MICROPHONE TRANSFER FUNCTION ADAPTATION USING A BI – QUAD FILTER AND DCL

J. Stergar[1], D. Miletić[1], C. Beaugeant[2], B. Tramblay[2]

[1]UNI Maribor, Faculty of EE and Computer Science, Maribor, Slovenia
[2]Siemens AG, formal ICM Mobile Phones, Munich, Germany

Key words: Bi-Quadratic filter, microphone transfer function, dynamic compression, adaptation, noise robustness, audio signal enhancement

Abstract: A suitable adaptation of the microphone in the audio path of a mobile device is a very sensitive tasks. The conformation of the frequency response characteristic to the GSM standards is inevitable. Achieving the highest correlation between the specified GSM frequency response specification with the microphone and speaker characteristic is a delicate matter. In this paper we will present tests performed for microphone transfer function adaptation on a mobile phone using a Bi-Quad filter. Therefore a cascading with a IIR filter of the II. order was applied. The main goal of our test was to evaluate the influence of a recursive filter in the audio input path to the recognition rate of the embedded recognizer. The focus of our tests was to simulate the cascaded components with only one substitute having similar (almost equivalent) frequency response characteristics. By doing so we assumed that the microphone used was a high quality microphone with a linear frequency response. Further enhancement of the audio quality was performed by a more sophisticated dynamic compression and limitation algorithm (DCL). For the evaluation tests the Motiv and Aurora 3 databases were applied to the audio input. The tests have shown that with appropriate adaptation of the cascaded components an improvement of the recognition rate is realizable. The recognition rate of the mobile phone embedded recognizer was enhanced for over 6% indicating that the proposed approach sensibly contributes to speech signal enhancement.

Adaptacija prenosne funkcije mikrofona z Bi – Quad filtrom in DCL

Kjučne besede: Bi-Quad filter, mikrofonska prenosna funkcija, dinamično zgoščanje, adaptacija, šumna robustnost, izboljšanje avdio signala

Izvleček: Ustrezna adaptacija mikrofonske prenosne funkcije avdio poti mobilnega telefona je zelo občutljiva naloga. Skladnost odzivne frekvenčne karakteristike z GSM standardi je obvezna. Doseči kar najvišje sovpadanje med standardiziranim GSM odzivom prenosnega telefona ter dejansko prenosno karakteristiko mikrofona in zvočnika je dokaj težavno. V tem članku bomo predstavili preskuse za adaptacijo mikrofonske prenosne funkcije mobilnega telefona z uporabo rekurzivnega filtra II. stopnje t.i. Bi-Quad filtra. Zato smo v kaskado z vhodno avdio potjo telefona vstavili ustrezen adaptacijski filter. Glavni cilj naših preskusov je bil oceniti vpliv kaskadiranega filtra na uspešnost razpoznavanja vsajenega razpoznavalnika telefona. Pri izvedenkah eksperimentih smo se osredotočili na kaskadiranje zgolj enega samega člena (rekurzivnega filtra) s podobno oz. ekvivalentno frekvenčno odzivno karakteristiko kot veleva GSM standard. Pri tem smo uporabili visokokakovosten mikrofon z linearnim frekvenčnim odzivom. Dodatno izboljšanje kakovosti avdia je bilo izvedeno z naprednejšim algoritmom dinamičnega zgoščanja in omejevanja (DCL). Za evalvacijo razpoznavanja vsajenega razpoznavalnika smo uporabili Motiv in Aurora govorni bazi z omejenim naborom ukaznih besed. Preskusi so pokazali, da lahko z ustrezno adaptacijo prenosne funkcije mikrofona izboljšamo uspešnost razpoznavanja. Z izboljšanjem govornega signala smo povečali uspešnost razpoznavanja vsajenega razpoznavalnika za 6%.

1 Introduction

In this article we will present tests regarding the influence of a II. order recursive filter implementation on speech recognition in a mobile device. This kind of the so called Bi-Quad filter is usually used in mobile devices in cascade with a microphone and a loudspeaker.

The purpose of cascading a filter in the audio path of a recognizer is to adjust the frequency response – the transfer function of microphone – to a specific GSM standard. Recursive filter and microphone or loudspeaker together must form a suitable transfer function which is defined with one of the GSM standards.

A cornerstone of quality in telecommunication is the acoustic quality of terminals. Nevertheless a combination of several factors influences this quality. Essential to the terminals are the physical characteristics of the transducers. This leads to possible distortions of signals like deviation from the ideal frequency response or even nonlinearities. Other factors are coupled with the environment and the use case in which the terminals are operated (e.g., noisy environment or hands-free mode), causing degradations like echo and unintelligible speech for the far-end listener. In principle, most effects can be reduced by elaborate acoustical or mechanical designs.

However, without the help of digital signal processing algorithms, the acoustic quality remains insufficient /3/.

Considerable amount of varying background noise is a problem for all mobile devices such as cell phones or speech controlled car systems. Automatic systems are much more sensitive to the variability of the acoustic signal than humans. Therefore the recognition error rates of speech recognition systems using standard methods usually rise considerably in these conditions /1/.

Besides the quest for robust features two main lines of research aimed at increasing performance of speech recognizers:

- speech signal enhancement and
- model adaptation.

Although some adaptation techniques achieve very good performance, their use in embedded systems is only of limited interest. This is due to the fact that recognizers operating in mobile phones are subject to constantly changing environments and little or no adaptation data.

In contrast, speech enhancement techniques require no training; therefore they are suitable for embedded systems and additionally provide "real-time" improvement of recognition rates. For a resource constrained mobile phone, speech signal enhancement has the added advantage that the same program code can be used to improve not only the recognition rates of the speech recognizer but also the quality of the speech signal for the far-end talker during a voice call. Of course, different tunings of the enhancement algorithm have to be found for both cases in order to optimize for a machine or a human being the listener /2/.

The first issue to deal with regarding speech enhancement in mobile phones is the quality and characteristics of transducers. The frequency characteristics of microphones and loudspeakers do not necessarily comply with the requirements given.

For acoustic shock prevention dynamic compression with signal limitation was additionally applied.

2 Dynamic Compressor and Limiter

Acoustic shock is reduced by limiting an input signal based on tones detected through frequency domain analysis. Further enhancement of the audio quality is performed by a more sophisticated dynamic compression and limitation (DCL) algorithm. The basic principle of the DCL is the following: by amplification of medium signal levels speech intelligibility is improved. For each consecutive frame, the power (PWR) of the signal is computed and weighted with the power of the previous frame. Depending on this energy, a first gain is applied on the signal according to the curve in Fig. 3. An amplification is applied for frames whose power belongs to the interval [Lim E, Lim L]. For low power signal, no amplification is applied and for high power frames, a limitation to Lim L is applied. In addition, a general gain C is applied to the signal according to the shift up head-room, so that a dynamic gain according to the dotted curve is applied to each consecutive frame.

The DCL provides speech signals which sound more 'direct' and 'present' by the reduction of the dynamic range and limitation of the signal to a maximum level. If such property is welcome for speech conversation scenario [1], it is also foreseen that speech recognizer could be positively influenced by an enhancement of such a speech presence: Amplifying speech region where the information of the signal is the most important is a priori a good way to increase the performance of speech recognition.

Moreover, for scenario where the Signal to Noise Ratio (SNR) is not too low, the influence of the noise is reduced. Indeed, in such scenario the noise level is lower the threshold Lim E and that "noise only" frames are less amplified than "speech + noise" frames. As a result, the noise period influence on speech recognizer is reduced compared to speech periods.

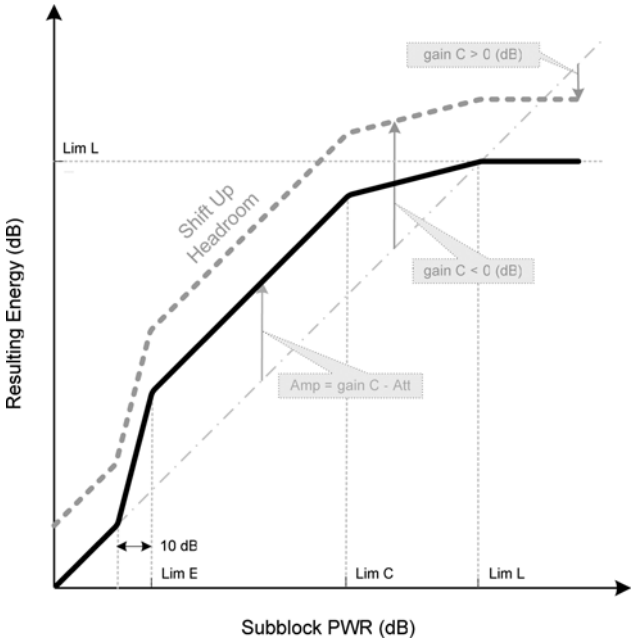

Fig. 1: The DCL function.

Likewise, sudden bursts of noise whose energy is higher then the threshold Lim L are limited by the DCL. It involves that their influence on speech recognizer is reduced as well.

Such remarks show that a priori the DCL should enhance the performance by enhancing the 'presence' of speech and reducing the influence of static and burst noise.

3 The Bi-Quadratic adaptation filter

More and more frequently occurs that the speech input signal exhibits a "flatter" spectrum, for example when a hands-free installation is used, employing a microphone with linear frequency response. Conventional recognizers are designed to be independent of the input with which they operate, and, they are without any knowledge of the characteristics of this input. If microphones with different characteristics are likely to be connected up to the mobile phone, or more generally if the recognizer is likely to receive acoustic signals exhibiting different spectral characteristics, there are cases in which the Very Smart Recognizer (VSR) embedded in our case, operates in a sub-optimal manner. In this context, a main purpose of the microphone transfer function adaptation is to improve the speech signal making it less dependent on the spectral characteristics.

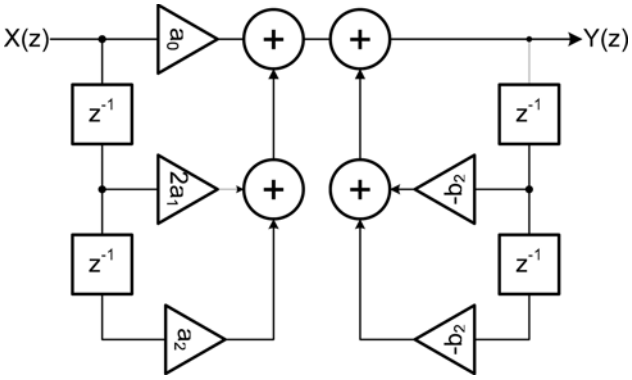

Fig. 2: Block diagram of a typical bi-quad filter.

The purpose of cascading a filter in the audio path of a recognizer is therefore to adjust/adapt the frequency response – the transfer function of microphone – to a specific standard.

The purpose of cascading a filter in the audio path of a recognizer is to adjust the transfer function of the microphone to a specific standard. A convolution function of the different characteristics of the filter and microphone in cascade is performed and together they must form a suitable transfer function which is defined with one of the GSM standards /5/.

$$H(z) = g\frac{a_0 + 2\cdot a_1 z^{-1} + a_2 z^{-2}}{1 + 2\cdot b_1 z^{-1} + b_2 z^{-2}} \qquad (1)$$

A II. order recursive filter – the Bi-Quadratic filter – was implemented to adapt the microphone transfer function. The Bi-Quad digital filter is a common name for a two-pole, two-zero recursive filter which name was derived from the transfer function structure of the filter. This kind of the Bi-Quadratic filter is typically used in mobile devices in cascade with a microphone and loudspeaker.

Practically this is an Infinite Impulse Response (IIR) filter of the second order. The transfer function for this kind of a filter can be defined from its block diagram (Figure 2).

It can be seen that this is a common two-pole, two-zero digital filter with a typical transfer function (Figure 3) /6/.

Using the shift theorem for z transforms, the difference equation for the Bi-Quad can be written by inspection of the transfer function as:

$$y(n) = \sum_{i=0}^{M} b_i x(n-i) - \sum_{j=1}^{N} a_j y(n-j) \qquad (2)$$

where x(n) denotes the input signal sample at time n, and y(n) is the output signal /8/.

In most fixed-point arithmetic schemes (such as two's complement, the most commonly used) there is no possibility of internal filter overflow. That is, since there is fundamentally only one summation point in the filter, and since fixed-point overflow naturally "wraps around" from the largest positive to the largest negative number and vice versa, then as long as the final result y(n) is "in range", overflow is avoided, even when there is overflow of intermediate results in the sum. This is an important, valuable, and unusual property of the used filter structure. There are twice as many delays as are necessary.

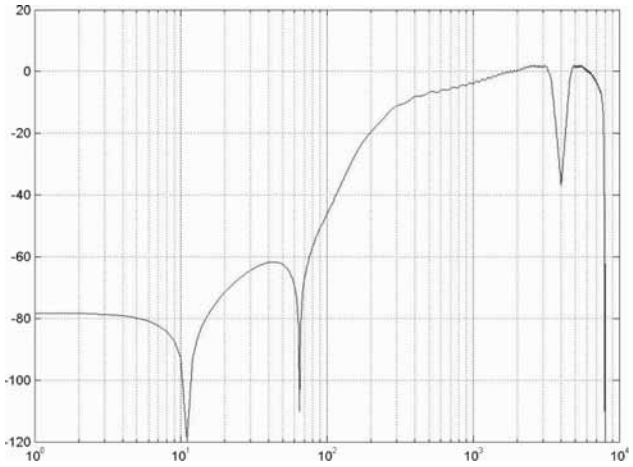

Fig. 3: A characteristic frequency response of the bi-quadratic filter (handset parameterisation).

As a result, the bi-quad structure is not canonical with respect to delay. In general, it is always possible to implement an N^{th}-order filter using only N delay elements. It is a very useful property of the Bi-Quad implementation that it cannot overflow internally in two's complement fixed-point arithmetic: As long as the output signal is in range, the filter will be free of numerical overflow. Most IIR filter implementations do not have this property /9/.

4 Used test material

4.1 The MoTiV database

The database MoTiV was recorded after the initiative between the industrial partners Philips, Siemens, Bosch, and Volkswagen in the subproject Man-Machine Interaction /7/. In total, 35 hours of hands-free multi-channel recorded speech data from about 640 drivers were collected in seven different mid- to upper-class ranged cars. All recordings were simultaneously made at least by two microphones, which had been fixed on the car ceiling at the A-beam and in the middle between driver and passenger.

For our experiments only a subset of recorded material was used. It consisted of 26 words (mostly command words) in German language with 100 diverse samples for each word. As mentioned all of these samples were recorded in different car environments therefore samples with different amount of noise were included. Beside that all samples were recorded by both genders. (Table 1).

The original samples had to be preprocessed (down-sampled) because of the embedded Very Smart Recognizer (VSR) limitations. Little memory and low computational power on the used mobile phone VSR supported only samples with f_s = 8 kHz for processing.

Table 1: The command words used from Aurora database.

Motiv		Aurora
ändern	löschen	
aus	Navigation	ZWEI
Ende	nein	ZWO
halt	Radio	SECHS
Hauptmenü	Start	FUENF
Hilfe	Stop	VIER
Information	stumm	ACHT
ja	suchen	DREI
Karte	Telefon	NEUN
Kassette	wählen	SIEBEN
Korrektur	weiter	EINS
lauter	wiederholen	NULL
leiser	zurück	

4.2 The Aurora 3 Database

The Aurora 3 database is a database of digits. This database is a subset of the SpeechDat-Car database in German language which has been collected as part of the European Union funded SpeechDat-Car project. It contains isolated and connected German digits spoken in the following noise and driving conditions inside a car: High/low speed good/rough road, stopped with motor running, town traffic /10/. Only digits from 0 to 9 are included. The samples are of different kind. A single sample consists of one or more digits. The frequencies of the digit appearance in the samples corpus differ. The used database was recorded on two different channels (ch0 and ch1). The ch0 is the primary channel where much less additive noise is present compared to channel ch1. The whole database has 3118 samples proportionally distributed in each channel. For the VSR recognition testing the samples had to be converted.

5 The Very Smart Recognizer

The Very Smart Recognizer (VSR) is a software-only speech recognition component especially designed for mobile terminals by Siemens AG. Featuring a modular architecture and flexible configuration it is particularly well suited for the support of voice commands /11/.

For the recognition experiments we used a speaker independent HMM (Hidden Markov Model) based VSR (V4.50). Speaker independent HMM-based technology offers command-and-control and digit dialling, i.e., recognition of commands and phone numbers without requiring a training phase. Natural number (e.g. twenty-five, ninety-eight) recognition applies the same technology while improving the usability of voice dialling. HMM based recognizers imply a higher implementation complexity and need appropriate speech databases in many languages. Today's memory and computational resources in 3G phones are facilitating deployment of such technology.

The VSR used for testing was in form of an executable and was activated with the appending parameters:

<melParamFile> <hmmFile> <vocabularyFile> <sampleFile>

In the options field a specific output file format specification for the recognition module was possible. In the melParamFile different kinds of parameters for VSR were included. The most important parameter for testing was the noise reduction parameter (NSR) set for all tests. The hmmFile carried a phonetic description for the used database. The vocabularyFile included a vocabulary of currently used database. The sampleFile represented the input data file with the recognition samples. This sample file was conveyed to the VSR input.

Fig. 4: The reference diagram for global recognition experiments.

6 Experiments

In the preliminary tests we applied the Bi-Quad module as speech enhancement pre-processing unit immediately before the VSR unit. For tests samples with different noise characteristics were deliberately selected (MoTiV/Aurora 3) for the evaluation of the recognition results with different DCL parameterisations.

The origin test framework consisted of the pre-processing module in cascade with the VSR module (Figure 4). This framework was used for separate tests using the Bi-Quad filter or DCL. Lastly a general I. order high-pass filter was also tested (Figure 5).

We started the tests with the objective of global recognition rates for each database, which has been later used as reference to other performed experiments. In addition a variety of experiments had been performed.

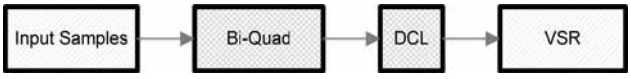

Fig. 5: The extended reference diagram supplemented with the DCL in cascade with the high-pass bi-quad filter.

The first step in all experiments as already mentioned was the evaluation of the global recognition rate for each database (VSR, Figure 10-13).

In the first experiment a Bi-Quadratic filter was inserted into the recognition path adapting the microphone transfer function. In the experiment our intent was to examine the speech recognition effectiveness/improvement of the VSR using different parameterisations of the Bi-Quad filter: high-pass (HP), handset and hands-free (Figure 7).

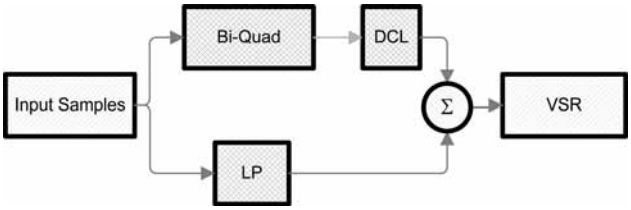

Fig. 6: The reference diagram combining the paths of the high-pass Bi-Quad + DCL with the low-pass branch.

In the second experiment a DCL module was inserted into a cascade before the VSR recognition stage replacing the Bi-quad (Figure 4). The Bi-Quad as well as the DCL executable were applied to every sample in the two databases. Therefore a comparison of the results with the global recognition rate could be performed (Figure 7).

In the second experiment another test was performed. A Bi-Quad filter was inserted before the DCL module and the high-pass, hands free and handset parameterisation of the cascaded Bi-Quad was examined (Figure 5). Our intent was to evaluate the influence of the inserted DCL on the VSR recognition.

In the third experiment another test scenario was applied. The recognition path was split into a high-pass filter branch in cascade with the DCL and a low-pass filter branch (Figure 6). The separated signals were summarised before the final recognition stage. Using the depicted method we separated the signal (information) from noise and applied the DCL only to the noisy part of the signal.

The goal of this experiment was to evaluate the impact of the DCL stage on the final recognition rate after summarisation.

All recognition rates (Word Correct Rates) were estimated with:

$$WCR\,(\%) = \frac{H}{N_r} = \frac{N_r - (D+S)}{N_r} \qquad (3)$$

where N_r represents the total number of words in the reference corpus, S the number of substituted words in the confusion matrix, D the number of words deleted from the confusion matrix, and H the number of correctly recognised words /12/. For all experiments an executable for the confusion matrix generation was used gathering the results for the word and global, recognition rates with the substitutions and deletions for each word tested.

The high/low pass filters as well as the BiQuad were realized in MATLAB environment.

7 Test results

All experiments were mainly performed with different DCL and filter parameterisations on the MoTiV database. The additional tests were performed with the Aurora 3 database and were used as a confirmation/rejection reference for the test results gained with the MoTiV database.

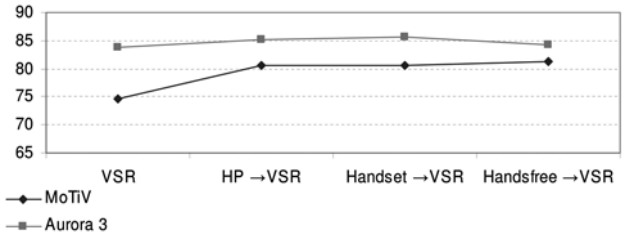

Fig. 7: The WCR for the MoTiV and Aurora 3 database using different Bi-Quad parameterisations.

Firstly our objective was to test the influence of the Bi-Quad adaptation of the transducers transfer function. The experiments show that all of the three parameterisations – high-pass, hands-free and handset contribute to increased the WCR of the implemented VSR in average over 6% (Figure 7). The promising recognition enhancement formed the foundation for the experiments with the DCL cascade. Our goal was to estimate if the cascading of the DCL preserves the gained improvement in recognition of the VSR and furthermore if only a single DCL parameterization would be sufficient for database independent recognition. Therefore in parallel supplementary experiments with the Aurora 3 database were performed.

With the applied DCL we also performed tests using different cut-off frequency ($f_{cut\text{-}off}$) variations. We estimated if any enhancements are possible for the different cut-off window. Slightly enhancements for some test scenarios using $f_{cut\text{-}off}$ = 350Hz were observed but the overall recognition results were best by using the $f_{cut\text{-}off}$ = 300Hz.

Further tests have been performed using the Aurora 3 database separately evaluating the global recognition rates for samples in both channels. The graphs indicate that the DCL parameterisation used for the MoTiV database almost preserves the gained recognition accuracy of the VSR (Figure 8). The recognition rate of the Aurora 3 database just slightly degrades in spite of the DCL parameterisation being optimised on the MoTiV database.

The experiments indicate that the used Bi-Quad considerably improves the overall recognition rate on the embedded VSR. Furthermore we can assume that the applied version of the DCL does not essentially degrade the recognition robustness.

Our deduction was already foreseen after making the DCL parameterization for the MoTiV database, since speech samples can differ in many parameters not just from loudness but most important from the level and characteristic of the added noise.

8 Conclusion

In order to make speech recognition more robust preprocessing influencing the physical characteristics of the mobile device transducers can be applied. The amount of varying background noise is a problem for all mobile devic-

es especially for automatic systems which are much more sensitive to the variability of the acoustic signal than humans. Therefore a pre-processing scenario was introduced improving speech recognition error rates with microphone transfer function adaptation limiting the background noise variations with a DCL preserving speech robustness.

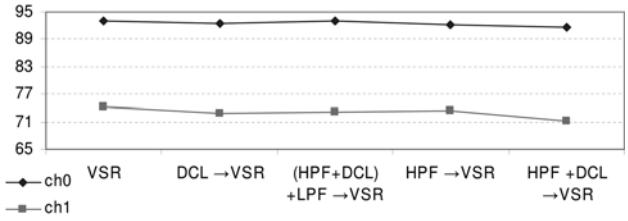

Fig. 8: The WCR for the supplemental experiments on the Aurora 3 database with the applied DCL.

We can observe that the applied Bi-Quad module considerable improves the Word Correct Rate (>6%) of the recognition module. Experiments also show that the DCL module applied for background noise variations elimination does not essentially degrade the gained improvement with the transfer function adaptation. Hence an assumption can be made that the introduced framework is preserving the speech robustness of the embedded VSR. There are strong indices that there can be a single DCL parameterization which would guarantee constant recognition results for arbitrary speech samples.

9 References

/1/ F. Hilger, H. Ney: Quantile Based Histogram Equalization for Noise Robust Large Vocabulary Speech Recognition, IEEE Transactions On Speech And Audio Processing, pp: 845 – 854, Volume 14, Issue 3, 2006.

/2/ S. Aalburg, C. Beaugeant, S. Stan, T. Fingscheidt, R. Balan, J. Rosca; Single- And Two-Channel Noise Reduction For Robust Speech Recognition In Car, Proceedings of ISCA Workshop, Multi-Modal Dialogue in Mobile Environments, Germany, 2002.

/3/ C. Beaugeant, M. Schönle, I. Varga; Challenges of 16 kHz in Acoustic Preand Post-Processing for Terminals, IEEE Communications Magazine, May 2006.

/4/ H. Gustafsson, I. Claesson, U. Lindgren; Low-Complexity Feature-Mapped Speech Bandwidth Extension," IEEE Transactions On Speech And Audio Processing, pp: 577- 588, Vol. 14, Issue 2, 2006.

/5/ 3GPP TS 26.131, "Technical Specification Group Services and System Aspects, Terminal Acoustic Characteristics for Telephony, Requirements (Release 6),", 2004.

/6/ U. Zölzer, Digitale Audiosignalverarbeitung, die 2. durchgesehene Auflage - Stuttgart : Teubner, 1997.

/7/ D. Langmann, H. R. Pfitzinger, T. Schneider, R. Grudszus, A. Fischer, M. Westphal, T. Crull, U. Jekosch, CSDC - The MoTiV Car Speech Data Collection. LREC98, 1998.

/8/ A. V. Oppenheim and R. W. Schafer; Digital Signal Processing, Englewood Cliffs, NJ, Prentice-Hall, 1975.

/9/ J. O. Smith; Introduction to Digital Filters with Audio Applications, W3K Publishing, http://books.w3k.org/, 2007.

/10/ Evaluations and Language resources Distribution Agency (http://www.elda.org/).

/11/ I. Varga et. all; ASR in Mobile Phones - An Industrial Approach, IEEE Transactions On Speech And Audio Processing, Vol. 10, No. 8, 2002.

/12/ I. McCowan, D. Moore, J. Dines, D. G.-Perez, M. Flynn, P. Wellner, H. Bourlard; On the Use of Information Retrieval Measures for Speech Recognition Evaluation, IDIAP Research Report 04-73, 2005.

/13/ 3GPP TS 26.071, AMR speech CODEC, http://www.3gpp.org/ ,2007.

Janez Stergar, Dejan Miletić
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova 17, 2000 Maribor, Slovenia
janez.stergar@uni-mb.si

Christophe Beaugeant, Bruno Tramblay
Siemens AG, ICM Mobile Phones
Grillparzerstrasse 10-18, 81675 Munich, Germany

Prispelo (Arrived): 27.03.08 *Sprejeto (Accepted):* 28.5.08