Automatic Voice Activity Detection in Different Speech Applications

Marko Tuononen
University of Joensuu
+358 13 251 7963
mtuonon@cs.joensuu.fi

Rosa Gonzalez Hautamäki
University of Joensuu
P.O.Box 111 FI-80101 Joensuu, Finland
+358 13 251 7902
rgonza@cs.joensuu.fi

Pasi Fränti
University of Joensuu
+358 13 251 7931
franti@cs.joensuu.fi

ABSTRACT
This paper presents performance evaluation of voice activity detectors (VAD) by long-term spectral divergence and simple energy-based scheme. Evaluation is made in the terms of false accept (FA) and false reject (FR) errors using four different types of materials, recorded under different transfer channels, scenarios and conditions. Performance of VADs is considered for forensics, speaker recognition and interactive speech dialogue applications. Performance is still far from perfect, but despite the numerous classification errors of the methods tested, especially with noisy data, the methods can be still useful.

Categories and Subject Descriptors
I.5.4 [**Pattern Recognition**]: Applications – *signal processing, waveform analysis*, C.4 [**Performance of systems**]: Performance Attributes, and C.3 **[Special. Purpose and application-based systems**]: Signal Processing Systems.

General Terms
Measurement, Performance, Experimentation.

Keywords
Voice activity detection, speech applications, unsupervised learning, voice biometric, and speaker recognition.

1. INTRODUCTION
Voice activity detector (VAD) aims at classifying a given sound frame as a speech or non-speech as demonstrated in Fig. 1. It is often used as a front-end component in voice-based applications such as automatic speech recognition [1], speech enhancement [2] and voice biometrics [3]. A common property of these applications is that only human actions (typically only speech) are of interest to the system, and it should therefore be separated from other background sounds. Although VAD is widely needed and reasonably well-defined, existing solutions do not satisfy all the requirements. The methods either must be trained for the particular application and conditions, or they can be highly unreliable if the conditions change. Demands for working solutions, however, are frequently requested by practitioners.

In this paper, we study the performance of *long-term spectral divergence* [4], and a simple *energy*-based VAD [3]. We consider their applicability in three specific applications. The first one is a voice-based dialogue system [5] where the user interacts with the system by sending voice queries via a mobile phone, and the system answers using synthesized voice. The second application is voice biometric [3] where the goal is to verify whether a speaker is the one he claims to be. The third one is a forensic application where one searches from hours of audio material whether there is any speech or other human activity in the recordings.

All these applications have slightly different demands and different definitions of what should be considered as "voice". In voice-based dialogue system, only speech is desired as input whereas other human voices are considered as background noise. In forensic application, on the other hand, the detection of any human activity in long recordings might be of interest. The first case is sometimes referred as *speech activity detection* (SAD) [6] to differentiate the need to recognize only spoken natural language in contrast to other human-made voices. However, usually the term voice activity detection is used for both cases, and we will use this term throughout the rest of the paper for convenience.

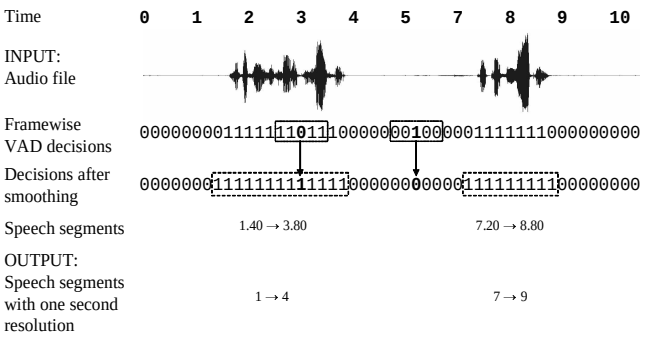

Figure. 1. Illustration of voice activity detection process.

E-FORENSICS 2008, January 21-23, Adelaide, Australia

DOI 10.4108/e-forensics.2008.2781

2. APPLICATIONS AND TEST MATERIAL

Next we describe our four test materials, two VAD methods, and three applications in question.

2.1 Applications

We consider three different applications. In forensic application, there are numerous recordings collected by eavesdropping, and automatic annotating would be needed to save manual work. It is important to find as many speech segments as possible, even at the cost of misclassifying some of the non-speech segments. Large variety of these recordings is a challenge for VAD. Environments, quality, and microphones differ a lot: from a hotel room so quiet that microphone can record speech from the neighboring room, to restaurant environment with so loud music that it is difficult to even understand what was said.

In voice biometrics, one wants to model the speaker only from those parts of a recording that contain speech. It is therefore important to use a conservative threshold to make sure that the frames used in the modeling actually do contain speech. If there is lack of speech material, a compromise might need to be taken between having enough training material, and not having too many non-speech frames included. It is also an open question whether frame level segments or longer smoothed segments should be used with speaker recognition system.

As an example of interactive voice-based dialogue system, we consider bus timetable system called Busman [7]. It provides bus route and timetable information for the city of Tampere, Finland. The system has both speech and "touch tone" interfaces, which may be used in a multimodal fashion. User can access the system using a standard mobile or landline phone. The purpose of voice activity detector in the system is to detect when the user is speaking, and to extract speech from the input. Efficient detection can improve performance especially in noisy environments [8].

2.2 Methods

We consider two approaches for the VAD:

1) a simple energy-based VAD as used in [3],
2) a method based on long-term spectral divergence [4].

The first approach measures the intra frame energy by calculating standard deviation of the frame and compares it to the expected SNR value (30 dB by default). A segment is defined as speech segment if it is within this marginal, and if it exceeds a given minimum energy level (-55 dB). This can be easily checked and it has found to work well for its purpose in voice biometric. The used thresholds have been optimised for NIST 2004 and NIST 2005 speaker recognition evaluation data. This can be problematic in applications where the background noise level varies.

The second approach measures long term spectral divergence (LTSD) between estimated speech spectra and noise spectra. It formulates the speech/non-speech decision rule by comparing the long-term spectral envelope to the average noise spectrum. It is also reasonably simple to implement and rather efficient compared to standard VAD methods according to the experiments made in [4]. The potential benefit over the energy-based VAD is that it is less dependent on the background noise level of the signal.

Other ideas based on spectral bandwidth, spectral crest factor, spectral tilt, zero crossing rate, modulation spectrum [9] and spectral entropy [10] were considered earlier in our laboratory but working solution was not found for the applications in question. Further investigations on these would be needed. Meanwhile, the only choice is to employ the best available tools and make the best use out of them.

The frame level classifications must be processed further because some of the applications need longer segments as output. We therefore smooth the frame-wise VAD decisions using median filtering and report the locations of the speech using one-second resolution as demonstrated in Figure 1. This is appropriate for the forensic and sbus-stop applications, although frame-wise resolution can still be used as such in the voice biometric application.

2.3 Materials

For testing, we will use the following materials: NIST 2005 data [11], Bus-stop data [5], our own laboratory recordings (prepared for this study), and NBI recordings (not publicly available). All tested files have 16-bit resolution and at least one second without speech at the beginning of the file, needed in the LTSD method. NIST-05 and Bus-Stop materials consist of phone conversations, and have therefore phone line quality. Lab and NBI materials are surveillance type of materials, and their technical quality varies between different recordings. Summary of the materials is presented in Table 1.

Table 1. Summary of the materials used.

Material	Files	Sampling rate	Duration / file	Total duration
NIST-05	15	8 kHz	5 minutes	01:14:45
Bus-Stop	94	8 kHz	1.5 – 9 minutes	03:08:13
Lab	1	44.1 kHz	over 4 hours	04:14:42
NBI	4	16 – 44.1 kHz	20 minutes – 2 hours	04:35:47

NIST-05 material is a subset of the 2005 NIST speaker recognition evaluation (SRE-05) corpus [11]. The material consists of conversational speech collected over the phone line. The level of background noise in this data set is small.

Bus-Stop material is a subset of the Bus-Stop timetable system [5] recordings. The dialogue system consists of both human speech and synthesized speech that provides bus schedule. The recordings contain a variety of background sounds, and user interface tones.

Lab material is a continuous recording from the lounge of our laboratory. Background noise in the recording is high and volume of the speech is very low most of the time.

NBI material is conversational forensic material from National Bureau of Investigation consisting of Finnish speakers. The files have been recorded under different conditions and places, and therefore, the volume of the speech varies a lot even within the same recording.

3. EXPERIMENTAL SETUP

We next define how to evaluate the quality of a given segmentation, and then study the parameter selection of the VAD methods.

3.1 Evaluating segmentation

For evaluating a given segmentation, we have formed correct segmentation (ground truth) manually, using one second resolution. While making this segmentation, the speech was defined to be human made sound if it contains a linguistic message. For example, sneezing is human-made sound but does not have any meaning, and therefore, it is not defined as speech. Furthermore, if volume or quality of the speech was so low that one cannot understand what was being said, it was classified as non-speech. This definition is relevant for the forensic application but might be slightly different in the other applications.

The quality of voice activity detection can be evaluated by two measures: *false acceptance* (FA) and *false rejection* (FR) rates:

$$FR = \frac{\text{Incorrectly classified speech segments}}{\text{Number of real speech segments}} \times 100 \cdot \quad (1)$$

$$FA = \frac{\text{Incorrectly classified nonspeech segments}}{\text{Number of real nonspeech segmets}} \times 100 \quad (2)$$

Equal error rate (EER) describes the performance of a system when FA and FR are equally important. It is a compromise between the two errors, and will therefore be used for a rough comparison of the performance between different parameter settings. To describe the performance across different FA and FR levels, we use the so-called *detection error trade-off* (DET) plots [12], which is a in principle *receiver operating characteristics* (ROC) plot in logarithmic scale. It is widely used in speech technology community to evaluate different detection systems, for example speaker verification systems.

3.2 VAD parameters

Sensitivity of the system is adjusted by changing the threshold of the speech/non-speech decision. A proper threshold can be selected manually by the user or automatically by the system. In automatic selection, the threshold can be trained on the basis of speech samples typical for the application. In this case, the material must be labelled beforehand and the acoustic and technical conditions in the training material should match to the ones to be used in the application. Alternative approach is to adapt the threshold to the data during the process. Selecting a threshold automatically is useful in batch processing tools, where user interaction should be minimized, or avoided completely.

In the case of manual threshold selection, a protocol is needed. The selection must be made interactively, but requiring minimum efforts from the user. One idea is to implement computer-guided selection that would provide the user with an initial segmentation. He can then analyze sample segments and give feedback on whether the result was correct or not. Alternatively, he could give overall comments either as “too many” or “too few” detections. The system can then adjust the threshold towards lower FA or lower FR, depending on the type of feedback. This kind of user-guided threshold would be necessary as long as perfect VAD method does not exist.

The LTSD based method uses a *calibration curve* to automatically set the correct threshold value for each recording individually. The threshold depends on the amount of noise energy observed from the beginning of the file. Automatic selection has been illustrated in Figure 2, where noise energy observed from the beginning of the recording is 50, and the corresponding threshold value is 10.

We consider three alternative parameter set-ups (Table 2). The original parameters are the ones recommended in [4]. The γ_0 and γ_1 refer to thresholds of the energy levels e_0 and e_1, example of which can be found in Figure 2. HO on the other hand, is the *hang-over threshold* and HO length is its corresponding window length. Parameters A were trained for our recording conditions. Parameters B were trained for NIST-05 and Bus-Stop, because the other two parameter set ups were too sensitive, and basically whole file was classified as speech. Training of the parameters was performed experimentally using a trial-and-error approach.

In all the tests, window size of 20 ms and window shift of 10 ms are used. In case of manual selection, threshold values 0–100 with step size 1 and *silence tolerance* values 1 (=filtering not in use), 50, 100, 200 and 400 are considered. Here, silence tolerance is the size of the window (in frames) for median filtering, see Figure 1.

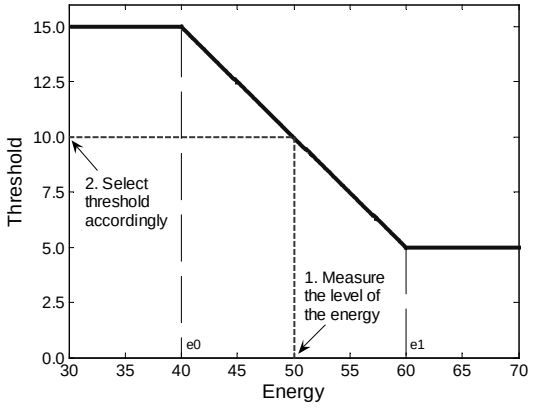

Figure. 2. Calibration curve and automatic selection of the threshold using the parameter set A in Table 2.

Table 2. Parameters for automatic thresholding with LTSD.

	Window length	γ_0	γ_1	e_0	e_1	HO	HO length
Original [4]	13	6	2.5	30	50	25	8
Parameters A	13	15	5	40	60	30	2
Parameters B	5	50	30	10	30	20	8

4. RESULTS

The main results are presented in Figure 3 using and the EER values are summarized in Table 3.

Comparison between the energy-based and LTSD methods is considered from three different view points: (1) FR and FA are equally important, (2) FR is more important, (3) FA is more important. The energy-based method works significantly better than the LTSD for NIST-05 material when we consider both the FR and FA errors equally costly. This is, because energy-based

method has been designed particularly for NIST materials. For the other materials the methods perform equally well.

When minimization of FR is considered more important, LTSD is better on Lab material. Otherwise, there are only minor differences. When FA is considered more important, the energy-based method has better performance overall, since it is better on NBI, Lab and NIST-05 materials.

We can clearly see that one should use energy-based method for NIST-05 material. With other materials the methods are equally good if both errors considered equally important. Results with the NBI material are worse than with other materials because the NBI recordings differ a lot from each other.

Benefit of the silence tolerance is arguable, because in some case (e.g. Bus-Stop) it can improve the result and in some cases (e.g. NBI) it can degrade it. In general, the LTSD method seems to benefit more with the use of silence tolerance parameter.

Automatic parameters of the LTSD seem to work only for the materials, which they were trained for. Parameters A work for Lab and NBI materials, whereas parameters B work for NIST-05 and Bus-Stop materials. The original parameters do not work with any of the materials, since we have used different materials than in [4].

4.1 Performance of the LTSD method

We have selected the most successful automatic parameters for each material, because this is the way the system would be use in real application. An alternative way to analyze the performance would be to use the threshold values at the EER point having equal proportion of false rejections and false acceptances.

The Bus-Stop material consisted of synthesized and human speech. We noticed that synthesized speech was detected significantly more reliably (FR=15%) than human speech (FR=71%). Backgrounds of the missed segments and the false alarms on Bus-Stop material are summarized in Figure 4. In the figure, “mixture of many sounds” refers to non-speech sounds observed outdoors that do not fit to any of the above categories.

On NIST-05 material missed segments consisted mostly to hesitation sounds and interjections. Although NIST-05 material is recorded under relatively clean conditions, those recordings are surprisingly hard for the LTSD method. This observation suggests that technical factors, like the quality of recording, have a significant effect.

The types of the falsely accepted segments for Lab and NBI recordings are summarized in Figure 5. We can see that LTSD method is mostly based on energy, since typically high energy noise was misclassified as a speech: door bangs and kitchen sounds in Lab recording, and all kind of bang and clatter sounds in general.

4.2 Applicability of VAD in speaker recognition system

Speaker verification tests are performed for NIST 2001, and NIST 2006 corpora [11]. Energy-based VAD is used for the tests, because according to the results on Figure 3 it is far more applicable for NIST-material.

We use 12 normalized MFCC features with their delta and double-delta coefficients. An *universal background model* (UBM) is trained from the development set of NIST 2001 corpus, and speaker models are created using GMM-UBM model of size 64. From Table 4, we see that VAD significantly improves speaker verification accuracy. In the case of NIST 2006 enhancement to the verification accuracy is enormous: from coin tossing to reasonable level.

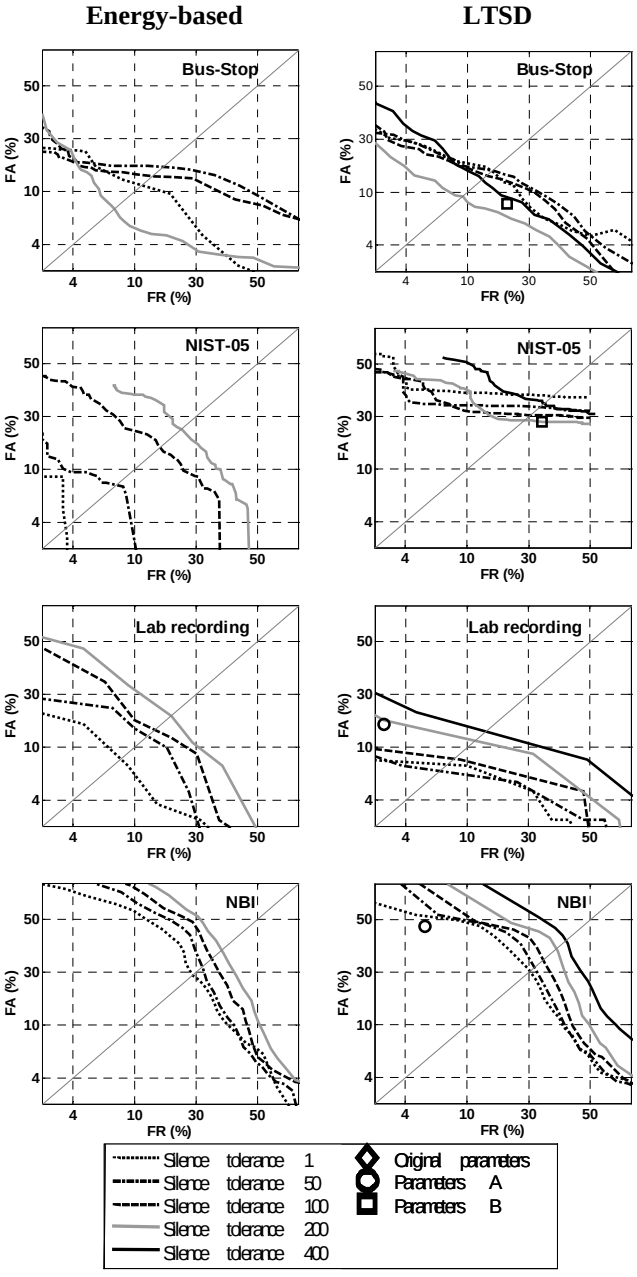

Figure. 3. Results of the energy-based (left) and the LTSD method (right) for voice activity detection.

Table 3. Equal Error Rates for each value of the silence tolerance.

		Bus-Stop	NIST-05	Lab	NBI
Energy	1	15 %	4 %	11 %	30 %
	50	21 %	10 %	17 %	33 %
	100	18 %	20 %	19 %	37 %
	200	10 %	26 %	23 %	40 %
LTSD	1	19 %	39 %	10 %	31 %
	50	20 %	35 %	9 %	33 %
	100	18 %	31 %	11 %	36 %
	200	13 %	29 %	17 %	41 %
	400	17 %	36 %	20 %	45 %

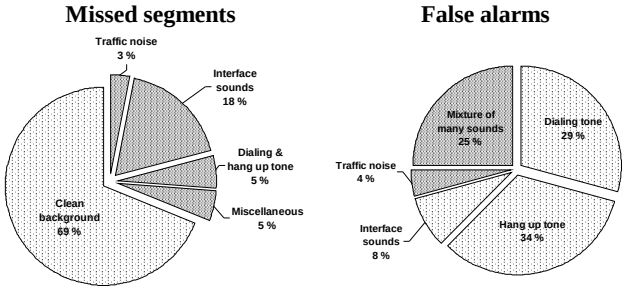

Figure. 4. Type of background in the missed speech segments (left), and segments that has been misclassified as a speech (right) on Bus-Stop recordings.

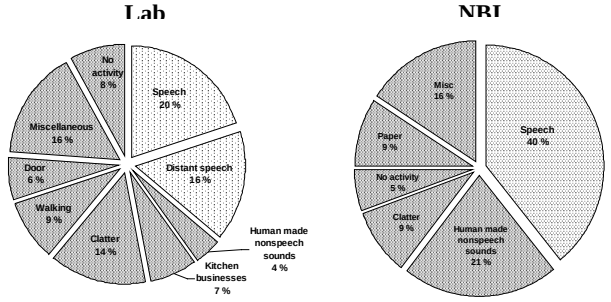

Figure. 5. Segments that have been classified as a speech on Lab (left) and on NBI (right) material.

Table 4. Effect of the energy-based VAD to the speaker verification performance (EER).

	VAD	No VAD
NIST 2001 evalset	15 %	23 %
NIST 2006 1conv-1conv	20 %	46 %

4.3 Applicability of VAD in forensics application

In the forensics application, it is important to find as many speech segments as possible, even at the cost of misclassifying some of the non-speech segments. We have therefore set an acceptable level of missed speech segments and the corresponding false acceptance rates are reported in Table 5.

We see from Table 5 that the energy-based VAD provides better results than the LTSD based method at the low FR side. However, false acceptance rates are still too high for the method to be really applicable here. For example, in the case of recording 3, if we accept 1 % of missed speech segment, we would still get 80 % of the non-speech segments!

Table 5. False Acceptance rates, when only few speech segments are missed. Silence tolerance value is 1.

Missed speech segments (FR)		5 %	1 %	0.2 %
Energy	File 1	17 %	27 %	28 %
	File 2	44 %	75 %	82 %
	File 3	48 %	80 %	83 %
	File 4	12 %	28 %	44 %
LTSD	File 1	23 %	30 %	32 %
	File 2	50 %	81 %	83 %
	File 3	52 %	90 %	93 %
	File 4	15 %	67 %	93 %

5. Summary

We considered an evaluation of the performance with four different types of material and with two different VAD methods. Materials were recorded using different transfer channels and in different scenarios and conditions. We evaluated performance in terms of FA/FR and DET-curves. Results are summarized in Table 6. The energy-based method is better with the NIST-05 data, most likely because it was tuned for earlier NIST corpora of similar quality. On the other materials the methods are equally good.

We found out that the LTSD method is mostly based on energy, since low volume speech was typically missed, and since high energy non-speech was typically misclassified as speech. VAD improves significantly the accuracy of the speaker verification system. Applicability in other applications could not been proven yet.

Table 6. Summary of the results. Best automatic parameters in EER sense are considered.

		Bus-Stop	NIST-05	Lab	NBI
Energy	EER	10 %	4 %	11 %	30 %
LTSD	FR	24 %	36 %	3 %	7 %
	FA	11 %	29 %	20 %	51 %
	EER	13 %	29 %	9 %	31 %

6. FUTURE WORK

Performance of voice activity detection methods is still far from perfect. The main problem is that the classification errors of the methods tested are relatively high with noisy data. In applications such as speaker verification, both methods are still useful and can provide better recognition accuracy. In forensic applications, on the other hand, automatic selection of the threshold is more problematic. It should be either trained on the type of material used, or a user-guided semi-automatic threshold selection should be provided.

Besides these, better VAD methods should still be developed. A way to make VAD more robust against high energy non-speech is to use in addition features, which are not so sensitive to energy peaks, for example the zero crossing rate [13] or the shape of the energy peak [14]. One experimental idea is to calculate fundamental frequency (F0) to find voiced frames, and use this information to make more reliable VAD decisions.

Because it seems to be that some training material is needed for learning proper parameters for our system, one should consider the possibility of using a statistical model-based VAD, e.g. [15]. One should also consider method to be used for training the parameters and fusing the features. Actually, current LTSD VAD is already model-based, since the parameters define the model. However, question is, if the model is extensive enough and the features are robust enough.

7. ACKNOWLEDGMENTS

The work was supported by the *National Technology Agency* of Finland (TEKES) as the projects *New Methods and Applications of Speech Technology* (PUMS).

8. REFERENCES

[1] F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, and H. Asoh, "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1727-1738, 2004.

[2] V. Gilg, C. Beaugeant, M. Schonle, and B. Andrassy, "Methodology for the design of a robust voice activity detector for speech enhancement," International Workshop on Acoustic Echo and Noise Control, Kyoto, Sept. 2003, pp. 131-134.

[3] R. Tong, B. Ma, K.A. Lee, C.H. You, D.L. Zhou, T. Kinnunen, H.W. Sun, M.H. Dong, E.S. Ching, and H.Z. Li, "Fusion of acoustic and tokenization features for speaker recognition," 5th International Symposium on Chinese Spoken Language Processing, Singapore, Dec. 2006, pp. 566-577.

[4] J. Ramírez, J.C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[5] M. Turunen, J. Hakulinen, and A. Kainulainen, "Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: similarities and differences," 10th European Conference on Speech Communication and Technology, Pennsylvania, Sept. 2006, pp. 1057-1060.

[6] J. Gruber, "A comparison of measured and calculated speech temporal parameters relevant to speech activity detection," *IEEE Transactions on Communications*, vol. 30, pp. 728-738, 1982.

[7] M. Turunen, J. Hakulinen, K.-J. Räihä, E.-P. Salonen, A. Kainulainen, and P. Prusi, "An architecture and applications for speech-based accessibility systems," *IBM Systems Journal*, vol. 44, pp. 485-504, 2005.

[8] L. Karray, A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environment," *Speech Communications*, vol. 40, pp. 261–276, 2003.

[9] L. Rabiner, and B.H. Juang, Fundamentals of speech recognition. Englewood Cliffs: Prentice Hall, 1993.

[10] P. Reneveyy, and A. Drygajloz, "Entropy based voice activity detection in very noisy conditions," 7th European Conference on Speech Communication and Technology, Aalborg, Sept. 2001. pp. 1887-1890.

[11] National Institute of Standards and Technology, NIST Speaker Recognition Evaluations. WWW-site, http://www.nist.gov/speech/tests/spk/ (27.11.2007).

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," 5th European Conference on Speech Communication and Technology, Sept. 1997, pp. 1895-1898.

[13] B.S. Atal, and L.R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 201–212, 1976.

[14] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 777-785, 1981.

[15] S. Basu, "A linked-HMM model for robust voicing and speech detection," IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, Apr. 2003, pp. 816-819.