Low-Complexity Voice Detector for Mobile Environments

Michal Ries, Bruno Gardlo, Markus Rupp
Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology
Gusshausstrasse. 25, A-1040 Vienna, Austria
Email: (mries, mrupp)@nt.tuwien.ac.at

Phillip De Leon
New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico USA 88003
Email: pdeleon@nmsu.edu

***Abstract*—Provisioning of mobile audio and video services is a difficult challenge since in the mobile environment, bandwidth and processing resources are limited. Audio content is normally present in most multimedia services, however, the user expectation of perceived audio quality differs for speech and non-speech content. Therefore, automatic voice or speech detection is needed in order to maximize perceived audio quality and reduce bandwidth and processing costs. The aim of this work is to find a low-complexity speech detector, suitable for detection of speech in a highly-compressed multimedia stream whose audio track may consist of speech, music, broadcast news, or other audio content. Finally, two methods for speech/non-speech detection are proposed and compared.**

I. Introduction

Massive provisioning of mobile multimedia services and higher expectations of end-user quality bring new challenges for service providers. One of the challenges is to improve the subjective quality of audio and audio-visual services. Due to advances in audio and video compression and wide-spread use of standard codecs such as AMR and AAC (audio) and MPEG-4/AVC (video), provisioning of audio-visual services is possible at low bit rates while preserving perceptual quality. The Universal Mobile Telecommunications System (UMTS) release 4 (implemented by the first UMTS network elements and terminals) provides a maximum data rate of 1920 kbps shared by all users in a cell and release 5 offers up to 14.4 Mbps in the downlink direction for High Speed Downlink Packet Access (HSDPA). The following audio and video codecs are supported for UMTS video services: for audio these include AMR speech codec, AAC Low Complexity (AAC-LC), AAC Long Term Prediction (AAC-LTP) [1] and for video these include H.263, MPEG-4 and MPEG-4/AVC [1]. The appropriate encoder settings for UMTS video services differ for various content and streaming application settings (resolution, frame and bit rate) [2].

End-user quality is influenced by a number of factors including mutual compensation effects between audio and video, content, encoding, and network settings as well as transmission conditions. Moreover, audio and video are not only mixed in the multimedia stream, but there is even a synergy of component media (audio and video) [3]. As previous work has shown, mutual compensation effects cause perceptual differences in video with a dominant voice in the audio track rather than in video with other types of audio [4]. Video content with a dominant voice include news, interviews, talk shows, etc. Finally, audio-visual quality estimation models tuned for video content with a dominant human voice perform better than a universal models [4]. Therefore, our focus within this work is on the design of automatic speech detection algorithms for the mobile environment.

In recent years, speech detection has been extensively studied [5], [6], [7], [8]. The proposed algorithms for speech detection differ in computational complexity, application environment, and accuracy. Our approach is to design a speech detection algorithm suitable for real-time implementation in the mobile environment. Therefore, our work is focussed on accurate and low complexity methods which are robust against audio compression artifacts.

Our proposed low-complexity algorithm is based on the kurtosis [9] and High Zero Crossing Rate Ratio (HZCRR) [10] extracted from the audio signal. The final speech or non-speech decision is based on hypothesis testing using a Log-Likelihood Ratio (LLR). The proposed method shows a good balance between accuracy and computational complexity. Furthermore, we have proposed a method based on Mel-Frequency Cepstral Coefficients (MFCCs) which provides significantly better accuracy but at the cost of increased computation. Finally, performance and complexity of these methods are compared.

The paper is organized as follows: In Section 2 we describe the objective parameters for speech detection. In Section 3 the design of speech detection algorithm is introduced. Performance evaluation of proposed algorithm and comparison with state-of-the-art methods are given in Section 4. In Section 5 we conclude the article and describe our future work.

II. Audio parameters

Due to the low complexity requirement of the algorithm, our investigation was initially focused on time-domain methods. Initial inspection of the various audio signals show significantly different characteristics in speech and non-speech signals (see Figures 1 and 2). Wide dynamic range of the speech signal (compared to non-speech signals) is clearly visible.

Both kurtosis and HZCRR features have been used in blind speech separation [12] and music information retrieval [10].

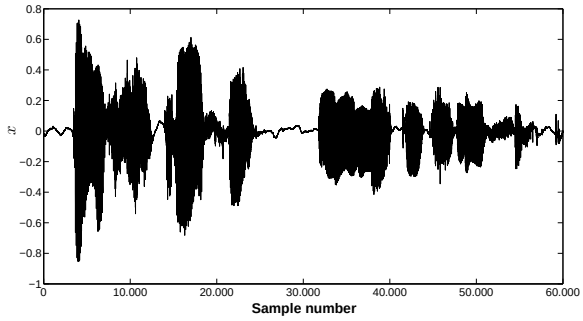

Fig. 1. Example of speech signal (time-domain).

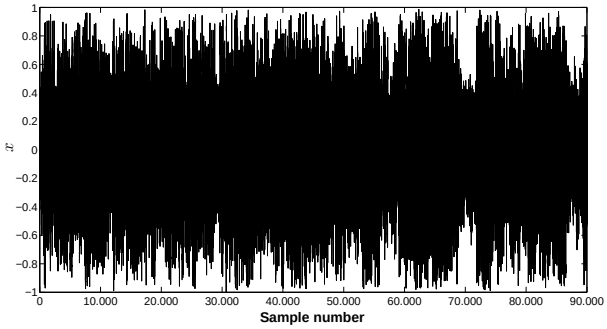

Fig. 2. Example of non-speech (time-domain).

Kurtosis of a zero-mean random process $x(n)$ is defined as the dimensionless, scale invariant quantity [1]

$$\kappa_x = \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^4}{\left(\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^2\right)^2}. \quad (1)$$

where in our case, $x(n)$ represents the n-th sample of an audio signal. A higher κ value is related to a more *peaked* distribution of samples as is found in speech signals (see Figure 3) whereas a lower value implies a flatter distribution as is found in other types of audio signals (see Figure 3). Therefore, kurtosis was selected as a basis for detection of speech. However, accurate detection of speech in short-time frames is not always possible by kurtosis alone.

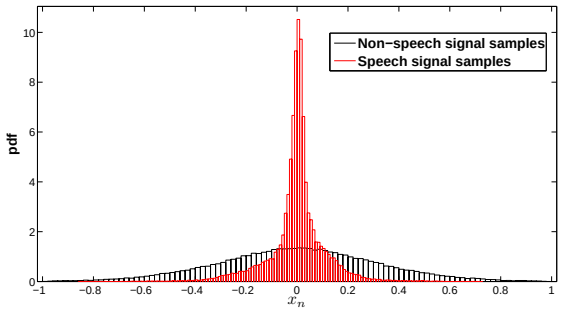

Fig. 3. Probability density function of the speech and non-speech audio samples

[1]The reader is cautioned that some texts define kurtosis as $\kappa_x = \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^4}{\left(\frac{1}{N}\sum_{n=1}^{N}(x(n)-\overline{x})^2\right)^2} - 3$ We shall however follow the definition in [9].

The second objective parameter under consideration is the HZCRR defined as the ratio of the number of frames whose Zero Crossing Rate (ZCR) is greater than $1.5\times$ the average ZCR in audio file as [10]

$$\mathrm{HZCRR}_M = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(\mathrm{ZCR}(n,M) - 1.5\overline{\mathrm{ZCR}}) + 1] \quad (2)$$

where $\mathrm{ZCR}(n,M)$ is the rate of the n-th, length-M frame (equation given below), N is the total number of frames, $\overline{\mathrm{ZCR}}$ is the average ZCR over the audio file. The ZCR is given by

$$\mathrm{ZCR}(n,M) = \frac{1}{M}\sum_{m=0}^{M-1}\mathbf{1}_{<0}\left[x(nM+m)x(nM+m+1)\right] \quad (3)$$

where m denotes the sample index within the frame and the indicator function is defined as

$$\mathbf{1}_{<0}(q) = \begin{cases} 1; & q<0 \\ 0; & q\geq 0. \end{cases}$$

According to our further experiences we use a frame length of 10 ms and the framing windows are overlapped by 50%. The 10 ms frame length [2] contains sufficient audio sample set for further statistical processing. Moreover, the longer framing window would increase the calculation complexity and length of investigated audio sequence nessesary for speech detection.

Figure 4 shows the ZCR curves for both speech and non-speech signals. The ZCR of the non-speech signal has a small amplitude range and low variance. The ZCR of the speech signal, on the other hand, has a wider amplitude range, large variance, and relatively low and stable baseline with occasional high peaks. However, many frames of the speech and non-speech signal have similar ZCRs and thus accurate detection of speech in short-time frames is also not possible with ZCR alone.

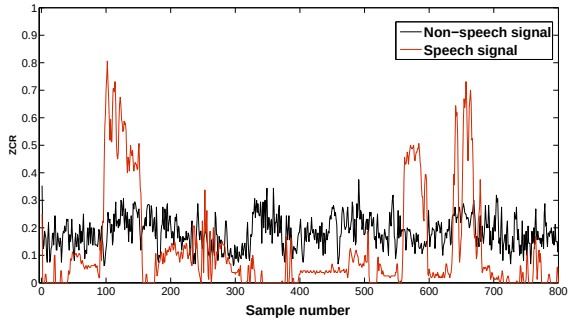

Fig. 4. Plot of the ZCR of the speech signal

A. Audio corpus

The training and evaluation of our speech detector was performed on a large audio corpus. Our corpus consists of 3032 speech and non-speech audio files (see details in Tables II and I). The speech part of corpus is in the German language

[2]e.g for SR = 32 kHz framing window contains M = 320 samples

and consists of ten speakers. The non-speech part of corpus consists of mainly music files of various genres (e.g. rock, pop, hip-hop, live music). All audio files were encoded using typical settings for the UMTS environment. Each audio file was encoded using three codec types at different sampling rates: AAC, AMR-WB at 16 kHz and AMR-NB at 8 kHz. Due to limitations of mobile radio resources, bit rates were selected in range 8–32 kbps. Encoded audio files with insufficient audio quality were excluded.

TABLE I
SPEECH AUDIO CORPUS

Codec	Encoding settings [BR@SR]	Number of audio files
AAC	16 kbps@16 kHz	1817
AMR-NB	7.9 kbps@8 kHz	1856
AMR-WB	12.65 kbps@16 kHz	1856

TABLE II
NON-SPEECH AUDIO CORPUS

Codec	Encoding settings [BR@SR]	Number of audio files
AAC	32 kbps@16 kHz	1169
AMR-NB	7.9 kbps@8 kHz	1172
AMR-WB	12.65 kbps@16 kHz	1176

For purposes of determining speech and non-speech detection parameters, 2273 audio without dominant voice and 3194 audio with dominant voice files were used in training. These files were selected from all codecs and encoding combinations. The rest of the audio corpus was used for performance evaluation.

Kurtosis and HZCRR$_M$ measurements on the training files is given in Figures 5 and 6. It can be seen that kurtosis is a better speech indicator than HZCRR$_M$, however, HZCRR$_M$ may be used as an additional indicator.

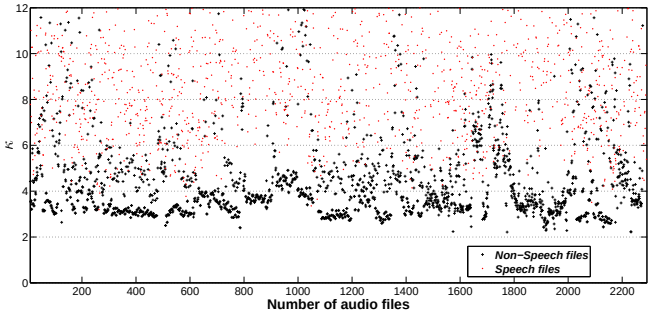

Fig. 5. Kurtosis values of speech and non-speech signals.

III. SPEECH DETECTOR

In order to reduce complexity, we propose a two-stage voice detection algorithm (see Figure 7). For the second stage (when the kurtosis is greater than the threshold), two solutions are proposed. The first has significantly lower complexity and is based on kurtosis and HZCRR$_M$ while the second solution, based on LLR, can provide higher accuracy but at the cost of higher complexity.

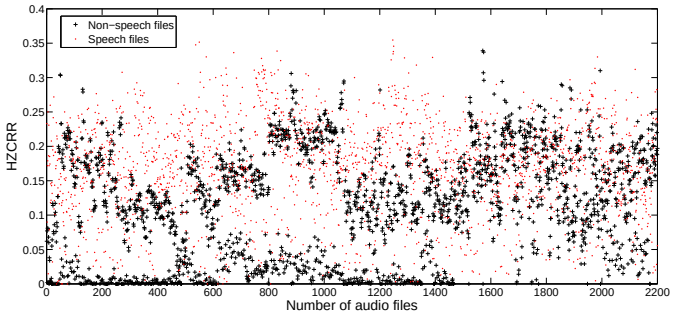

Fig. 6. HZCRR$_M$ values of speech and non-speech signals.

During the first stage, for the first solution based on κ and HZCRR$_M$, non-speech audio frames are detected by a simple decision based on whether the kurtosis is less than threshold (c_0) of 4.96 (see Figure 5). The first stage is capable of recognizing 62.3% of the non-speech frames from our corpus with a 97% accuracy rate.

Furthermore, for the MFCC based solution, the threshold (c_0) for the second stage was set at κ = 4 using the Least Absolute Errors optimization technique. All sequences with $\kappa \leq 4$ are recognized as non-speech sequences. By the first stage are recognized 40% of non-speech sequences from our corpus with 99.7%. precision.

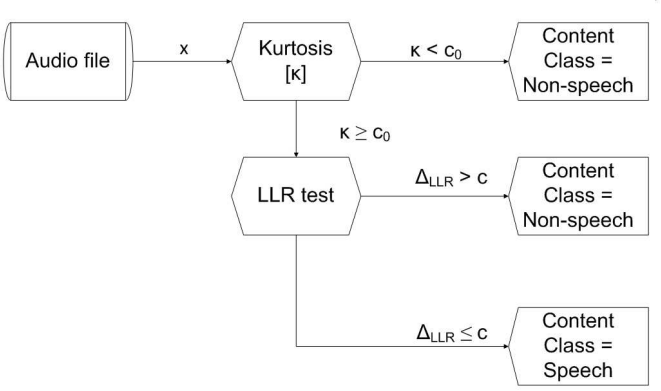

Fig. 7. Two-stage speech detector

A. Feature Vector based on κ and HZCRR$_M$

For the second stage (first solution), we derive a more general decision rule based on a hypothesis test (LLR) and we use both kurtosis and HZCRR$_M$ of the frame as elements in a feature vector

$$X = \begin{bmatrix} \kappa \\ \text{HZCRR} \end{bmatrix}.$$

For speech signals, we denote the mean vector for the speech feature vectors as μ_s and covariance matrix as Σ_s and for non-speech feature vectors, we denote the mean vector as μ_m and covariance matrix as Σ_m. Furthermore, the LLR test is performed on the first 20 frames, in order to reduce computational complexity. The log-likelihood ratio is calculated as

follows

$$\Delta = \frac{\sum_{i=1}^{20} \log\Big\{ \frac{1}{\sqrt{(2\pi)^2 \|\Sigma_s\|}} \exp(-\frac{1}{2}(X_i - \mu_s)\Sigma_s^{-1}(X_i - \mu_s)^T) \Big\}}{\sum_{i=1}^{20} \log\Big\{ \frac{1}{\sqrt{(2\pi)^2 \|\Sigma_m\|}} \exp(-\frac{1}{2}(X_i - \mu_m)\Sigma_m^{-1}(X_i - \mu_m)^T) \Big\}} \quad (4)$$

If the LLR is greater than the decision threshold, $c = c_1 = 2.2$ (see Figure 7), we declare a non-speech frame otherwise we declare a speech frame.

B. Feature Vector based on Mel-Frequency Cepstrum Coefficients

For the second stage (second solution), we consider the use of MFCCs extracted from the frame as the feature vector. MFCCs are widely used in speech and audio as a feature vector in a variety of applications. The algorithm in [13] is used for calculation of the first 14 MFCCs. Thus the covariance matrix is 14×14 and mean vector is 14×1. The LLR test is performed on the first 20 frames. The LLR is calculated as

$$\Delta = \frac{\sum_{i=1}^{20} \log\Big\{ \frac{1}{\sqrt{(2\pi)^{13} \|\Sigma_m\|}} \exp(-\frac{1}{2}(X_i - \mu_m)\Sigma_m^{-1}(X_i - \mu_m)^T) \Big\}}{\sum_{i=1}^{20} \log\Big\{ \frac{1}{\sqrt{(2\pi)^{13} \|\Sigma_s\|}} \exp(-\frac{1}{2}(X_i - \mu_s)\Sigma_s^{-1}(X_i - \mu_s)^T) \Big\}} \quad (5)$$

If the LLR is greater than the decision threshold, $c = c_2 = 1.04$ (see Figure 7), we declare a speech frame otherwise we declare a non-speech frame.

IV. Performance evaluation and comparison

We evaluate both two-stage algorithms: feature vector composed of kurtosis and HZCRR$_M$ and feature vector composed of MFCCs. The first algorithm is a relatively low-complexity solution based on time-domain audio parameters, κ and HZCRR$_M$. The second algorithm provides a more sophisticated solution based on MFCCs. The performance and complexity (measured in terms of computation time) of both methods was evaluated using 1770 speech files and 1181 non-speech files. The audio corpora for training and evaluation were approximately the same size. The overall accuracy of both proposed methods exceeds 92% (see Table III) for speech and non-speech content averaged over all codecs. The precision of second algorithm, however, clearly outperforms the first but at increased computation cost.

Content	Codec	κ & HZCRR	MFC
Non-speech	AAC	92.70 %	98.27 %
	AMR-NB	99.06 %	100 %
	AMR-WB	85.71 %	96.85 %
Speech	AAC	89.27 %	98.51 %
	AMR-NB	94.94 %	100 %
	AMR-WB	90.30 %	98.21 %
Overall		92.78 %	98.21 %

TABLE III

In order to evaluate complexity, the computation time was measured using 6091 audio files (3759 speech files, 2332 non-speech files). The algorithms were executed in MATLAB environment on a Core 2 Duo processor. In order to obtain the accurate results, the test was repeated ten times. Table IV gives the average computation times. The first algorithm is approximately $2\times$ faster then the second algorithm. The efficiency reflects the amount of processed files per second (see Table IV). The computing time and efficiency results show that both methods allow for fast detection of speech frames and are suitable for real time implementation in mobile devices.

Method	Time[s]	Efficiency [files/s]
κ & HZCRR	106.46	57.20
MFCC	233.89	26.04

TABLE IV
TIME NEEDED FOR CONTENT ESTIMATION

V. Conclusion

The goal of this work was to design a speech detector for mobile environment. The design was focused on accurate, low complexity methods, which are robust against audio compression artifacts. Both proposed algorithms show very good accuracy (92%) and relatively low complexity. However, the method based on kurtosis and HZCRR$_M$ is $2\times$ faster (lower complexity).

VI. Acknowledgement

The authors would like to thank mobilkom austria AG for supporting their research. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG.

References

[1] 3GPP TS 26.234 V6.13.0:"Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs," Mar. 2008.
[2] M. Ries, "Video Quality Estimation for Mobile Video Streaming," Doctoral thesis, INTHFT, Vienna University of Technology, Vienna, Austria, Oct. 2008.
[3] S. Tasaka, Y. Ishibashi, "Mutually Compensatory Property of Multimedia QoS," in Proc. of IEEE International Conference on Communications 2002, vol. 2, pp. 1105–1111, NY, USA, 2002.
[4] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, M. Rupp,"Audivisual Quality Estimation for Mobile Streaming Services," in Proc. of 2nd Int. Symp. on Wireless Communications (ISWCS), pp. 173–177, Siena, Italy, Sep. 2005.
[5] D. Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amador, X. Menendez-Pidal, "A robust speech detection algorithm for speech activated hands-free applications," in Proc. of ICASSP'99, pp. 2407–2410, Mar. 1999.
[6] J. Junqua, C. B. Mak, B. Reaves, "A Robust Algorith for Word Boundary Detection in the Presence of Noise," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 3, Jul. 1994.
[7] L. Mauuary, J. Monne, "Speech/non-Speech Detection for speech Responses Systems," in Proc. of Eurospeech93, Berlin, pp. 1097-1 100, September 1993.
[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, Apr. 1990.
[9] M. G. Bulmer, Principles of statistics, New York: Dover Publications, 1967.
[10] L. Lu, H. Jiang, and H. J. Zhang, "Content analysis for audio classification and segmentation," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 7, Oct. 2002.
[11] C. H. Chen, Signal processing handbook, New York: Dekker, 1988.
[12] P. De Leon, "Short-Time Kurtosis of Speech Signals with Application to Co-Channel Speech Separation," in Proc. IEEE Int. Conf. Multimedia and Expo, NY, USA, 2000.
[13] D. P. W. Ellis, (2005) PLP and MFCC in Matlab, Accessed on 10th December 2008. [Online]. Available: http: http://www.ee.columbia.edu/∼dpwe/resources/matlab/rastamat/melfcc.m