Recognition Error Types

Experiments are conducted to investigate the behavior of VFR through the analysis of recognition error types. In the experiments, only the "Babble Noise" subset of Test Set A is used which consists of 3308 words.

Table III shows the analysis for clean speech for the FFR baseline, Cep-VFR, the *a posteriori* SNR weighted energy based VFR, the MSNE based SS and the combination of SNR-LogE-VFR with MSNE-SS and MVA methods. It is common that noise robustness algorithms increase the WER for clean speech and as shown in Table III, this is the case for the VFR methods as well. The number of insertion errors, however, still decreases after applying the VFR methods.

TABLE III

NUMBER OF CORRECT WORDS (H), DELETIONS (D), SUBSTITUTIONS (S) AND INSERTIONS (I) ON CLEAN SPEECH (A SUBSET OF TEST SET A WITH 3308 WORDS IN TOTAL).

Methods	H	D	S	I	% WER
FFR Baseline	3285	10	13	10	1.0
Cep-VFR	3180	45	83	3	4.0
SNR-LogE-VFR	3263	11	34	7	1.6
MSNE-SS	3273	14	21	15	1.5
MSNE-SS+SNR-LogE-VFR	3265	12	31	4	1.4
MSNE-SS + MVA+ SNR-LogE-VFR	3260	14	34	5	1.6

Table IV shows the same analysis for 10 dB datasets corrupted by "Babble Noise", respectively. It is interesting to note: 1) the number of correctly recognized words steadily increases after applying VFR, MSNE-SS and MVA, 2) the number of substitutions steadily decreases after applying VFR, MSNE-SS and MVA, 3) the number of insertions decreases very substantially applying VFR and MVA, and 4) the number of deletions increases slightly after applying VFR and MVA. It is clear that the most significant performance improvement comes from the reduction of insertion errors. The same trends are observed for 0 dB dataset.

TABLE IV

NUMBER OF CORRECT WORDS (H), DELETIONS (D), SUBSTITUTIONS (S) AND INSERTIONS (I) ON 10 dB SPEECH CORRUPTED BY "BABBLE NOISE" (A SUBSET OF

TEST SET A WITH 3308 WORDS IN TOTAL).

Methods	H	D	S	I	% WER
FFR Baseline	2700	102	506	1065	50.6
Cep-VFR	2952	79	277	294	19.6
SNR-LogE-VFR	2772	138	398	65	18.2
MSNE-SS	2885	78	345	828	37.8
MSNE-SS+SNR-LogE-VFR	2917	103	288	100	14.8
MSNE-SS + MVA+ SNR-LogE-VFR	2965	169	174	14	10.8

To achieve maximum absolute performance, noise robust features that are compatible with standard MFCCs can be evaluated and a nice overview is provided in [26]. Among others, subspace based approaches such as linear discriminative analysis [27] are of particular interest.

V. SOURCE CODING IN DSR

To take advantage of the resources available over networks, DSR employs the client-server architecture and submits the computation-intensive ASR decoding task to a powerful server [1]. Specifically, speech features estimated for ASR are compressed and transmitted through networks to a server. In the server the features are decoded and used for recognition. This architecture relieves the burden of computation, memory and energy consumption from low-resource devices. As a side effect, the distributed solution requires data compression.

As shown in previous sections, the VFR method aims at a high time resolution for fast changing events and a low time resolution for steady regions. The same philosophy is applied as well in data compression in DSR (and Voice-over-IP). Frame selection in the feature extraction process optimized over a certain period in the VFR analysis is likely of benefit to the data compression which follows right after the feature extraction.

This motivates us to use the VFR method for data compression. The target bit rate for DSR is simply realized by choosing a proper frame rate. For the purpose of comparison, we optimize the SNR-LogE-VFR, by constraining the range of the frame selection search, to give a comparable performance on clean speech to the ETSI-DSR FFR baseline. After applying split vector quantization, this gives a DSR front-end with a bit rate of approximately 3.5 kbps (SNR-LogE-VFR-DSR) and its recognition results are shown in Table V.

A bit rate of approximately 1.5 kbps is implemented as well and to restore the original frame rate for the match between the

frame rate and the applied HMMs, frame repetition is applied in the server. The work in [15] shows that there is a strong correlation between the number of states of the back-end HMM models and the frame rate used in the front-end and a mismatch between the two introduces a significant increase in ASR WER. The mismatch can as well be removed by using a smaller number of HMM states, at the expense of additional acoustic model sets.

An efficient compression method in DSR is the two-dimensional Discrete Cosine Transform (2D-DCT) based code [28]. More recently, the group of pictures concept (GoP) from video coding was applied to DSR to achieve a variable-bit-rate interframe compression scheme [29]. The results for these methods (2D-DCT and GoP) are cited and presented in Table V. Since there may exist mismatches in training/testing between the various simulation systems in the references, the comparisons are indicative only.

Note that the ETSI-DSR standard uses a split vector quantization for data compression without exploiting interframe information [36].

Experimental results in Table V show that the VFR based data compression is significantly superior to the 2D-DCT method and the GoP one.

TABLE V

PERCENT WER ACROSS THE DATA COMPRESSION METHODS FOR TEST SET A. THE RESULTS FOR 2D-DCT AND GOP ARE CITED FROM [28] AND [29], RESPECTIVELY.

Methods	kbps (payload)	$0 \sim 20$ dB (Average)	Clean
ETSI-DSR	4.40	39.8	1.0
2D-DCT	1.45	40.5	1.6
GOP	2.57	N/A	2.5
GOP	1.27	N/A	2.6
SNR-LogE-VFR-DSR	3.50	33.7	1.0
SNR-LogE-VFR-DSR	1.50	32.8	1.2

VI. VOICE ACTIVITY DETECTION

Widely used in real-world speech systems, voice activity detection attempts to detect the presence or absence of speech in a segment of an acoustic signal [30]. The detected non-speech segments can subsequently be abandoned to improve the overall performance of these systems. For instance, DSR makes use of a VAD to avoid unnecessary processing and transmission of silence regions and thus save on computational resources and on network bandwidth. As another example, speech recognition systems can take advantage of a VAD to reduce recognition error rates as demonstrated by the combination of the energy weighted cepstral distance based VFR and a VAD [9].

In general, voice activity detection is realized in two key steps: First, some features are calculated from a segment of the

acoustic signal; secondly, a classifier is applied to the features to categorize the segment as speech or non-speech.

Speech features used for VAD include both classical ones, e.g. energy and zero crossing rate, and more sophisticated ones, e.g. entropy [31] and Mel-filter bank outputs [32] that have recently been proposed. In terms of classification, techniques such as support vector machines [33], Gaussian mixture models [34] and decision trees [35] have been used. The simplest technique is a threshold based approach in which the decision is made by comparing the calculated value(s) against certain threshold(s).

Two different VAD algorithms are used by the ETSI advanced front-end for different purposes [36]. The first one is energy based and is used for noise estimation, while the second marks each 10 ms frame in an utterance as speech/non-speech so that the information can be used for frame dropping at the server recognizer. Only the second algorithm is further analyzed in this paper. It has two stages: a frame-by-frame stage consisting of three measures (whole spectrum, spectral sub-region and spectral variance) and a decision stage analyzing the pattern of buffered measurements for making the VAD decision.

Voice-over-IP standards include VAD algorithms as well. The G.729 VAD algorithm uses the following features: full- and low-band frame energy, a set of line spectral frequencies and the frame zero-crossing rate [37]. The G.723.1 VAD algorithm compares the energy of the inverse filtered signal with a threshold [38].

It is a unsolved problem to develop VAD methods that are accurate in both clean and noisy environments. Further, accuracy, latency and complexity are considered key metrics for measuring and comparing VAD methods. Complexity is important since a VAD applies to various applications which often involve low-resource devices.

The ETSI advanced front-end VAD performs very well in noisy environments, but very poor in noiseless conditions. A comparison in [39] also shows that the advanced front-end VAD is primarily suitable for stationary noise environments. The Mel-filter bank outputs based VAD [32] is highly accurate for clean speech, but its performance in noisy environments is worse than the advanced front-end VAD in terms of frame error rate. Both are significantly superior to the G.729 and G.723.1 VAD algorithms.

As shown in Section III.B, the *a posteriori* SNR weighted energy based VFR method is able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like vowels and no frames to silence, even for very low SNR signals. This motivates us to further process the selected frames for speech/non-speech classification, leading to a high-accuracy and low-complexity VAD method that performs well in both clean and noisy environments.

A. VAD Decision Based on VFR Selected Frames

A moving average is applied to the frames selected by the VFR algorithm as detailed in Subsection II.B. The moving average $M(n)$ is calculated on the basis of a 10 ms frame shift and is measured as the average number of frames within the moving average window as follows.

$$M(n) = \frac{1}{m_1 + m_2 + 1} \sum_{m=-m_1}^{m_2} frame\_selection(\gamma \times (n + m)) \tag{5}$$

where the function $frame\_selection(t)$ represents whether the t-th frame is selected or not in the frame selection process: The value is 1 if selected and 0 if not. The constant $\gamma = 10$ maps the 1 ms frame shift for VFR frame selection to the 10 ms frame shift for VAD.

It is a central moving average when $m_1 = m_2$, a prior moving average when $m_2 = 0$, and a biased moving average when $m_1 \neq m_2$. The latency of the VAD method is controlled by adjusting m_2.

The output of the moving average $M(n)$ is compared against a threshold T_{vad}: If $M(n) > T_{vad}$, the current frame is classified as speech; otherwise, the current frame as non-speech.

The use of *a posteriori* SNR was introduced in a very recent work in [40] together with *a priori* SNR and predicted SNR as principal parameters of support vector machine based VAD. The method presented in this paper, however, is based on three factors: energy distance, *a posteriori* SNR weighting and accumulation. Further the method first conducts frame selection as done in VFR analysis and then applies VAD decision on the frame selection results. Alternatively, the *a posteriori* SNR weighted energy can be used for VAD decision directly.

B. Generation of Frame-Based Reference VAD

The frame-by-frame reference VAD is generated from forced-alignment speech recognition experiments. Whole word models are trained on clean speech data for all digits using the HTK recognizer as described in III.A.

The trained word models are used for performing forced-alignment for the 4004 utterances (clean speech) from which all utterances in Test Set A, B and C are derived from by adding noise. The forced-alignment results are used to set the time boundaries for speech segments to create a frame-based reference VAD.

C. Illustrative VAD Results

To provide some insight about the VAD process, the intermediate and final VAD results for two input speech signals are depicted in Fig. 3. In this experiment a 37-point central moving average is applied. The utterances are the English digits "five nine four" in noiseless and 5 dB noisy environments. The figure presents the results on waveform, spectrogram, frames selected by the proposed method and VAD experiments.
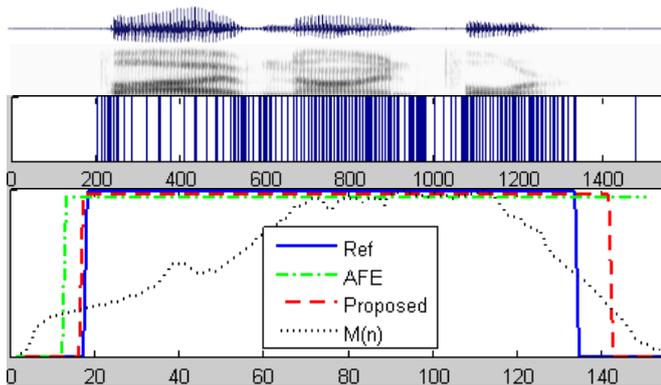
(a)
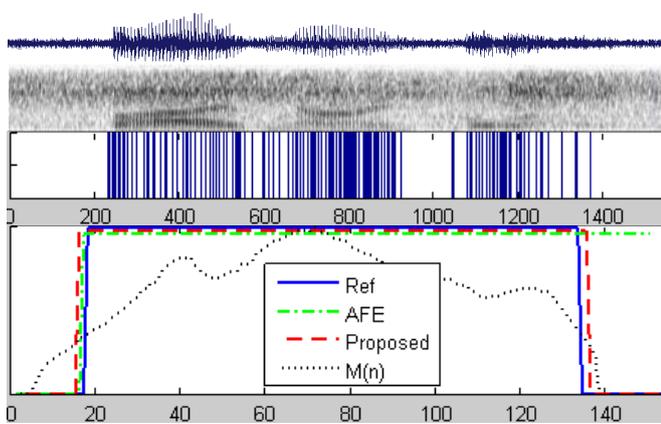
(b)

Fig. 3. VAD experiment for the English digits "five nine four": (a) For clean speech: waveform (the 1st panel), spectrogram (the 2nd panel), frames selected by the proposed method (the 3rd panel), VAD results (the 4th panel: the solid blue for the reference VAD, the dash-dot green for the advanced front-end VAD, the dashed red for the proposed VAD and the black dotted for the moving average $M(n)$); (b) for 5 dB speech with the same order of panels as in (a).

The illustration shows that the VAD method performs well even for a speech signal of 5 dB in terms of both frame selection and VAD decision. Also it is observed that the VAD result of the proposed method is more precise than that of the advanced front-end.

D. Performance Comparison

Frame error rates for several VAD methods on the Aurora 2 Test Set A, B and C are presented in Table VI. Results for G.729, G.723.1 and MFB VAD methods are cited from [32].

TABLE VI

PERCENTAGE OF FRAME ERRORS OBTAINED BY SEVERAL METHODS ON AURORA 2 DATABASE TEST SET A, B AND C. THE RESULTS FOR G.729 VAD, G.723.1

VAD AND, MFB VAD ARE CITED FROM [32].

Methods	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Ave.
G.729	12.8	24.5	26.1	27.4	29.1	32.2	35.2	26.8
G.723.1	19.5	21.3	23.3	24.4	26.3	26.6	28.6	24.3
MFB	6.9	15.4	17.7	20.1	22.8	26.2	31.1	20.0
DSR AFE	18.4	15.2	15.0	14.6	14.5	15.6	22.1	16.5
Proposed	8.1	8.3	9.0	10.6	13.5	19.5	28.2	13.9

Table VI shows that the average frame error rate of the proposed method is significantly lower than those of the referenced methods. The proposed method is substantially superior to the G.729 VAD [37] and the G.723.1 VAD [38] in all conditions, to the Mel-filter bank VAD [32] except for noiseless condition and to the advanced front-end VAD [36] except for very low SNR signals (0 dB and -5 dB).

Speech recognition experiments were carried out on Test Set A for the proposed VAD. The obtained WERs are 28.9% and 1.0% for noisy and clean speech, respectively, indicating its effectiveness in speech recognition.

E. *Latency Experiments*

Since a 37-point central moving average is applied in the proposed VAD method, this gives a latency of 18 frames. To eliminate this latency, further experiments are conducted by using a prior moving average that depends on preceding data only. The use of a prior moving average will result in wrongly classifying the first several speech frames as non-speech. To handle this problem, an adaptive threshold is applied as follows.

$$T_{vad}(n) = T_{vad} - \frac{1}{3}(m_1 - \sum_{m=-m_1}^{-1} vad\_decision(n+m)) \tag{6}$$

where the function $vad\_decision(n)$ is the VAD decision at the n–th frame: 1 for speech and 0 for non-speech. The results of this modified VAD method (Modification 1) are presented in the second row of Table VII.

Further, in Step 2 of the algorithm presented in Section II.B, the average weighted distance $\overline{D(t)}$ in (4) is calculated over an entire utterance, i.e. relying on knowledge of future observations. To avoid the latency caused by this, the following estimation of $\overline{D(t)}$ is applied:

$$\overline{D(t)} = \lambda \cdot \overline{D(t-1)} + (1-\lambda) \cdot D(t) \tag{7}$$

This modification is combined with the use of a prior moving average (i.e. Modification 1), resulting in an implementation of the method with 0 frame latency (Modification 2). Its results with $\lambda = 0.9995$ are shown in the third row of Table VII.

The performance of the method with 0 frame latency degrades significantly as compared to the one with latency, demanding a further research.

TABLE VII

PERCENTAGE OF FRAME ERRORS OBTAINED BY THE PROPOSED METHOD WITH MODIFICATIONS ON AURORA 2 DATABASE TEST SET A ACROSS SNR VALUES.

Methods	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Ave.
Modification 1	10.3	11.6	12.7	14.2	16.2	20.1	26.5	15.9
Modification 2	10.7	11.2	12.4	14.1	16.8	21.5	28.7	16.5

VII. CONCLUSIONS

The contributions of this paper are multifold. First, the accumulative *a posteriori* SNR weighted energy distance based VFR was presented. In terms of frame selection, the method is able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like vowels and no frames to silence, even for very low SNR signals.

Secondly, the VFR method was applied to noise-robust ASR and was combined with spectral- and cepstral-domain methods. Encouraging results were obtained. Further experiments were conducted to investigate the behavior of the VFR through the analysis of ASR error types. It was found that the decrease in the number of insertion errors is the most significant reason for ASR accuracy improvement. The secondary reasons are the increase of the number of correct words and the decrease of the number of substitutions. The number of deletions, however, slightly increases.

Moreover, the VFR method was applied to two new applications namely DSR and VAD. The employment of the VFR method in DSR for data compression results in an efficient and scalable DSR coding scheme. Its employment in VAD derives an accurate VAD method.

Importantly, the proposed method has the advantage of a low complexity.

Future work includes applying the accumulative *a posteriori* SNR weighted energy distance directly for VAD decision by bypassing the frame selection step and applying a weighted prior moving average over the distance measure. The VFR method is expected to benefit from a combination with other advanced methods as well, especially model based noise-robustness methods.

REFERENCES

[1] Z.-H. Tan and B. Lindberg (eds.), *Automatic Speech Recognition on Mobile Devices and Over Communication Networks*. Springer-Verlag, London, 2008.

[2] J. Cohen, "Embedded speech recognition applications in mobile phones: status, trends, and challenges," in *Proceedings of ICASSP 2008*, Las Vegas, USA, 2008.

[3] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals, 2nd.* Edition, Wiley-IEEE Press, 1999.

[4] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in Proc. IEEE ICASSP, pp. 3264–3267, 2000.

[5] P. Le Cerf and D. Van Compernolle, "A new variable frame rate analysis method for speech recognition," IEEE Signal Processing Letters, 1(12), pp. 185–187 1994.

[6] S.J. Young and D. Rainton, "Optimal frame rate analysis for speech recognition," IEE Colloquium on Techniques for Speech Processing, Dec 1990.

[7] K. M. Pointing and S. M. Peeling, "The use of variable frame rate analysis in speech recognition," *Computer Speech and Language*, vol. 5, no. 2, 1991, pp. 169–179.

[8] A. Sasou, "Singing voice recognition considering high-pitched and prolonged sounds," in Proc. Eusipco 2006, Florence, Italy.

[9] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR", in Proc. IEEE ICASSP, 2004.

[10] J. Epps and E. Choi, "An energy search approach to variable frame rate front-end processing for robust ASR," in *Proc. Eurospeech 2005*, Lisbon, Portugal, 2005.

[11] E. L. Bocchieri and J. G. Wilpon, "Discriminative analysis for feature reduction in automatic speech recognition," in Proc. IEEE ICASSP, 1992.

[12] C. Cerisara, S. Demangea and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer Speech & Language*, vol. 21, no. 3, July 2007, pp. 443-457.

[13] Yoma, N.B., McInnes, F.R., Jack, M.A., "Weighted Viterbi algorithm and state duration modelling for speech recognition in noise", in *Proc. ICASSP 1998*, Seattle, WA, 1998.

[14] Zheng-Hua Tan, Paul Dalsgaard and Borge Lindberg, "Partial splicing packet loss concealment for distributed speech recognition," *IEE Electronics Letters*, vol. 39, no. 22, October 2003, pp. 1619-1620.

[15] Z.-H. Tan, P. Dalsgaard and B. Lindberg, "Exploiting temporal correlation of speech for error-robust and bandwidth-flexible distributed speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, May 2007, pp. 1391-1403.

[16] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, Paris, France, 2000.

[17] S. J. Young et al., HTK: Hidden Markov Model Toolkit V3.2.1, Reference Manual. Cambridge, U.K.: Cambridge Univ. Speech Group, 2004.

[18] J. Macias-Guarasa, J. Ordonez, J. M. Montero, J. Ferreiros, R. Cordoba and L. F. D. Haro, "Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition," in *Proc. Eurospeech 2003*, Geneva, Switzerland, September 2003.

[19] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, 1979, pp. 113-120.

[20] J. Droppo, L. Deng and A. Acero, "A comparison of three non-linear observation models for noisy speech features," in *Proc. Eurospeech 2003*, Geneva, Switzerland, September 2003.

[21] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, Apr 1981, pp. 254- 272.

[22] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol.25, no.1–3, 1998, pp.133-147.

[23] R. Sarikaya and J.H.L. Hansen, "Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition," in *Proc. Eurospeech* 2001, Aalborg, Denmark, Sepember 2001.

[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, 2001, pp. 504-512.

[25] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, 2007, pp. 257-270.

[26] Q. Zhu and A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," Computer, Speech, and Language, vol. 17, no. 4, October 2003, pp. 381-402.

[27] R. Haeb-Umbach, and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in Proc. *ICASSP 1992*, San Francisco, CA, USA, 1992.

[28] W.-H Hsu and L.-S. Lee, "Efficient and robust distributed speech recognition (DSR) over wireless fading channels: 2D-DCT compression, iterative bit allocation, short BCH code and interleaving", in *Proc. IEEE ICASSP 2004*, Montreal, Quebec, Canada, 2004.

[29] B.J. Borgstrom and A. Alwan, "A packetization and variable bitrate interframe compression scheme for vector quantizer-based distributed speech recognition", in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[30] J. Ramirez, C. Segura, C. Benitez, A. Torre and A. Rubio, "A new Kullback-Leibler VAD for speech recognition in noise," *IEEE Signal Processing Letters*, vol. 11, no. 2, 2004.

[31] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, September 2001.

[32] D. Vlaj, B. Kotnik, B. Horvat and Z. Kacic, "A computationally efficient mel-filter bank VAD algorithm for distributed speech recognition systems", *EURASIP Journal of Applied Signal Processing* 2005:4, 487-497.

[33] Dong, E., Liu, G., Zhou, Y. and Zhang, X., "Applying support vector machines to voice activity detection," in *Proc. ICSLP 2002*, Denver, USA, 2002.

[34] Shah, J.K., Iyer, A.N., Smolenski, B.Y. and Yantorno, R.E., "Robust voiced/unvoiced classification using novel features and Gaussian mixture model," in *Proc. ICASSP 2004*, Montreal, Quebec, Canada, 2004.

[35] Shin, W.-H., Lee, B.-S., Lee, Y.-K. and Lee, J.-S., "Speech/non-speech classification using multiple features for robust endpoint detection," in *Proc. ICASSP 2002*, Orlando, Florida, USA, 2002.

[36] ETSI, "Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm," ES 202 050 v1.1.1, 2002.

[37] ITU, "Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear-prediction (CS-ACELP) Annex B: A silence compression scheme," ITU Recommendation G.729, 1996.

[38] ITU, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. Annex A: Silence compression scheme," ITU Recommendation G.723.1, 1996.

[39] M. Fujimoto, K. Ishizuka and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in Proc. *ICASSP 2008*, Las Vegas, Nevada, USA, 2008.

[40] J.W. Shin, J.-H- Chang and N.S. Kim, "Voice activity detection based on statistical models and machine learning approaches," to appear in Computer, Speech and Language.

Zheng-Hua Tan (M'00-SM'06) received the B.S. and M.S. degrees in electrical engineering from Hunan University, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 1999.

He is an Associate Professor in the Department of Electronic Systems at Aalborg University (AAU), Denmark, which he joined in May 2001. Prior to that, he was a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology (KAIST), Korea. He was also an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, China.

His research interests include speech recognition, noise robust speech processing, multimedia signal and information processing, multimodal human-computer interaction, and machine learning. He has published extensively in these areas in refereed journals and conference proceedings. He edited the book *Automatic Speech Recognition on Mobile Devices and over Communication Networks* (Springer-Verlag, 2008). He serves as an Editorial Board Member for *Elsevier Computer Speech and Language*, and the *International Journal of Data Mining, Modelling and Management*. He has served/serves as a program co-chair, session chair, tutorial speaker and organising committee member in major international conferences.

Børge Lindberg (M'94) was born in Denmark, November 1959. In 1983 he received the MSc degree in electrical engineering from Aalborg University (AAU), Denmark. From 1983 he was a research assistant at AAU, from 1986 he was at Jydsk Telefon, R & D Lab, Århus, Denmark and from 1992 he was a research assistant at AAU (from 1993 at the Center for PersonKomunikation (CPK)). In 1995 he was a visiting researcher at the Defence Research Agency, Great Malvern, UK, studying methods for predicting the performance of automatic speech recognition systems. Since 1996 he has been an Associate Professor at the Department of Electronic Systems, AAU, and since 2006 he has been the Head of this department. A

large part of his automatic speech recognition research has been done in collaboration with Danish and European companies on the basis of public funding. From 2001 to 2006 he was the chairman of the EU COST Action 278 on Spoken Language Interaction in Telecommunication. His current research interests include speech recognition with a focus on robustness and acoustic modeling techniques.