Modified LTSE-VAD algorithm for applications requiring reduced silence frame misclassification

Iker Luengo, Eva Navas, Igor Odriozola, Ibon Saratxaga, Inmaculada Hernaez, Iñaki Sainz, Daniel Erro

University of the Basque Country
Alda. Urquijo s/n 48013 Bilbao (SPAIN)
{iker.luengo, eva.navas, igor.odriozola, ibon.saratxaga, inma.hernaez, inaki.sainz, daniel.erro}@ehu.es

Abstract

The LTSE-VAD is one of the best known algorithms for voice activity detection. In this paper we present a modified version of this algorithm, that makes the VAD decision not taking into account account the estimated background noise level, but the signal to noise ratio (SNR). This makes the algorithm robust not only to noise level changes, but also to signal level changes. We compare the modified algorithm with the original one, and with three other standard VAD systems. The results show that the modified version gets the lowest silence misclassification rate, while maintaining a reasonably low speech misclassification rate. As a result, this algorithm is more suitable for identification tasks, such as speaker or emotion recognition, where silence misclassification can be very harmful. A series of automatic emotion identification experiments are also carried out, proving that the modified version of the algorithm helps increasing the correct emotion classification rate.

1. Introduction

One of the first steps in many speech processing systems is usually the voice activity detection (VAD), which identifies the time intervals with speech signal and those with only background noise. The outcome from the VAD algorithm allows discarding the silent frames from further processing, as they carry no speech information, therefore improving the performance of the rest of the system.

The quality of a VAD system is measured in terms of silence frames detected as speech or *silence misclassification rate* (ER0) and speech frames detected as silence or *speech misclassification rate* (ER1). Given a certain algorithm, both measures are somehow related, and reducing one of them results in the increase of the other. Therefore, it is necessary to reach a compromise with acceptable values in both error rates.

VAD algorithms are usually designed to have a very low ER1, at the cost of a higher ER0. In other words, they are designed to misclassify as few speech frames as possible, even though that means that a considerably large number of silence frames enter into the system. However, this behaviour is not always the most desirable, and the optimum working point depends on the application the VAD is used for.

In applications in which preserving the linguistic message of the speech is essential, a very low ER1 is required, since classifying speech frames as silence would mean to loose part of the message. This is the case of speech coding and automatic speech recognition (ASR) systems. Speech coding algorithms can use the VAD outcome to efficiently encode silent frames and to reduce the required bandwidth, as it is done in the EVRC (3GPP2, 2004), G.729 (ITU-T, 2007) and GSM (ETSI, 1997) systems. In ASR, the VAD is used to discard the silent frames from further processing, as they could confuse the recogniser and increase the error rates. For example, the *advanced front-end for distributed speech recognition* (AFE-DSR) defined by (ETSI, 2003) uses a VAD this way.

However, there are some applications that rely on the global acoustic characteristics of the speech, and for which the uttered message is not that important. Speaker, gender and emotion recognition are examples of this kind of applications. In these systems, silence frames are discarded, and all the rest are usually gathered in order to estimate a speaker, gender or emotion model, assuming that all speech frames come from the same distribution. Silent frames classified as speech corrupt the estimated distribution of the features, providing unreliable and weak models. Therefore, low ER0 is required. However, loosing some speech frames is not critical. Furthermore, speech frames classified as silence probably have a very low energy or are corrupted by noise, so discarding them and retaining the more robust frames can be beneficial (Krishnakumar et al., 2003). Therefore, in these applications a moderately higher ER1 can be bearable or even preferable.

Aside the already mentioned standard VAD implementations, a number of general-purpose algorithms have been proposed. Among them the LTSE-VAD presented by (Ramirez et al., 2004) stands out because of its simplicity, adaptability and good results. It obtains a very low ER1 even in noisy signals while maintaining a rather acceptable low ER0 in comparison with other algorithms. Nevertheless the achieved ER0 may still be too high for some applications.

In this work we propose a modification of the original LTSE-VAD algorithm that leads to a better performance when a low ER0 is required. The modified algorithm is compared to the original one and to other standard algorithms in terms of speech and silence frame errors (ER0 and ER1) as well as their detection error trade-off (DET) curves (Martin et al., 1997). Finally an experiment on emotion identification is carried out with both the original and modified algorithms showing a better performance with the proposed system.

2. Original LTSE-VAD algorithm

In broad lines, the original LTSE-VAD algorithm presented by (Ramirez et al., 2004) computes the divergence between the long term spectral envelope (LTSE) of the current frame and the mean spectrum of the noise in order to decide whether the frame contains speech or not. A divergence larger than a given threshold means that the spectral characteristics of the frame and the noise are different enough to classify it as speech. Otherwise, it is classified as silence. When the noise level is low, the speech and silence parts are easily distinguishable, but with a noisy signal the difference is not so clear. Therefore the decision threshold is adaptive and depends on the noise level, so that the algorithm gets good results with different noise levels.

Let $s[n]$ be the input signal, which is windowed with fixed-length overlapping windows, resulting in L frames $x(l)$ with $l = 1 \dots L$. Let $X(k,l)$ be the spectrum amplitude for frequency bin k in frame l, estimated with a discrete Fourier transform (DFT), with $k = 1 \dots K$. The LTSE of order N for frame $x(l)$ is defined as:

$$LTSE(k,l) = \max_{-N \leq j \leq N} \{X(k, l+j)\} \tag{1}$$

The LTSD between the frame and the estimated noise spectrum $N(k)$ is defined as:

$$LTSD(l) = 10 \cdot \log_{10} \left(\frac{1}{K} \sum_{k=1}^{K} \frac{LTSE^2(k,l)}{N^2(k)} \right) \tag{2}$$

Estimation of the noise spectrum $N(k)$ and noise power P_N can be done during a short initialisation step averaging T frames without vocal activity (and thus with only background noise), if the initial silence is known to be long enough.

$$N(k) = \frac{1}{T} \sum_{l=1}^{T} X(k,l) \tag{3}$$

$$P_N = \frac{1}{KT} \sum_{l=1}^{T} \sum_{k=1}^{K} X^2(k,l) \tag{4}$$

The LTSD value of each frame is compared to a given threshold γ. If the LTSD is larger than this threshold, it is labelled as speech, otherwise it is labelled as silence. The capability of the system to detect the vocal activity correctly depends on the signal to noise ratio (SNR). With high SNR the acoustic characteristics of speech and background noise are clearly different, so the threshold γ can be set at a fairly large value. With lower SNR the power and spectrum of frames with and without vocal activity are more similar, and it is more difficult to distinguish them. That means that for low SNR γ should be small in order to make the algorithm more flexible.

Therefore, the value of the threshold γ is set according to the noise level P_N:

$$\gamma(l) = \begin{cases} \gamma^m & P_N(l) \leq P_N^m \\ \gamma^M & P_N(l) \geq P_N^M \\ \dfrac{P_N(l) - P_N^m}{P_N^m - P_N^M}(\gamma^m - \gamma^M) + \gamma^m & \text{other} \end{cases} \tag{5}$$

where $P_N(l)$ is the estimated noise power for frame l, and P_N^m and P_N^M are the minimum and maximum considered values for the noise power, while γ^m and γ^M are the predefined thresholds for these extreme noise values respectively. In order to make the algorithm adaptive to time-varying noise levels, the estimated noise spectrum $N(k)$ and noise power P_N are updated with each frame classified as non-speech using a factor α_N:

$$N(k,l) = \begin{cases} \alpha_N \cdot N(k, l-1) + (1 - \alpha_N) \cdot X(k,l) & \text{if silence} \\ N(k, l-1) & \text{if speech} \end{cases} \tag{6}$$

$$P_N(l) = \begin{cases} \alpha_N \cdot P_N(l-1) + (1 - \alpha_N) \cdot P_X(l) & \text{if silence} \\ P_N(l-1) & \text{if speech} \end{cases} \tag{7}$$

with $P_X(l)$ the power of frame l. The initial values $N(k,0)$ and $P_N(k,0)$ are obtained during the initialisation step using equations (3) and (4). This adaptive behaviour, together with the use of a noise-level dependent threshold, are the characteristics that give the algorithm its robustness and great accuracy.

Finally, a hangover mechanism is implemented in order to delay the speech to silence transitions during HO frames. This mechanism is turned off if the LTSD exceeds a given threshold $LTSD_0$, as this would mean that the difference is clear enough not to need the hangover at all.

3. Modified LTSE-VAD algorithm

Making the threshold depend only on the noise level means that the speech level is considered to be the same in all cases. The difference between the LTSE of the signal and the noise spectrum does not depend on the noise level itself but on the SNR: the noise level may increase, but if the speech level increases as well, the relation (and the LTSD) may remain the same. It is expected that modifying the original LTSE-VAD algorithm in order to make the threshold dependent on the SNR instead of the noise level will improve results under certain conditions.

Therefore, we propose to estimate the threshold value according the the SNR level instead:

$$\gamma(l) = \begin{cases} \gamma^m & SNR(l) \leq SNR^m \\ \gamma^M & SNR(l) \geq SNR^M \\ \dfrac{SNR(l) - SNR^m}{SNR^m - SNR^M}(\gamma^m - \gamma^M) + \gamma^m & \text{other} \end{cases} \tag{8}$$

with $SNR(l)$ the SNR value for frame l, and SNR^m and SNR^M the minimum and maximum considered SNR values.

Using the SNR as a parameter to define the proper value for γ involves having a mechanism to estimate the SNR value for each frame. In addition to equations (6) and (7) to estimate the noise spectrum and power for each frame in an adaptive process, we also estimate the speech level adaptively:

$$P_S(l) = \begin{cases} \alpha_S \cdot P_S(l-1) + (1 - \alpha_S) \cdot P_X(l) & \text{if speech} \\ P_S(l-1) & \text{if silence} \end{cases} \tag{9}$$

Finally the adapted SNR for frame l is calculated as:

$$SNR(l) = 10 \cdot \log_{10}(P_S(l)) - 10 \cdot \log_{10}(P_N(l)) \tag{10}$$

Also, the hangover mechanism is completely turned off, as delaying the speech to silence transition may increase the ER0 type of errors.

4. Accuracy experiments

In order to check the effect of the applied changes a series of accuracy experiments were carried out comparing the original and modified LTSE-VAD algorithms. At the same time, the performance of the VAD algorithms included in ITU G.729 (ITU-T, 2007) and ETSI AFE-DSR (ETSI, 2003) standards are evaluated and compared. The AFE-DSR standard uses two VAD algorithms, one for the noise-reduction system and another one for the frame-dropping mechanism. Both are evaluated.

For these experiments the Spanish SpeeCon database (Iskra et al., 2002) was used. This database contains more than 1000 recordings in different environments (car, office and public place). Each recording was done with four different microphones: a close-talk headset (channel *C0*), a lavalier (channel *C1*), a medium distance cardioid microphone (0.5-1 meter, channel *C2*) and a far distance omnidirectional microphone (channel *C3*). Each of these channels represents a different SNR, *C0* being the cleanest (around 20 dB) and *C3* the noisiest (0 dB) scenarios.

The signals in the database were recorded in raw format at 16 kHz sample rate and 16 bit per sample. All recordings were downsampled to 8 kHz prior to the experiments. The reference speech and silence labelling was performed manually.

4.1. VAD accuracy experiments

As a first experiment the considered algorithms were evaluated in terms of ER0 (silence frames detected as speech), ER1 (speech frames detected as silence) and TER (total error rate). For these experiments G.729 and AFE algorithms used their standard values, while the original LTSE algorithm used the values proposed by (Ramirez et al., 2004): $N = 12$, $\gamma^m = 6$, $P_N^m = 30$, $\gamma^M = 2.5$, $P_N^M = 50$, $\alpha_N = 0.95$, $LTSD_0 = 25$, $HO = 8$. The modified algorithm was implemented with: $N = 12$, $\gamma^m = 8$, $SNR^m = 5$, $\gamma^M = 15$, $SNR^M = 20$, $\alpha_N = 0.95$, $\alpha_S = 0.95$.

Table 1 shows the performance of these algorithms for each of the scenarios in the database. Among the standard systems G.729 is the one with highest error rates both in terms of ER0 and ER1. The AFE systems obtain a much better result in terms of ER1, especially AFE-FD. Since this algorithm is used to discard silence frames entering the ASR, it is very conservative with speech frames and is adjusted to loose as few speech samples as possible.

The original LTSE algorithm also performs significantly well in terms of ER1, with results comparable to those obtained by AFE-FD. But at the same time it manages to reduce the ER0 value, showing the benefits of the adaptive algorithm. Accordingly, the TER is also reduced with respect to the standard algorithms. Nevertheless, the ER0 is over 30% in all scenarios, which may be fatal in some applications.

The proposed changes provide the best results in terms of ER0, with values between 10% and 20%, depending on the noise level. At the same time the ER1 level is kept under 7% in all cases, which makes this algorithm suitable also for applications were ER0 type of errors is not critical but may have some importance. In fact, looking at the total error rate, the proposed algorithm gives the lowest errors, with more than 90% of the frames correctly classified in *C0*, *C1* and *C2* scenarios.

	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C0	3.63	0.03	0.62	0.05	0.78
C1	9.28	0.23	1.98	0.49	4.77
C2	18.19	0.48	4.83	0.53	6.75
C3	17.22	1.41	8.30	1.34	5.04

(a) Error rate in speech frames (ER1)

	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C0	56.06	63.88	58.23	38.57	15.23
C1	70.23	54.75	55.96	33.04	8.62
C2	59.54	52.10	38.10	38.82	10.25
C3	70.49	50.10	47.65	34.88	22.55

(b) Error rate in silence frames (ER0)

	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C0	28.98	30.49	28.11	18.68	7.77
C1	38.74	26.24	27.73	16.22	6.63
C2	38.16	25.09	20.69	19.02	8.44
C3	42.94	24.61	27.05	17.54	13.50

(c) Total error rate (TER)

Table 1: Comparison of ER0, ER1 and TER for the considered algorithms.

4.2. DET curves

The proposed algorithm gets a lower ER0 partly at the cost of increasing the ER1 value. This same effect could be obtained modifying the LTSD threshold to a higher value and changing the working point of the algorithm. The key is to reduce significantly ER0 while increasing ER1 only a little. To see if the proposed changes really improve the performance of the system at different working points, Figure 1 shows the DET curves of the original LTSE and the proposed algorithms. The working points corresponding to the values of Table 1 are also represented.

It can be seen that the convenience of the applied changes depends on the selected working point. For low ER1 (under 1% in clean speech and under 5% in noisy speech) the original algorithm obtains the best performance, with the lowest values of ER0. This result seems reasonable, as the algorithm was developed precisely to give good results at low ER1 values.

On the other hand, for low ER0 values (below 10%) the proposed algorithm provides better results, with a lower corresponding ER1. This means that the proposed algorithm is more suitable for systems in which low ER0 is needed.

4.3. Automatic emotion identification experiments

As stated before, having a low ER0 value is especially important in speaker or emotion identification systems. Silence frames provide no information about the speaker or about the emotion, and they only increase the confusion in the system. On the contrary, ER1 errors are not so harmful, at least if they are kept below a reasonable level. Furthermore, speech segments that are classified as silence are

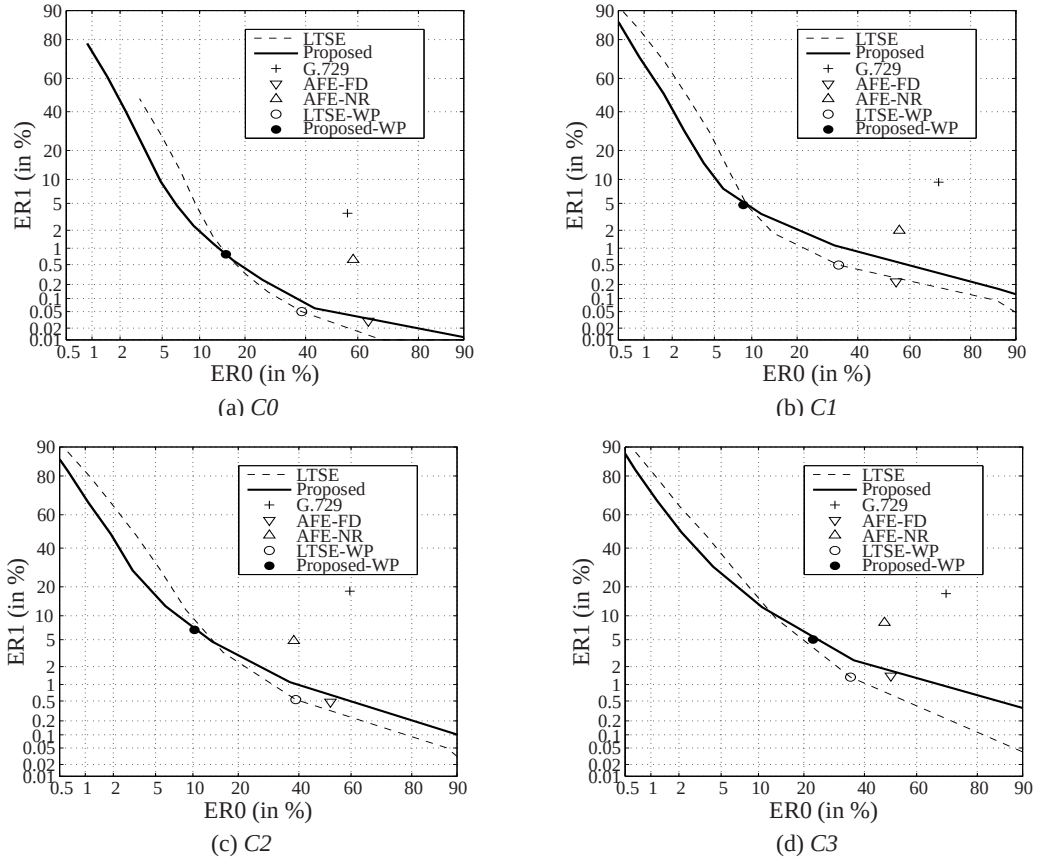

Figure 1: DET curves and working points of the considered algorithms for the four experimental scenarios.

likely to have low intensity or to be corrupted by the background noise. Therefore, discarding these frames can also make the identification system more robust. Of course, if the ER1 level increases too much, useful information will be lost.

In order to confirm whether the proposed changes are beneficial for this kind of systems or not, automatic emotion identification experiments were carried out using the *AIBO* emotional database (Batliner et al., 2006). This database contains approximately 18,000 recordings of 51 children while they were playing with the Sony-AIBO pet robot[1]. The data was collected in two different schools, *Mont* and *Ohm*, and the developers of the database suggest using the recordings from *Ohm* for the training and the ones from *Mont* for the testing phases. This way the speaker independence of the results is guaranteed.

According to the description given in (Batliner et al., 2006), five labellers assigned an emotional label to each *word* in the recordings. Afterwards, an heuristic algorithm was applied to these word-level annotations in order to obtain the final label for each recording. The database also provides a measure of the agreement among the labellers. Several recordings have an agreement below 50%, which means that less than 50% of the labels assigned to the words of those recordings agree with the final label estimated by the heuristic algorithm. We have considered that these sentences have uncertain emotional content, as not even humans reached to an agreement about the conveyed emotion.

[1]http://support.sony-europe.com/aibo/

	Anger	**Emphatic**	**Neutral**	**Positive**	**Total**
Train	424 (6,0%)	630 (9,0%)	5589 (79,4%)	398 (5,7%)	7041
Test	292 (4,8%)	371 (6,1%)	5377 (87,7%)	93 (1,5%)	6133

Table 2: Distribution of the recordings in the *AIBO* database, once items with uncertain emotion were discarded.

Therefore, we decided to discard them for this experiment. At the end we got 13,174 recordings, representing four different emotional states: anger, emphatic, neutral and positive. The recordings are distributed as shown in Table 2.

Although prosodic parametrizations have been traditionally used for the identification of emotions (Burkhardt and Sendlmeier, 2000; Banse and Scherer, 1996; Paeschke, 2004), some studies show that spectral information can also be useful for this task (Vlasenko et al., 2007; Kim et al., 2007; Casale et al., 2007). Hence, we decided to use both types of features and compare how the proposed changes affect to them. Therefore, two different parametrizations were defined:

- LFPC features (Nwe et al., 2003) and their first and second derivatives, as representative of spectral envelope information.
- Prosodic primitives, i.e., intonation and intensity curves, together with their first and second derivatives.

In both cases a new feature vector was extracted every

10 ms. As the value of F_0 is not defined for unvoiced frames, prosodic primitives were divided into two distinct streams, one for the voiced frames (parametrized with intonation and intensity features) and another one for the unvoiced frames (parametrized only with intensity features). For comparison purposes, LFPC features were also divided into voiced and unvoiced streams. 64 mixture Gaussian mixture models (GMM) (Paalanen et al., 2006) were used as classifiers in all cases.

As shown in Table 2, emotions are not balanced in the database, with most of the recordings being labelled as neutral. Therefore, the unweighted average recall (UAR) was used to measure the correct identification rate, instead of the more traditional weighted average recall (WAR) (i.e., accuracy).

$$UAR = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{N}\sum_{i=1}^{N}\frac{A_i}{M_i} \tag{11}$$

$$WAR = \frac{\sum_{i=1}^{N} A_i}{\sum_{j=1}^{N} M_j} = \sum_{i=1}^{N}\frac{A_i \cdot M_i}{M_i \sum_{j=1}^{N} M_j} = \sum_{i=1}^{N} P_i \cdot \Pr\{c=i\} \tag{12}$$

with N the number of emotions, P_i the accuracy for emotion i, M_i the number of test recordings for this emotion and A_i the number of test recordings for emotion i that are correctly classified. $\Pr\{c=i\} = \frac{M_i}{\sum_{j=1}^{N} M_j}$ represents the *a priori* probability that a test recording belongs to class i. Therefore, WAR is equivalent to the weighted average of the accuracies of each emotion, using the *a priori* probabilities of the emotions as weighting factor. The measure provided by the UAR is more meaningful when the test examples are unbalanced, as it takes into account the fact that the most represented class is more likely to get higher accuracy.

Table 3 shows the results of these experiments, both for the original LTSE algorithm and for the modified version. As it can be seen, the modified VAD algorithm helps increasing the correct classification rate in all cases, but its effect is more noticeable in the unvoiced streams. Obviously, silence frames are unvoiced, so when these frames are detected as part of speech, they change the feature distribution of the unvoiced stream, increasing the confusion in the classifier. The lower ER0 of our algorithm helps preventing this effect, providing a higher identification rate. The improvement in the emotion classification with voiced frames can be due to the increase in the ER1. In most cases, the speech frames that are discarded by the VAD algorithm have a very low intensity, and are probably corrupted by the background noise. Therefore, a moderate increase in the ER1 also helps making the models more robust.

	LFPC-V	**LFPC-UV**	**PP-V**	**PP-UV**
LTSE	57.7	50.2	49.4	42.1
Prop.	58.9	54.0	50.0	44.7

Table 3: Emotion identification results using the original and modified LTSE-VAD algorithms. Values represent UAR in percentage.

5. Conclusions

Usually, VAD algorithms are designed to have very low ER1 errors, i.e., to misclassify as few speech frames as possible, even though that means that as much as 30% of the silence frames are not detected. This design is the most suitable for applications in which the linguistic message should be kept at all costs (e.g., ASR). However, there are some applications in which the important information is on the acoustic characteristics, and not on the message. Speaker recognition or emotion identification are good examples of these. For this kind of applications, using silence frames during the modelling and identification steps results in an increase of the error rate. Therefore, achieving a low ER0 is necessary. Furthermore, a moderately higher ER1 may also improve the results, as the speech frames that will be detected as silence will probably have a very low intensity or will be corrupted by noise.

We have proposed some changes to the LTSE-VAD algorithm presented in (Ramirez et al., 2004), so that the VAD decision takes into account the SNR of the signal, and not only the background noise level. Silence and speech detection experiments have shown that the proposed algorithm achieves a lower ER0, while maintaining a reasonably low ER1 at the same time, even in noisy environments. Furthermore, taking into account both the speech and silence misclassification, the proposed algorithm gets the lowest TER levels among all the algorithms considered.

In order to confirm that the modified algorithm does indeed provide a better framework for emotion or speaker recognition tasks, a series of automatic emotion identification experiments have been conducted. Results show a significant increase in the correct classification rate when unvoiced frames are used, and a moderate increase when using voiced frames.

6. Acknowledgements

The work presented in this paper has been partially funded by the Spanish Government under grant TEC2009-14094-C04-02 (BUCEADOR project) and by the Basque Government under grant IE09-262 (BERBATEK project).

7. References

3GPP2. 2004. *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*.

Rainer Banse and Klaus R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Pathology*, 70(3):614–636.

Anton Batliner, Stefan Steidl, Bjrn Schuller, Dino Seppi, Kornel Laskowski, Thurid Vogt, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson. 2006. Combining efforts for improving automatic classification of emotional user states. In *Information Society - Language TechnologiesConference (IS-LTC)*, pages 240–245, Ljubljana (Slovenia).

Felix Burkhardt and Walter F. Sendlmeier. 2000. Verification of acoustical correlates of emotional speech using

formant-synthesis. In *ISCA Tutorial and Research Workshop on Speech and Emotion*, pages 151–156, Belfast.

Salvatore Casale, Alessandra Russo, and Salvatore Serano. 2007. Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*, 49(10):801–810.

ETSI. 1997. *ES 301 249: Digital cellular telecommunications system (Phase 2); Voice Activity Detector (VAD) for Enhanced Full Rate (EFR) speech traffic channels (GSM 06.82 version 4.0.1)*.

ETSI. 2003. *ES 202 050: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*.

Dorota Iskra, Beate Grosskopf, Krzysztof Marasek, Henk van del Heuvel, Frank Diehl, and Andreas Kiessling. 2002. SPEECON – speech databases for consumer devices: database specification and validation. In *Language Resources and Evaluation Conference (LREC)*, pages 329–333, Las Palmas, Spain.

ITU-T. 2007. *Recommendation G.729 Annex B: A silence compression scheme for G.729 optimized for terminals conforming to ITU-T Recommendation V.70*.

Samuel Kim, Panayiotis G. Georgiou, Sungbok Lee, and Shrikanth Narayanan. 2007. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *IEEE Workshop on Multimedia Signal Processing*, pages 48–51, Crete.

S. Krishnakumar, K.R. Prasanna Kumar, and N. Balakrishnan. 2003. Pitch maxima for robust speaker recognition. In *ICASSP*, volume 2, pages 201–204, Hong Kong.

Alvin F. Martin, George R. Doddington, Terri Kamm, Mark Ordowski, and Mark A. Przybocki. 1997. The DET curve in assessment of detection task performance. In *Eurospeech*, pages 1895–1898, Rhodes, Greece.

Tin Lay Nwe, Say Wei Foo, and Liyanage C. de Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech Communication*, 41(4):603–623.

Pekka Paalanen, Joni-Kristian Kamarainen, Jarmo Ilonen, and Heikki Klviinen. 2006. Feature representation and discrimination based on gaussian mixture model probability densities - practices and algorithms. *Pattern Recognition*, 39(7):1346–1358.

A. Paeschke. 2004. Global trend of fundamental frequency in emotional speech. In *Speech Prosody*, pages 671–674, Nara, Japan.

Javier Ramirez, Jose C. Segura, Carmen Benitez, Angel de la Torre, and Antonio Rubio. 2004. Efficient voice activity detection algorithms using long term speech information. *Speech Communication*, 42:271–287.

Bogdan Vlasenko, Bjrn Schuller, Andreas Wendemuth, and Gerhard Rigoll. 2007. Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. *Lecture Notes on Computer Science*, 4738:139–147.