Journal of ELECTRICAL ENGINEERING, VOL. 65, NO. 6, 2014, 325–332

INTELLIGIBILITY ASSESSMENT OF IDEAL BINARY–MASKED NOISY SPEECH WITH ACCEPTANCE OF ROOM ACOUSTIC

Vladimír Sedlák — Daniela Ďuračková
Roman Záluský — Tomáš Kováčik *

In this paper the intelligibility of ideal binary-masked noisy signal is evaluated for different signal to noise ratio (SNR), mask error, masker types, distance between source and receiver, reverberation time and local criteria for forming the binary mask. The ideal binary mask is computed from time-frequency decompositions of target and masker signals by thresholding the local SNR within time-frequency units. The intelligibility of separated signal is measured using different objective measures computed in frequency and perceptual domain. The present study replicates and extends the findings which were already presented but mainly shows impact of room acoustic on the intelligibility performance of IBM technique.

K e y w o r d s: ideal binary mask, speech separation, intelligibility, room acoustic

1 INTRODUCTION

The ideal binary mask is very useful technique for separation mixed signals. There are many speech and audio applications where the desired signal is corrupted by highly correlated noise sources. Separating such signals from their mixture has often been considered as one of the most challenging research topics in the area of speech enhancement [1]. Our present work is aimed at investigation of influence of environment and the origin of the mixed signals on the intelligibility. In many of previous study is shown only impact of noise level, masker type or binary mask error on performance of this technique, but in this article is presented also impact of environment in which are signals acquired. It means the distance between receiver and source, room dimension and reflection coefficients are accepted in evaluation process. All this parameters are simply described using the room impulse response (RIR) which is generated for our test room and test conditions.

The approaches appropriate for solving this issue (source separation) can be divided into two groups: 1) model based method, and 2) source driven or computational auditory scene analysis (CASA)-based method. The first group, model-based separation system is based on statistical models including vector quantization [2], Gaussian mixture models [3] and Hidden Markov models [3]. The CASA-based methods search auditory scenes in the time-frequency domain which are probably to come from the same sources of speech signals by exploiting the characteristics of human auditory system [4]. The CASA-based methods rely on extracting psychoacoustics cues from the given mixed signals and work in two stages: segmentation and grounding. The ideal binary mask has been set as a computational goal in CASA algorithms and has also been adopted in "missing feature" speech recognition technique. The ideal binary mask takes value of zero and one and is briefly described in the next section. Currently the research groups working on speech-separation problem especially focus on the topic of how to separate speech signal from interfering sounds, including other speech [5].

The rest of paper is structured as follows: In the next section, the main problem definition is described. It includes specification of ideal binary mask and specification of objective metrics for measuring the intelligibility. The next section is devoted experiments and analyzes. There are described methods, goals of individual analyses and individual results. The last section concludes this article and summarizes achieved results.

2 PROBLEM DEFINITION

This section is devoted to single channel source separation of mixed signals. Especially to technique based on ideal binary mask. Because of is suitable to compare achieved results under different test conditions the value of intelligibility have to be measured. Therefore, in this section are presented objective metrics for intelligibility assessment. It is focused on perceptual evaluation of speech quality (PESQ), segmental version of SNR (SNRS) and on the sort-time intelligibility measure (STOI).

2.1 Single Channel Source Separation

The approaches for signals separation can be divided based on number of microphones which are included in process of signal acquisition. In case of only one microphone is available a process of separation is called the single channel source separation. Separating different speech

* Slovak University of Technology in Bratislava, Slovakia, Faculty of Electrical Engineering and Information Technology, Institute of Electronics and Photonics, Ilkovičova 3, 812 19 Bratislava, {vladimir.sedlak, daniela.durackova, roman.zalusky, tomas.kovacik}@stuba.sk

DOI: 10.2478/jee-2014-0054, Print ISSN 1335-3632, On-line ISSN 1339-309X

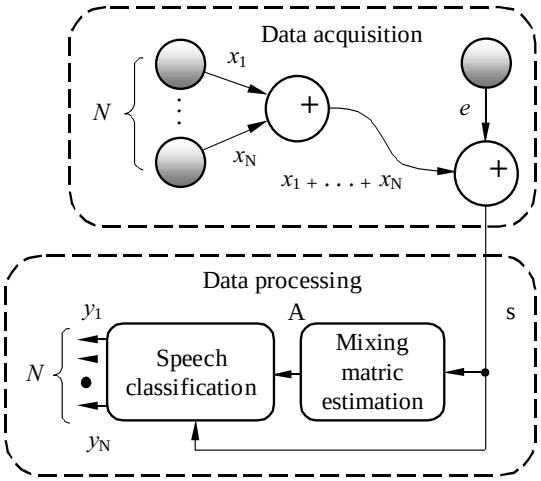

Fig. 1. Flow chart of single-channel source separation

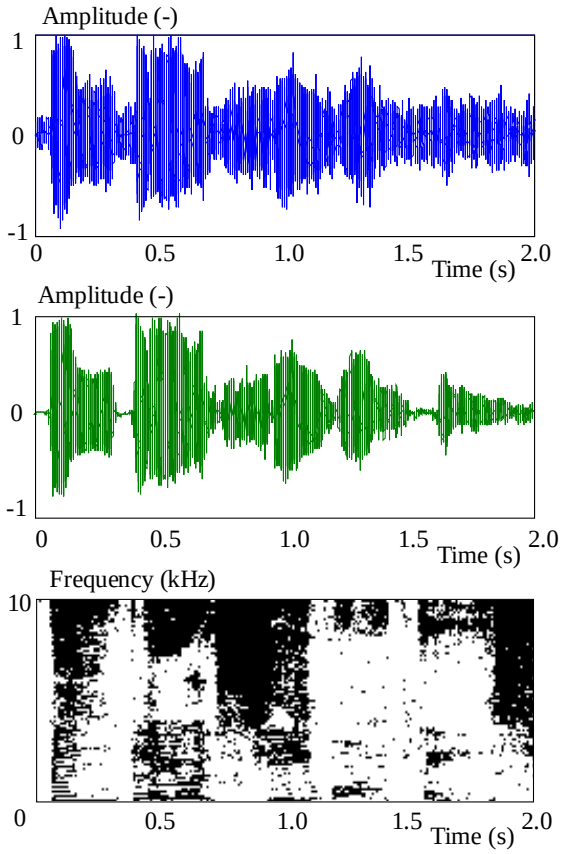

Fig. 2. Waveforms for: original (brighter) and input signal with $SNR = 5$ dB – top; original (brighter) and separated signal – middle; Illustration of ideal binary mask for acquiring separated signal with $LC = 0$ dB

or audio signals using single-channel approach is very challenging topic and can be defined as follows:

$$\mathbf{S} = \mathbf{AX} \tag{1}$$

where $\mathbf{S}$ is the observed input matrix of mixtures $\mathbf{S} = [s_1(t), s_2(t), \ldots, s_M(t)]$ and $\mathbf{X}$ represents matrix of original source signals $\mathbf{X} = [x_1(t), x_2(t), \ldots, x_N(t)]$. The symbol $\mathbf{A}$ is a mixing matrix which is corresponding to the mixing conditions and M is the number of inputs (microphones). In a single-channel separation case ($M = 1$), the number of mixtures is one and equation (1) can be simplified to equation (2) where N represents the number of sources which contribute to input signal $s(t)$, $e(t)$ is additive noise and $x_n(t)$ is the n^{th} source signal at time t.

$$s(t) = \sum_{n=1}^{N} x_n(t) + e(t)\,. \tag{2}$$

Formula (2) is graphically represented by top block of Fig. 1 which shows flow chart of single channel source separation (SCSS) and is divided into two main blocks. This block shows process of data acquisition, in other words the creation of mixtures and signal capture. Sources or original signals produce input signal $s(t)$ which can be written in vector notation as $s = g(x_1, x_2, \ldots, x_N)$, where g is some possibly non-linear and stochastic mixing process. The bottom block represents data processing to achieve or estimate original signals from mixture and usually is based on data filtering, data decomposition and grouping or on source modeling. The aim is estimating the mixing matrix which is applied on the input signal to produce estimated values of desired signals $y_n(t)$.

2.2 Ideal Binary Mask

An ideal binary mask (IBM) is defined in the T–F domain as a matrix of binary numbers. The T–F representation makes it possible to utilize the temporal and spectral properties of speech and is obtained using *eg* the short-time Fourier transform or a Gammatone filter bank. The IBM is defined by comparing the SNR within each T–F unit again a local criterion (LC) or threshold measured in units of decibels [6]. Only the T–F units with local SNR exceeding LC are assigned 1 in the binary mask what in mathematically described by formula (3). Where $T(j,k)$ is the power of the target signal, $M(j,k)$ is the power of the masker signal, LC is a local SNR threshold, j the time index, and k the frequency index. The LC value is the threshold for classifying the T–F unit as target or masker and determines the amount of target and masker signal in the processed signal, if the binary mask is applied to the mixture. In the most CASA-based method the LC is set to 0 dB.

To calculate the IBM the unmixed signals must be available, what is problem in real-life application, or the IBM can be estimated from mixed signal using methods based on a Bayesian classification of speech features, pitch continuity information, sound localization cues and others. In picture number 2 are depicted results obtained by IBM which was computed from original (unmixed signals). In the top part of figure is depicted original signal (two seconds long male utterance) and mixed signal which is generated from this original signal by adding babble with SNR 5 dB. Both signal are sampled at 25 kHz and normalized to maximum value of signals. In the middle

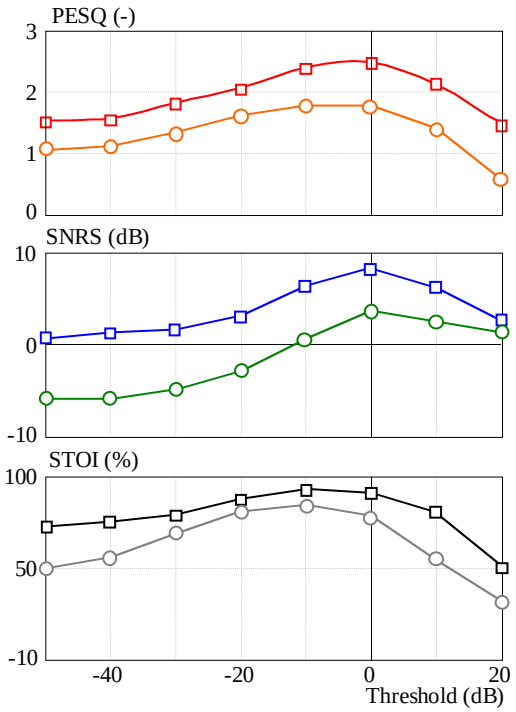

Fig. 3. Intelligibility of IBM-processed mixtures masked by multi-talker babble at −5 dB (brighter color) and 5 dB SNR as function of local threshold used for generating the IBM

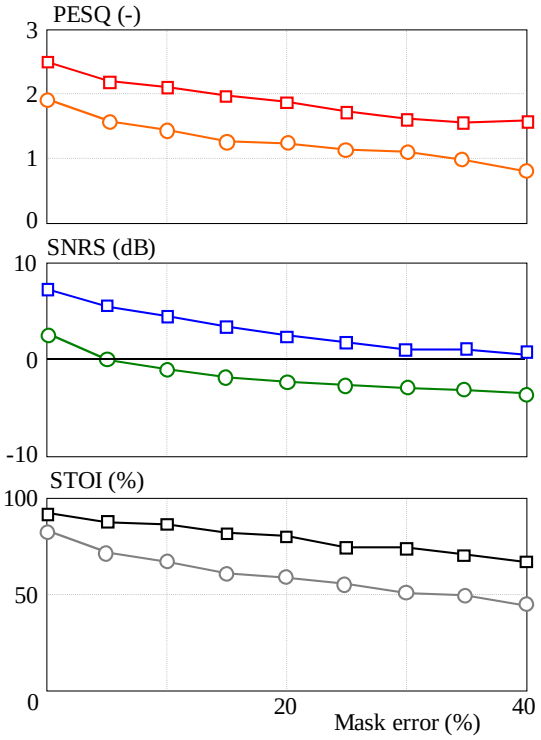

Fig. 4. Intelligibility of IBM-processed mixtures masked by multi-talker babble at −5 dB (brighter color) and 5 dB SNR as function of the overall percentage IBM error

part is depicted original (lighter color) and output (separated) signal which shows a marked improvement in quality of speech. And finally in the bottom part is shown illustration of IBM where black and white color indicates 1 and 0, and the threshold is set to 0 dB. The whole procedure for acquisition of mask will be described in the next section.

$$IBM(j,k) = \begin{cases} 1\,, & \text{if } \frac{T(j,k)}{M(j,k)} > LC\,, \\ 0\,, & \text{otherwise.} \end{cases} \tag{3}$$

2.3 Measures for Intelligibility Assessment

A necessary condition for development of source separation algorithms is the ability to measure the performance or quality of result. In general, the separation quality can be measured by comparing separated signals with reference sources (objective methods) or by listening to the separated signals (subjective methods). Subjective methods are based on ratings by human listeners according to the categories (Excellent, Good, Fair, Poor and Bad) defined in a subjective test and finally the statistical analysis is applied to these ratings to reach value of speech quality. The most commonly used methods for measuring the subjective quality of speech transmission over voice communication systems have been standardized by the International Telecommunications Union and mostly are based on 5 categories.

Objective methods can be classified into intrusive (reference) measures and non-intrusive (non-reference) measures. The intrusive measures compare the output signal (distorted signal) with the original signal, which is usually called the reference signal. The non-intrusive methods do not require a reference signal because the speech quality is determined only by the output speech signal. In general, objective speech quality measures can be categorized into three domains: time domain, spectral domain or perceptual domain. The objective measures are more preferred than subjective measures since they are more convenient and time saving, and can be repeatedly utilized for different input data sets.

A very sophisticated objective measure is the perceptual evaluation of speech quality (PESQ) metric. Procedure to compute this metrics is divided in two stages. The PESQ measure is recommended by ITU-T P.862 for speech quality assessment of 3.2 kHz handset telephony and narrow-band speech codec. The first is a time alignment stage that aligns the separated signal and reference signal. In the next stage a psychoacoustics model is used to calculate an auditory representation of the signals, followed by a cognitive model that calculates final score based on the differences between signals [7].

Formula (4) represents segmental version of SNR (SNRS), what is time domain measure. It is the ratio of energies of the reference signal $s(k)$ and the error between the separated $\hat{s}(k)$ and reference signal. Symbol k represents index of signal frame and symbol K is the total number of frames. Other objective measure with shows high correlation witch the intelligibility of noisy and time-frequency weighted noisy speech is short-time objective intelligibility measure (STOI) and was presented in [8].

Very simplified is described by formula (5) where J represents the number of frequency bins, K is the number of frames and $d_{j,k}$ represents correlation coefficient of j^{th} frequency bin of k^{th} frame.

$$SNRS = \frac{10}{K} \sum_{k=1}^{K} \log_{10} \frac{|s(k)|^2}{|s(k) - \hat{s}(k)|^2}\,, \tag{4}$$

$$STOI = \frac{1}{JK} \sum_{j,k} d_{j,k}\,. \tag{5}$$

3 EXPERIMENTS

This section is devoted for experiments which deal with the ideal binary-masked noisy signals. The goal is to show relationships between intelligibility of the output speech, measured by some of mentioned metrics, and some input parameter as IBM computing, input signals or environment. Similar experiments have been already presented *eg* in [9] but without acceptance of room acoustics and because we have used different dataset of speech, we have decided to make suchlike experiments on our samples. It could be also useful for results comparison and verification of intelligibility measures because in mentioned paper were performance measured by subjective methods. It means we have as well verified performance depending on the local SNR threshold, masker signal type and accuracy of mask.

3.1 Speech database

Speech samples were taken from database presented in [10], which consists of 100 utterances from each one of 34 speakers. This database was primary collected to support the use of common material in speech perception and automatic speech recognition studies but it was also used in the different signal processing tasks. Sentences are simple, syntactically identical phrases such as "place red at C 1 now". The noisy samples were taken from AURORA database [11], which was developed primary for performance verification of algorithms for adaptive noise cancelling. It contains noisy samples from different places *eg*: car, restaurant, exhibition hall or airport. All have been recorded at 20 kHz and then downsampled to 8 kHz. Room acoustic was modeled by room impulse response generator [12] and is based on image method. Generator enables the user to control the reflection order, room dimension and microphone directivity. This way generated response in then used for create reverberant signal from original "clean" signal

3.2 Signal processing

The IBM was computed from three types of signals: the target signal, the masker signal and the resulting mixture. It means the IBM was not estimated directly from resulting mixture as it is in real applications because the experiments are focused only on the binary masking technique and not on estimating on IBM. The same concept is presented as well in [9]. As were mentioned earlier the IBM technique is based on T–F representation of signals. That is the reason why the signals are divided into 20 ms frames with 50 % overlap between frames in the first step. Next the fast Fourier transform (FFT) is used to transforms frames into frequency domain. In this point the T–F representation is done. A local value of SNR of each T–F unit is determined by comparing energies between target and masker. The result value is compared again a threshold value T to determine whether to retain the T–F unit or to eliminate it. For separating target signal from mixture is then the magnitude spectrum multiplied by computed IBM and the inverse FFT is applied to this modified spectrum. Speech is synthesized in each 20 ms frame using overlap-and-add method.

3.3 Analyses

As was mentioned earlier a procedure for acquiring of IBM is based on comparison of the local SNR and the threshold. That is the reason why first was analyzed the impact of threshold value on intelligibility, The local SNR threshold T was varied from -50 dB to 20 dB for two input signals which were corrupted by babble noise (masker signal). The SNR of input (mixed) signal was set to 5 dB and -5 dB. Changing of the threshold affects the amount of energy how much the original signal have to exceed masker signal to mark current T–F unit as one. Achieved results are depicted in Fig. 3 and show expected trend which can be compared with results presented in [9]. What is important is plateau region around 0dB and gradual decline of intelligibility with increasing or decreasing of the T. In case when the threshold is set to 0 dB only those T–F units are retained in which is energy of the original signal higher than energy of masker signal. Increase of the T causes that less T–F units are retained because the difference between energies of original and masker signal has to be higher than in case of lower threshold value. On the other site, decrease of the T causes that more T–F units are retained. But the problem is that there are also retained units where masker signal energy is higher than original signal energy. The plateau region in a similar experiment in [9] was chosen from -25 dB to 5 dB and in presented analyze it is from -10 dB to 5 dB. This difference could be caused by different sound samples. Interesting is also comparison of trends of individual evaluation measures which are very similar although in [9] were used the subjective measures and in this analyze the objective measure.

Another question is how exactly should be estimated the IBM in order to do not affect the intelligibility. So this is the reason why we decide to analyze intelligibility depending on the mask error. And because the IBM was computed directly from input signals (not estimated from mixed signal) the error was inserted artificially. The advantage of it is that the total error value in the IBM can

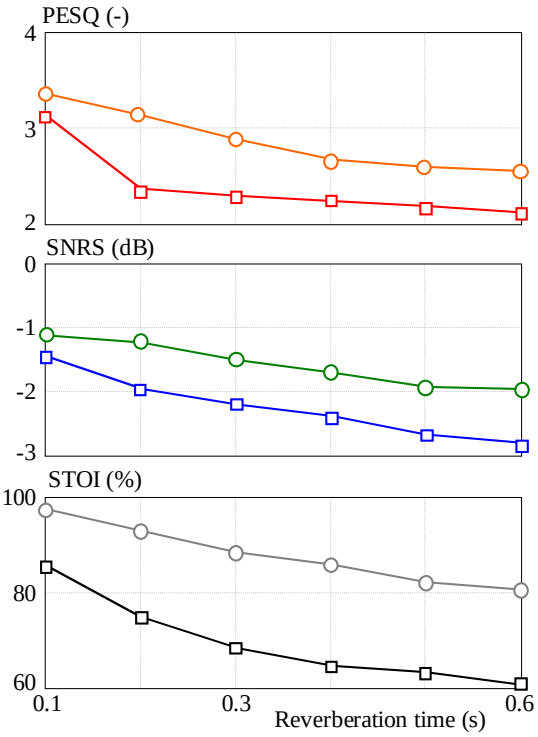

Fig. 5. Intelligibility of IBM-processed mixtures mixed with multi-talker babble at 20 dB SNR for different distances (1m (brighter color) and 3 m) between source and microphone as function of reverberation time

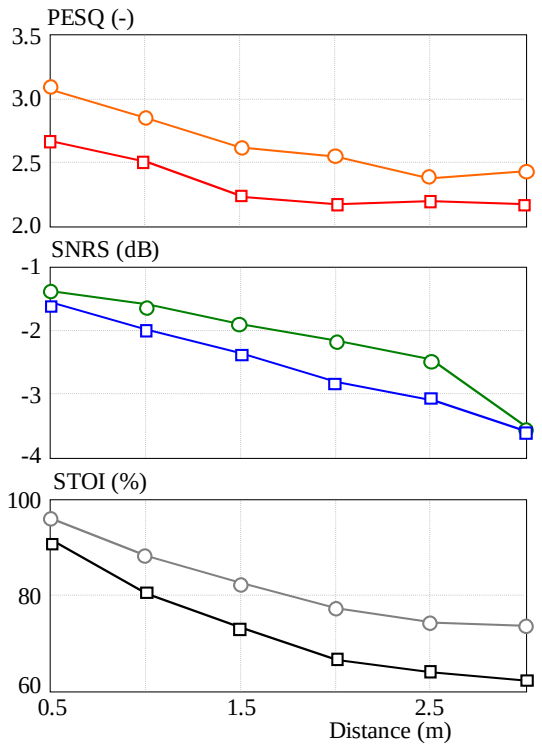

Fig. 6. Intelligibility of IBM-processed mixtures mixed with multi-talker babble at 20 dB SNR as function of distance between source and receiver for two values of RT_{60}: 300 ms and 600 ms (brighter color)

Table 1. Intelligibility of IBM-processed mixtures masked by six different types of masker signals: babble, one speaker, two speaker, noise in car, noise at airport and noise at station

Masker Type	Value (dB)	SNRS (dB)	PESQ (−)	STOI (%)
babble	−5	2.67	1.89	82.1
	5	7.48	2.48	92.2
one speaker	−5	6.51	2.31	88.9
	5	11.32	2.99	95.7
two speakers	−5	4.84	2.28	88.9
	5	10.21	2.92	95.6
noise in car	−5	4.74	1.96	83.1
	5	9.51	2.54	92.4
noise at airport	−5	4.67	1.92	84.4
	5	9.37	2.53	92.8
noise at station	−5	3.96	1.97	81.1
	5	8.72	2.5	92.3

be controlled very exactly. The mask error means how many percent of T–F units is labeled wrongly (*ie*, 0 was labeled as 1 and vice versa). In case of 10 %-error condition were 10 % of T–F units re-labeled. At the beginning the IBM was created the same way as in the previous analysis. In the next step specific count of T–F units was labeled again but with opposite value, what represents an error. The indexes of these re-labeled units were selected randomly. In this analyze the amount of error was varied from 0 % to 40 % with 5 % step. This way changed mask was then used to separate the same signals as in the previous analyze. The achieved results are shown in Fig. 4. The intelligibility of separated signal decreasing with increasing error value what is an excepted trend. Interesting is finding that this trend is nearly linear and for all evaluation measures is similar. From results can also be seen that quality of input signal (measured as SNR) influences the intelligibility of separated signal in form of offset.

All previously presented analyses used for masking only multi-talker babble, but the origin of masker can also have strong impact on performance as was presented in many previous articles. We decided to verify this argument using samples from AURORA database [12] and using speech of other speakers. Our goal was show how the origin of the masker signals can affect overall intelligibility of IBM-processed mixtures and to verify if all chosen objective measures exhibit equal trend with changing of masker signal. It is also interesting to observe whether the IBM technique is more effective when the masking has informational and energetic components or when the masking is purely energetic (noise). The IBM was generated from mixtures masked by six different masker signals: multi-talker babble, speech (one speaker), speech (two speakers), and noise in car, noise at airport and noise at station. The values of these maskers were set to −5 dB and 5 dB SNR and local threshold value was set to 0 dB. Achieved results (evaluated by SNRS, PESQ and STOI) are summarized in Tab. 1 and show that the IBM technique is more effective, in terms of improving intelligibility, when target speech is masked by speech that when

it is masked by noise. Interesting is finding that multi-talker babble had nearly equal effect like noise although it is composed of human speech. The best results, for all presented intelligibility measures, were achieved when the target signal was masked by speech of one speaker. Adding other speech of different speaker into masker signal caused decrease in speech intelligibility. The intelligibility of processed mixtures masked by noise was nearly equal, only the noise form station had a little bit stronger effect when the intelligibility was measured using SNRS. Otherwise all presented measured showed the same trend.

In all previously presented analyses have not been taken account the environment in which the sound propagates. However, this environment can have a significant effect on the signal, *eg* reflections, therefore the following analysis deals with the impact of reverberation time on the overall quality. Using the room impulse generator [12] was created imaginary testing room which was described by its impulse response and served for affecting the initial signals (without reflections). Reverberant speech samples were generated by convolving anechoic speech corrupted by noise at different SNR levels with impulse response of testing room. In an effort to minimize the effect of magnitude indeterminacies, the root-mean-square value of all speech signals was equalized to the same root-mean-square value. The results presented in Fig. 5 were obtained at constant distance between the source and the receiver (1 m), for two different SNR values (-5 dB — brighter color, 5 dB — darker color). The test room dimensions was set to $6 \times 3 \times 4$ meters (length $\times$ width $\times$ height), the number of samples of the room impulse response was 1024, high-pass filter was disabled, maximum frequency order was not used, and reverberation time was varied from 0.1 to 0.6 second. All intelligibility measures show the same trend: their value decrease with increasing of reverberation time. This trend is expected because with the increasing value of $\mathrm{RT}_6 0$ is raised also amount of reflected energy which affects the direct part of signal. The PESQ score is inversely proportional to the reverberation time but the sensitivity is not very large. It was changed from 2.42 to 2.28 for 5 dB SNR. Note that the PESQ score was not developed to determine the speech quality in a reverberant environment. But it in [13] a good correlation between this measure and subjective evaluation of intelligibility of reverberant speech was shown. Better sensitivity on change of reverberation time show SNRS which was changed from -2.2 dB to -7.1 dB (5 dB SNR) for analyzed values of RT_{60}. The short-time objective intelligibility measure also exhibits the comparable sensitivity, from 91 % to 81 % for 5 dB SNR. The noise in inputs signals manifests like offset as in the previous experiments. This analysis is closely related to the following analysis because instead of changing reverberation time was varied the distance between the source and receiver. All other parameters necessary for generating room impulse response was the same and the distance was varied from 0.5 m to 3 m. Achieved results are depicted in Fig. 6 and were done for two values of RT_{60}: 300 ms and 600 ms (brighter color). Same as above the root-mean-square value of all speech signals was equalized to the same root-mean-square value to minimize the effect of magnitude indeterminacies. Expected dependence between intelligibility and distance between source and receiver is reduction in intelligibility with increase in distance. This trend can be observed in Fig. 6 so our experiments meet expectations. The decrease is caused by decrease of energy of direct part of signal if the distance is increasing.

Experiment also shows how the reverberation time can affect intelligibility of speech when the distance is changing. For RT_{60} set to 300 ms the separation procedure shows better performance (or intelligibility of separated signal) in comparison with RT_{60} set to 600 ms. This fact confirms our previous experiments (depicted in Fig. 5) which show that intelligibility is inversely proportional to the reverberation time. When we compare all evaluation measures we can see that the most sensitive measures is STOI, but for overall evaluation is it important that all measures show the same trend. The PESQ score is decreasing from 3.06 to 2.41, the SNRS from -1.38 dB to -3.55 dB, the STOI from 96.3 % to 73.3 % for distance varying from 0.5 m to 3 m and RT_{60} set to 300 ms. For higher values of reverberation time are these measures shifted to lower values.

Because of room impulse response generator [12] takes into consideration the microphone type (or microphone directionality) the last experiment was focused on analyzing impact of microphone directionality on speech intelligibility. Different kinds of polar patterns are implemented in this generator and can be chosen using input parameter. The signal attenuation $A(\theta)$, where θ denotes the directional of arrival, is calculated using formula (6). The polar pattern is controlled by parameter ? and is described in Tab. 2. The angle in which the microphone is pointing can be adjusted by external parameter but in our experiments the microphone pointed towards the positive x-axis.

$$A(\theta) = \alpha + (1 - \alpha)\cos\theta\,. \tag{6}$$

The last analyze was done with the same basic parameters as two experiments above, it means: room dimension, number of samples and disabled high-pass filter. The distance between source and microphone was constant and set to one meter, the reverberation time was 300 ms and 600 ms and microphone directionality was varying from 0 to 1. The microphone type can also have impact on overall intelligibility because parameter ?, used to specify microphone directionality, affects amount of reflected energy which is recorded by this microphone. So if the directionality is increasing the amount of reflected energy is also increasing and intelligibility is decreasing. Achieved results are depicted in Fig. 6 for two different RT_{60} values: 300 ms and 600 ms (brighter color). As we could see the RT_{60} had the same effect to the intelligibility as in

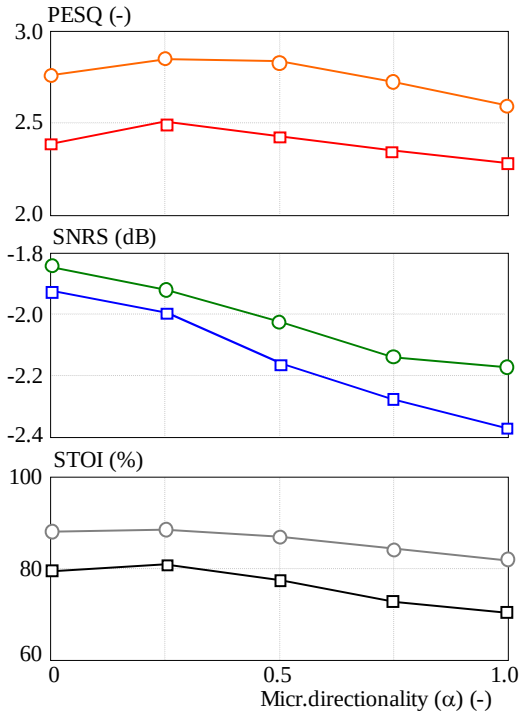

Fig. 7. Intelligibility of IBM-processed mixtures mixed with multi-talker babble at 20 dB as function of microphone directionality for two values of RT_{60}: 300 ms and 600 ms (brighter color)

Table 2. Polar patterns and corresponding values of α for RIR generator

Directivity Pattern	α
Omnidirectional (Monopole)	1
Subcardiod	0.75
Cardiod	0.5
Hypercardiod	0.25
Bidirectional (Dipole)	0

the previous analyze (it exhibits like offset in intelligibility). Based on polar plots the best intelligibility should be achieved in case hypercardiod or cardiod microphone type, because their characteristics have the most directionality. On the other side the monopole should have the worst intelligibility of recorded speech because this type does not take account direction of sound and finally the dipole should be between monopole and cardiod. When we compare these expectations with the achieved results we can see that our expectations are meet only in the case PESQ and STOI. If the intelligibility was measured using SNRS, the speech recorded by dipole showed higher intelligibility than the speech recorded by hypercardiod type of microphone what is different from our expectations. Generally we can say that chosen type of microphone has impact on overall intelligibility of separated speech but its sensitivity is not very strong as we can see in Fig. 6. The top score for PESQ is 2.81 and the lowest score is 2.55, for SNRS the top is -1.9 dB and bottom -2.2 dB and for STOI is the top 89 % and bottom 81 %, all for RT_{60} set to 300 ms. This sensitivity could be also affected using other parameters (room dimension, distance between source and receiver), not only using RT_{60} because all these have impact on reflections.

4 DISCUSSION

The present study extended and replicated the findings and analyses presented in [9]. There are a number of differences and a number of similarities between the procedures and the results of the two studies. The main difference was in evaluation of intelligibility since in this study were used objective methods instead of subjective methods. Generally are these methods less accurate so for higher credibility three different approaches (measures) were used. Although the PESQ measure assesses overall loudness differences between input and processed signal it has been shown good correlation with subjective ratings, *eg* in [14]. Other used measure was STOI also with good correlation with subjective ratings and last one was SNRS. The SNRS was used only for check because it is based on totally different approach, belongs to basic measures and generally is used for measuring speech quality not intelligibility. We have assumed that if the intelligibility performance curves (especially PESQ and STOI) have the same trend than this trend would be similar with the curve achieved using subjective methods.

Experiment 1 verified how value of local threshold affects intelligibility. The results are nearly the same as in similar analysis in [9]. The difference is only in plateau region, in our experiment it was chosen from -10 dB to 5 dB, what could be caused by different data set. The second experiment showed dependence on accuracy of binary mask. The pattern of performance was similar for all three measures and also for different value of noise. Scores were high when binary mask error was near 0 % and dropped thereafter. Next experiment confirmed fact that the IBM technique is more effective when desired speech is masked by speech than when it is masked by noise. It means this technique is more effective by informational masking than by energetic masking. All other experiments were focused to reveal impact of room acoustic on intelligibility performance. Using room impulse generator was created responses for different values of reverberation times (varied from 0.1 s to 0.6 s) and different distances between source and microphone (varied from 0.5 m to 3 m). Achieved results showed how intelligibility performance gradually decreasing with increasing values of these parameters for all three measures. This fact could be caused by dependence of energy of reflected sounds on these parameters. Therefore, for a greater distance or reverberation time the energy of these reflected sounds is increasing and affecting the input signal, what is the reason why the intelligibility decreasing. With energy of reflected sound is also connected directivity of microphone used for recording, what was the last analysis. Outcomes showed that using the microphone with higher directivity positively affects indelibility of separated signal.

5 CONCLUSION

By measuring intelligibility of ideal binary-masked speech (usually corrupted by noise or reflections), we have shown how this intelligibility depends on parameters such as: local threshold value, amount and origin of noise in the input signal, accuracy of the mask estimation, dimensions of room and distance between source and receiver, reverberation time and directionality of microphone. For evaluation of intelligibility we have not chosen subjective measures as it was presented in many previous studies but objective measures, because of these are less time consuming and can be repeatedly utilized for different input data sets (but less accurate). To minimize this disadvantage we have used three different objective metrics: SNRS, PESQ and STOI. For achieving of impact of room acoustics we have used RIR generator and based on desired parameters we have generated impulse responses which have been applied on origin mixtures. By applying IBM processing on this way modified mixtures, we have extended the findings presented in [9] showing similar analysis without acceptance of room acoustic but evaluated using subjective methods.

Acknowledgment

This work is resulting from the project VEGA 1/0987/12 sponsored by Ministry of Education, Slovak Republic. The authors would like to thank Emanuel Habets for making available the source code of RIR generator at his homepage.

References

[1] MOWLAEE, P.—CHRISTENSEN, M—JENSEN, S.: The IEEE Transactions on Audio, Speech and Language Processing.

[2] ASGARI, M.—FALLAH, M.—MEHRIZI, E.—MOSTAFAVI, A.: A VQ-Based Single Channel Audio Separation for Music/Speech Mixtures, Proceeding of International Conference on Computer Modelling and Simulation, Cambridge, UK, 2009, pp. 223–227.

[3] WANG, H.—WANG, Y.—WANG, B.—ZHU, B.—MA, S.: Single Channel Polyphonie Music Signal Separation Based on Bayesian Harmonic Model, Proceeding of International Conference on Image and Signal Processing, Shanghai, 2011, pp. 2784–2787.

[4] LEE, Y.—LEE, I.—KWON, O.: Single-Channel Speech Separation Using Phase-Based Methods, The IEEE Transactions on Consumer Electronics **4** No. 56 (2010), 2453–2459.

[5] ANATHAKRISHNAN, K.—DOGANCAY, K.: Recent Trends and Challenges in Speech-Separation Systems Research – A Tutorial Review, Proceeding of TENCON, Singapore, 2009, pp. 1–6.

[6] KJEMS, U.—BOLDT, J.—PEDERSEN, M.: Role of Mask Pattern in Intelligibility of Ideal Binary-Masked Noisy Speech, Journal of Acoustical Society of America **126** No. 3 (2009), 1415–1426.

[7] SHIRAN, N.—SHALLOM, D.: Enhanced PESQ Algorithm for Objective Assessment of Speech Quality at a Continuous Varying Delay, Proceeding of Workshop on Quality of Multimedia Experience, San Diego, USA, 2009, pp. 157–162.

[8] TAAL, C.—HENDRIKS, R.—JENSEN, J.: An Algorithm for Intelligibility Prediction of Time-Frequency Weifhted Noisy Speech, The IEEE Transactions on Audio, Speech and Language Processing **19** No. 7 (2011), 2125–2136.

[9] LI, N.—LOIZOU, C. P.: Factors Influrncing of Ideal Binary-Masked Speech: Implications for Noise Reduction, Journal of Acoustical Society of America **123** No. 3 (2008), 1673–1682.

[10] COOKE, M.—BARKER, J.: An Audio-Visual Corpus of Speech Perception and Automatic Speech Recognition, Journal of Acoustical Society of America **120** No. 5 (2006), 2421–2424.

[11] PEARCE, D.—HIRS, H.: The AURORA Experimantal Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions, Proceeeding of ICSLP, Beijing, China, 2000, pp. 29–32.

[12] HABETS, E.: Room Impulse Response Generator, Available: http://home.tiscali.nl/ehabets/rir generator.html, Dec. 2013.

[13] KOKKINAKIS, K.—LOIZOU, P. C.: Evaluation of Objective Measures for Quality Assessment of Reverberant Speech, Proceeding of ICASSP, 2011, pp. 2420–2423.

[14] MA, J.—HU, Y.—LOIZOU, P. C.: Objective Measures for Predicting Speech Intelligiility in Noisy Conditions Based on New Band-Importance Function, Journal of Acoustical Society of America **125** No. 5 (2009), 3387–3405.

Received 18 February 2014

Vladimír Sedlák, (Ing) was born in Myjava, Slovak Republic, in 1985. In 2010 he graduated from the Faculty of Electrical Engineering, STU Bratislava. His masters thesis was focused on the neural networks and creation of the simulator for verification their behavior. At the presence he is PhD student at FEI, STU Bratislava. His research interests include the analysis and implementation of algorithms for separation of mixed signals. He is also interested in the area of digital signal processing, embedded systems and human-computer interaction.

Daniela Ďuračková (Prof, Ing, PhD) recived her MSc and PhD degrees from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEEIT SUT) in 1974 and 1981, respectively. Since 1991 she has been an associate professor and since 2005 a professor at the Microelectronics Department (since 2011 the Institute of the Electronics and Photonics) of the FEEIT SUT in Bratislava. The main field of her research and teaching activities has moved from semi-conductor devices towards the design of analog and digital ASICs and neural network implementation on chip.

Roman Záluský (Ing, PhD) was born in Bratislava, Slovakia, in 1983. In 2012 he graduated from the Faculty of Electrical Engineering, STU Bratislava. His doctoral thesis was focused on the neural networks and its hardware implentation. At the presence he is with FEI, STU Bratislava. The aim of his research is hardware implementation of neural networks and digital signal processing.

Tomáš Kováčik (Ing) was born in Brezno, Slovak Republic, in 1987. In 2012 he graduate from the FEI, STU Bratislava. His master's thesis was focused on FPGA and robot control system development. At the presence he is PhD student at FEI, STU Bratislava. His research aims on image recognition and application algorithm of computer vision and image recognition on FPGA. Another part of his research aims to digital signal processing.