ACCURACY OF MP3 SPEECH RECOGNITION UNDER REAL-WORD CONDITIONS
Experimental Study

Petr Pollak and Martin Behunek
Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague, Czech Republic

Keywords: Speech recognition, MPEG compression, MP3, Noise robustness, Channel distortion.

Abstract: This paper presents the study of speech recognition accuracy with respect to different levels of MP3 compression. Special attention is focused on the processing of speech signals with different quality, i.e. with different level of background noise and channel distortion. The work was motivated by possible usage of ASR for off-line automatic transcription of audio recordings collected by standard wide-spread MP3 devices. The realized experiments have proved that although MP3 format is not optimal for speech compression it does not distort speech significantly especially for high or moderate bit rates and high quality of source data. The accuracy of connected digits ASR decreased consequently very slowly up to the bit rate 24 kbps. For the best case of PLP parameterization in close-talk channel just 3% decrease of recognition accuracy was observed while the size of the compressed file was approximately 10% of the original size. All results were slightly worse under presence of additive background noise and channel distortion in a signal but achieved accuracy was also acceptable in this case especially for PLP features.

1 INTRODUCTION

Automated speech recognition (ASR) represents a field which is more present in everyday human life in growing number of applications as in voice operated control of consumer devices, automated information services, or general conversion of uttered speech into text record. The systems for automatic transcription of speech to text are currently well developed for all important world languages. It is possible to meet today dictation software for standard PC enabling users to input texts into documents without using the keyboard, e.g. Dragon dictate, or for mobile devices, e.g. (Nouza et al., 2009). Further, the transcription of broadcast news is currently a very important task solved by many research teams (Chen et al., 2002), (Gauvain et al., 2002). Probably the most popular is the transcription of news, but there are also other applications such as automated subtitling of TV programmes, e.g. parliament meetings (Vaněk and Psutka, 2010) or sportscasts (Psutka et al., 2003). Special attention is also paid to the transcription and indexing of large audio archives enabling better search within them in the future (Makhoul et al., 2000), (Byrne et al., 2004).

When audio records are transcribed on-line, e.g. the above-mentioned subtitling of TV programmes, ASR systems work with full quality input signal. On the other hand, when they work off-line, recordings can be saved in formats of different quality and typically, MP3 format (more precisely MPEG Layer III) represents one of the most frequently used formats for the saving of sound files in compressed form (Bouvigne, 2007), (Brandenburg and Popp, 2000). It is well known that this format uses psychoacoustic models reducing the precision of components less audible to human hearing so it makes it possible to decrease the size of the sound file to 10% while CD quality is preserved. Although this format has been developed especially for saving the music, it is standardly used also in simple dictation devices or mobile phones enabling recording and saving audio speech files. Some works in MP3 speech data recognition have been already done. The recognition of spontaneous speech from large oral history archives published in (Byrne et al., 2004) used signals saved in MP3 format but rather high bit-rate (128 kbps) was used in this case. In (Barras et al., 2001) the study of automatic transcription of compressed broadcast audio with different bit-rates was done. The comparison of various compression schemes was realized in this study, however, the quality of the signal was rather better.

Pollak P. and Behunek M..
ACCURACY OF MP3 SPEECH RECOGNITION UNDER REAL-WORD CONDITIONS - Experimental Study.
DOI: 10.5220/0003512600050010
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2011), pages 5-10
ISBN: 978-989-8425-72-0

This paper presents the results of an experimental study analyzing ASR accuracy working with respect to different quality of compressed data as current ASR systems need to work accurately under real conditions and often under presence of additive background noise or channel distortion. It depends mainly on the quality and the position of the microphone used. The task of solving robust recognition with low error rate of possibly distorted speech from real environment is thus also a blossoming research field. Standard features used most commonly in ASR systems have typically modified versions increasing their robustness in real environment (Fousek and Pollák, 2003), (Bořil et al., 2006), (Rajnoha and Pollák, 2011) but these methods are designed usually for uncompressed data. That is the main reason why our study focuses mainly on the analysis of the information lost in compressed speech signals when varying levels of additive noise and channel distortion appear in speech signal. The results of this study are supposed to be helpful for further application of automated speech recognition from MP3 speech files.

2 MP3 SPEECH RECOGNITION

We use rather small vocabulary ASR for the purpose of this study. Of course, such small vocabulary system is much simpler than complex large vocabulary continuous speech recognition (LVCSR) system which is supposed to be used in a target application. On the other hand, we want to analyze mainly the sensitivity of ASR system to loss of information after MP3 compression without a dependency on further blocks such as complex statistic language model in LVCSR system and it is the main reason for the choice of digit recognizer in experiments of this study.

2.1 Speech Compression Scheme

MP3 compression was developed for the compression of music audio files (Brandenburg and Popp, 2000), (Bouvigne, 2007) and it is known that this compression gives slightly worse results for speech signal. The masking and attenuation of some frequency components can yield to a suppression of a phone at the beginning or at the end of the word, sometimes interword pause shortening can appear. Less naturalness of decoded utterance is then the main effect of this fact and consequently, the accuracy of speech recognition can decrease too. Algorithms, which have been designed and optimized for speech signal, are represented mainly by G.729 (ITU-T, 2007), AMR (ETSI, 2007), or Speex (Valin, 2007). These encoders are based typically on CELP algorithm, but they are used rather in telecommunications and they do not appear so frequently in standard audio devices.

Consequently, although speech signals can be compressed in a better way, our attention was paid just to MP3 compression in this study because the long-term goal of our work was mainly in off-line mode of ASR operation on compressed speech data from wide-spread audio consumer devices. The study was realized with signals from the database SPEECON recorded in real environment with full-precision PCM coding. The MP3 compression was then simulated by publicly available software LAME (Cheng and et. al., 2008) which made it possible to simulate also different levels of MP3 compression bit-rate.

2.2 Small Vocabulary ASR Setup

Current ASR systems are usually based on Hidden Markov Models (HMM). HMM based recognizer consists typically of 3 principal function modules: feature extraction (parameterization), acoustic modelling, and decoding. Generally known principles of HMM based recognizer are not explain in this paper, as they are known or can be found in many sources, e.g. in (Huang et al., 2001), (Young and et al., 2009), and others. Only a brief description of our ASR setup (see block scheme in Fig 1), which is relevant for the next parts of this paper, is presented.

Concerning the parameterization module, two sets of features are most standardly used in ASR systems: mel-frequency cepstral coefficients (MFCC) and perceptually based linear predictive cepstral coefficients (PLP). All our experiments were carried out just with these two feature sets. Both of them use their own non-linear filter-bank which in both cases models the perceptual properties of human auditory system, for more detail see (Huang et al., 2001), (Young and et al., 2009), (Psutka et al., 2001), or (Hermansky, 1990). Exactly 12 cepstral coefficients plus logarithm of frame energy commonly with the 1st and 2nd derivatives formed feature vector in our experiments. MP3 compression was involved within our system as optional preprocessing module before standard feature extraction, see Fig 1.

Acoustic modelling was based on monophone HMM models, i.e. phones modelled without any context to neighbouring phones. Finally, HMM models of 44 Czech monophones with 3 emitting states were used as standard subword acoustic element modelling in ASR. As used phone models were context independent, their higher variability was modelled by 32 mixtures of Gaussian emitting function of HMM models and 3 streams were also used for modelling of static,

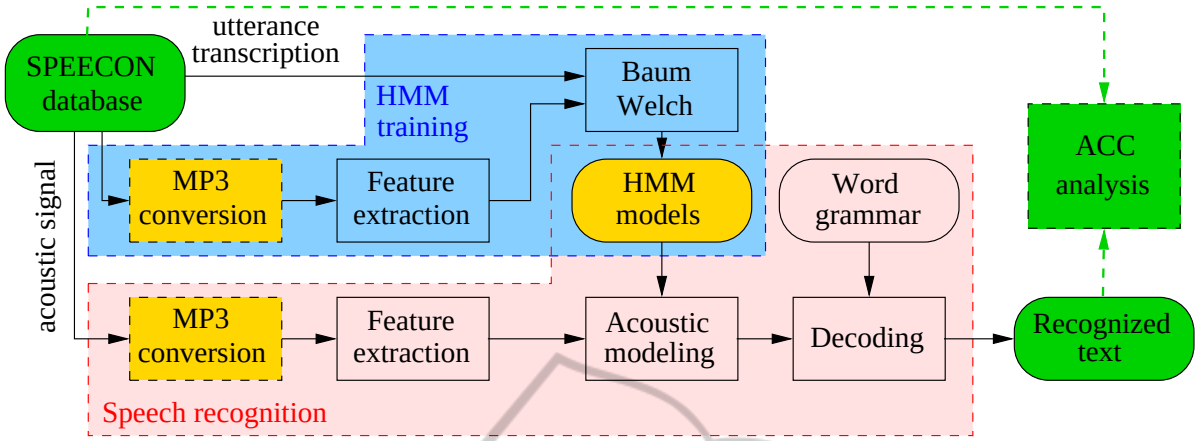

Figure 1: Block scheme of experimental setup.

dynamic and acceleration features respectively.

Connected digit ASR was used within this study so that it was possible to use simple grammar in the phase of decoding. On the other hand, though just basic digits from 0 to 9 can appear in the utterance, the number of digits in the utterance can vary and they can be pronounced with or without pauses and with possible repetitions. Finally, it means that our digit recognizer should be sufficiently general for our experiments and we can assume that it simulates well operating mode of target practical application.

2.3 Training of Speech Recognizer

Training of HMM models, which are composed of Gaussian emitting functions representing probability distribution of a feature in given state and from probabilities of transitions between particular states, is performed on the basis of iterative embedded training procedure from large training data (Baum-Welch algorithm). The size of the training database containing speech signals with precisely annotated content must guarantee sufficient appearance of each acoustic element. The training procedure, which has been based on Czech SPEECON data in our case, is illustratively represented by blue part of block scheme in Fig. 1, more details can be found in (Young and et al., 2009) or (Huang et al., 2001).

Our acoustic models with 32 Gaussian mixtures were trained iteratively in 23 steps with flat start for each parameterization and also for many operating conditions. As training data had to match these conditions, we have finally obtained comprehensive set of HMM models for particular channels, for different bit-rates in MP3 encoding, and for feature set used.

2.4 Implementation of ASR

Both small vocabulary ASR and the training of acoustic HMM models were realized by tools from publicly available HTK Toolkit (Young and et al., 2009) which is often used world-wide for the realization of HMM based recognition. For readers without detail knowledge of HTK Toolkit, typical and core tools of HTK Toolkit are *HCopy* as parameterization tool, *HERest* as the tool for the training by Baum-Welch algorithm, or *HVite* as Viterbi based word recognizer.

The computation of PLP cepstral coefficients was performed by *CtuCopy* tool (Fousek and Pollák, 2003) and (Fousek, 2006), providing some extensions of *HCopy* from the standard set of HTK Toolkit.

3 EXPERIMENTS AND RESULTS

Experiments described in this part comprise the core contribution of this study, which is mainly in the analysis of ASR performance for MP3 compressed speech data under different real-word conditions.

3.1 Speech Data Description

All experiments were carried out with speech signals from Adult Czech SPEECON database (ELRA, 2009). It is the database (DB) of sentences, digits, command, names, etc. recorded by 550 speakers under different conditions, i.e. in offices or home environment, at public places, or in the car. For this study, only well-balanced subset of adult speaker data from office environment were used, i.e. 90% of data for training and 10% for testing. It contains signals with similar and not so strong background noise.

Table 1: Description of channels recorded in SPEECON database.

Channel ID	*Microphone type*	*Distance*	*Additive noise*	*Channel distortion*
CS0	head-set	2 cm	-	-
CS1	hands-free	10 cm	+	-
CS2	middle-talk	0.5-1 m	++	+
CS3	far-talk	3 m	+++	++

Speech data in SPEECON DB are raw 16 bit linear PCM sound files sampled by 16 kHz sampling frequency. These signals were then encoded into MP3 sound files with different bit-rates. The MP3 compression was simulated by successive encoding and decoding by the above-mentioned *lame* software encoder/decoder (Cheng and et. al., 2008).

Although only data from one environment were used in our experiments, the influence of additive noise and channel distortion could be analyzed, because signals in SPEECON DB were recorded in 4 channels which differed in microphone type and its position, see (Pollák and Černocký, 2004). Following Tab. 1 describes the properties of particular channels. Although different types and quality of microphones were used, it was mainly the distance from the speaker's mouth that played the key role in the quality of recorded speech signal. Finally, signals from close talk head-set channel CS0 are then almost without additive noise and reasonable channel distortion appears only in signals from channels CS2 and CS3. These data can simulate well real MP3 recordings made by standard devices in various environments.

3.2 Results of Experiments

The accuracy (ACC) of the digit recognizer was measured standardly on the basis of errors on word level (Young and et al., 2009) and the following sections describe obtained results.

3.2.1 Analysis of Optimum Segmentation for MP3 Speech Data

Within the first experiment, the influence of short-time analysis setup on target accuracy of MP3 recognition was analyzed. In accordance with phonetically based assumptions as well as default settings used in (Young and et al., 2009), the optimum length of the frame for short-time acoustic analysis is 25 ms with the period of 10 ms for uncompressed speech data, while for MP3 compressed data the segmentation with frame length 32 ms and frame period 16 ms gives the best results for both studied feature sets.

The reasons for this effect lie in the first modules of both feature extraction algorithms which realize short-time Fourier analysis followed by non-linear filter banks computing perceptually based power spectra. Due to the decrease of short-time frame length, frequency resolution of Discrete Fourier Transform decreases too and consequently the masking and deletions of some frequency components within the MP3 compression scheme increase the estimation error of power spectrum at the output of the filter bank.

Table 2: ASR accuracy (ACC) dependence on varying segmentation for MFCC features and WAV or MP3 signals.

Features	WAV	MP3
MFCC_1608_hmm23	95.55	54.39
MFCC_2510_hmm23	96.89	76.31
MFCC_3216_hmm23	95.22	93.21

Table 3: ASR accuracy (ACC) dependence on varying segmentation for PLP features and WAV or MP3 signals.

Features	WAV	MP3
PLP_1608_hmm23	95.11	72.64
PLP_2510_hmm23	96.33	81.76
PLP_3216_hmm23	95.11	93.10

The results of this experiment for both MFCC and PLP features are in Tab 2 and 3. MP3 compression was realized with bit-rate 160 kbps and results are presented for the CS0 channel. It can be supposed that this error at the output of filter bank increases for shorter frame length also when uncompressed speech signal is more corrupted by additive noise, which is the case of channels CS1, CS2, and CS3.

Abbreviations used in the following tables describe the feature extraction used, e.g. MFCC_3216_hmm23 means MFCC features computed from 32 ms long frame with the period of 16 ms. The flag "hmm23" specifies HMM models obtained after 23rd training step.

3.2.2 Influence of MP3 Compression for Particular Channels

Within the second experiment, the influence of recognition accuracy in particular channels on different MP3 bit-rates was analyzed. All these experiments were realized with optimum segmentation parameters, i.e. 32 ms frame length and 16 ms frame period. Achieved results are presented numerically in the following tables and for quick illustrative overview also in figures showing the same data in graphical form.

Tab. 4 and Fig. 2 present results obtained with

Table 4: ASR accuracy (ACC) dependence on varying MP3 bit-rate for MFCC features computed for all channels.

MP3 bit-rate	CS0	CS1	CS2	CS3
WAV	95.22	92.44	89.54	61.18
160 kbps	93.21	83.31	42.83	30.03
64 kbps	93.21	84.43	43.60	31.59
48 kbps	93.33	85.54	43.83	32.26
40 kbps	93.33	86.43	43.94	31.59
32 kbps	89.77	88.54	44.75	33.48
24 kbps	89.32	37.71	38.71	27.92
8 kbps	21.02	16.35	12.57	6.90

Table 5: ASR accuracy (ACC) dependence on varying MP3 bit-rate for PLP features computed for all channels.

MP3 bit-rate	CS0	CS1	CS2	CS3
WAV	95.11	89.43	88.88	64.63
160 kbps	93.10	82.09	78.75	27.36
40 kbps	92.66	87.32	83.09	28.48
24 kbps	92.32	88.21	80.09	28.70
8 kbps	62.40	36.15	24.25	11.01

mel-frequency cepstral coefficients. Looking at the results achieved for CS0 channel we can see that for rather high quality signal the MP3 compression has just minimum effect up to bit-rate 24 kbps. For other channels the trend is always similar but the absolute values of the achieved ACC are lower according to our assumptions. We can also see that for channels CS2 and CS3 containing higher level of background noise and stronger channel distortion, the ACC falls rapidly already for rather high bit-rates of MP3 compression. Such results disable in principle the recognition of MP3 data collected under similar conditions. On the other hand it must be mentioned that all experiments in this study were carried out with basic feature setup, i.e. no algorithm for additive noise suppression or channel normalization was used.

Tab. 5 and Fig. 3 show similar results for the recognition with perceptually based linear predictive cepstral coefficients. The same trends have been observed again, so in the end the experiments were realized just with 4 different bit-rates. In comparison with MFCC, better performance can be observed for PLP for channels CS1 and CS2. Especially the results for channel CS2 represent acceptable values of ACC for MP3 compressed data up to the bit-rate of 24 kbps (80.09% as for MFCC it was 38.71%) which is similar as for high quality CS0 channel. In principle it allows the practical usage of ASR of MP3 compressed data collected by middle-distance microphone, e.g. it can be the case of MP3 recorder placed on the table.

Finally, we computed sizes of compressed data so we could compare the level of compression (in percent) with the achieved accuracy of speech recognition. These results are shown in Fig. 4 where we can observe minimum decrease of ASR accuracy up to 20% compression of sound file. Strong downfall appears as far as beyond 10% compression. These results were obtained for MFCC features and the high quality CS0 channel.

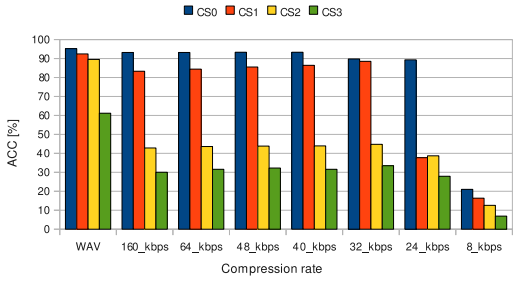

Figure 2: ASR accuracy (ACC) dependence on varying MP3 bit-rate for MFCC features computed for all channels.

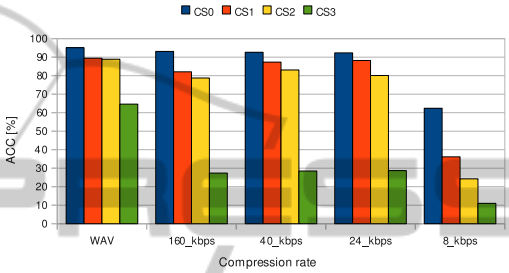

Figure 3: ASR accuracy (ACC) dependence on varying MP3 bit-rate for PLP features computed for all channels.

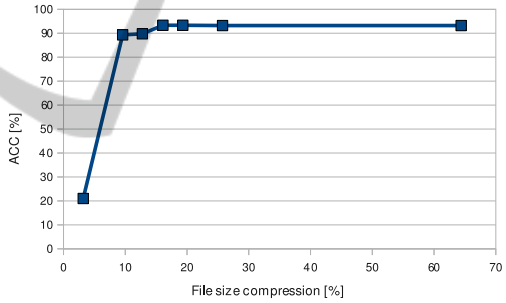

Figure 4: Dependence of ASR accuracy (ACC) on file size reduction (MFCC features and channel CS0).

4 CONCLUSIONS

The analysis of speech recognition using MP3 compressed audio data was done. The achieved results proved acceptable accuracy of speech recognition of MP3 compressed speech data. The most important contributions are summarized in the following points.

- ACC decreases rapidly for shorter frame length of short-time features when MP3 speech is rec-

ognized. It is affected by perceptual masking in MP3 compression scheme and decreasing of short-time Fourier analysis frequency resolution used in computation of MFCC and PLP features.

- Generally, the loss of accuracy is very small up to bit-rate 24 kbps. In this case the size of compressed data is just 10% of full precision linear PCM and ACC decreased by 6% for MFCC and only by 3% for PLP features.
- The results are worse for noisy channels where 50% decrease of ACC can be observed for MFCC features, comparing standard PCM and 24 kbps MP3 speech signal from desktop microphone. This decrease is just about 8% for PLP features.
- Realized experiments proved that MP3 compressed speech files used in standardly available consumer devices such as MP3 players, recorders, or mobile phones, can be used for off-line automatic conversion of speech into text without critical loss of an accuracy. PLP features seem to be preferable for speech recognition in this case.

ACKNOWLEDGEMENTS

This research was supported by grants GAČR 102/08/0707 “Speech Recognition under Real-World Conditions” and by research activity MSM 6840770014 “Perspective Informative and Communications Technicalities Research”.

REFERENCES

Barras, C., Lamel, L., and Gauvain, J.-L. (2001). Automatic transcription of compressed broadcast audio. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 265–268, Salt Lake City, USA.

Bouvigne, G. (2007). MP3 standard. Homepage. http://www.mp3-tech.org.

Bořil, H., Fousek, P., and Pollák, P. (2006). Data-driven design of front-end filter bank for Lombard speech recognition. In *Proc. of ICSLP 2006*, Pittsburgh, USA.

Brandenburg, K. and Popp, H. (2000). An introduction to MPEG layer 3. *EBU Technical Review*.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Pichney, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., and Zhu, W.-J. (2004). Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. on Speech and Audio Processing*, Vol.12(No.4):420–435.

Chen, S. S., Eide, E., Gales, M. J. F., Gopinath, R. A., Kanvesky, D., and Olsen, P. (2002). Automatic transcription of broadcast news. *Speech Communication*, Vol.37(No.1-2):69–87.

Cheng, M. and et. al. (2008). LAME MP3 encoder 3.99 alpha 10. http://www.free-codecs.com.

ELRA (2009). Czech SPEECON database. Catalog No. S0298. http://www.elra.info.

ETSI (2007). Digital cellular telecommunications system (Phase 2+) (GSM). Test sequences for the Adaptive Multi-Rate (AMR) speech codec. http://www.etsi.org.

Fousek, P. (2006). CtuCopy-Universal feature extractor and speech enhancer. http://noel.feld.cvut.cz/speechlab.

Fousek, P. and Pollák, P. (2003). Additive noise and channel distortion-robust parameterization tool. performance evaluation on Aurora 2 & 3. In *Proc. of Eurospeech 2003*, Geneve, Switzerland.

Gauvain, J.-L., Lamel, L., and Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, Vol.37(No.1-2):89–108.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, Vol.87(No.4):1738–1752.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing*. Prentice Hall.

ITU-T (2007). International Telecommunication Union Recommendation G.729, coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction(CS-ACELP). http://www.itu.int/ITU-T.

Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., and Srivastava, A. (2000). Speech and language technologies for audio indexing and retrieval. *Proc. of the IEEE*, Vol.88(No.8):1338–1353.

Nouza, J., Červa, P., and Ždánský, J. (2009). Very large vocabulary voice dictation for mobile devices. In *Proc. of Interspeech 2009*, pages 995–998, Brighton, UK.

Pollák, P. and Černocký, J. (2004). Czech SPEECON adult database. Technical report.

Psutka, J., Müller, L., and Psutka, J. V. (2001). Comparison of MFCC and PLP parameterization in the speaker independent continuous speech recognition task. In *Proc. of Eurospeech 2001*, Aalborg, Denmark.

Psutka, J., Psutka, J., Ircing, P., and Hoidekr, J. (2003). Recognition of spontaneously pronounced TV ice-hockey commentary. In *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 83–86, Tokyo.

Rajnoha, J. and Pollák, P. (2011). Asr systems in noisy environment: Analysis and solutions for increasing noise robustness. *Radioengineering*, Vol.20(No.1):74–84.

Valin, J.-M. (2007). The speex codec manual. version 1.2 beta 3. http://www.speex.org.

Vaněk, J. and Psutka, J. (2010). Gender-dependent acoustic models fusion developed for automatic subtitling of parliament meetings broadcasted by the Czech TV. In *Proc. of Text, Speech and Dialog*, pages 431–438, Brno, Czech Republic.

Young, S. and et al. (2009). *The HTK Book, Version 3.4.1*. Cambridge.