PAN ARCHIVES OF ACOUSTICS
Vol. **39**, No. 3, pp. 411–420 (2014)

DOI: 10.2478/aoa-2014-0045

Two-Microphone Dereverberation for Automatic Speech Recognition of Polish

Mikolaj KUNDEGORSKI[(1),(3)], Philip J.B. JACKSON[(2)], Bartosz ZIÓŁKO[(3)]

[(1)] *School of Engineering and Computing Sciences, Durham University*
Durham, UK; e-mail: mikolaj.kundegorski@gmail.com

[(2)] *Centre for Vision, Speech and Signal Processing, University of Surrey*
Guildford, Surrey GU2 7XH, UK; e-mail: p.jackson@surrey.ac.uk

[(3)] *Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology*
al. Mickiewicza 30, 30-059 Kraków, Poland; e-mail: bziolko@agh.edu.pl

(*received November 29, 2013; accepted August 2, 2014*)

Reverberation is a common problem for many speech technologies, such as automatic speech recognition (ASR) systems. This paper investigates the novel combination of precedence, binaural and statistical independence cues for enhancing reverberant speech, prior to ASR, under these adverse acoustical conditions when two microphone signals are available. Results of the enhancement are evaluated in terms of relevant signal measures and accuracy for both English and Polish ASR tasks. These show inconsistencies between the signal and recognition measures, although in recognition the proposed method consistently outperforms all other combinations and the spectral-subtraction baseline.

Keywords: speech enhancement, reverberation, ASR, Polish.

1. Introduction

Speech is the primary method of communication for humans. Over many years of research several successful automatic speech recognition (ASR) systems have been developed, predominantly for the English language (Hinton *et al.*, 2012). They allow hands-free control over a computer (essential for disabled users), text dictation, meeting transcription (judge hearings, interviews), virtual personal assistants, and more. Although ASR performance can be impressive, correctly recognizing spontaneous speech, or speech in an adverse acoustical environment, remains an issue (Fukumori *et al.*, 2013).

ASR performance depends mainly on the quality of the speech data. Alas, in most real-life situations the desired speech is collected with background noise, other speech signals, and reverberation, all of which degrade its intelligibility. In contrast with quasi-stationary background noise that can be modelled during non-speech intervals, the latter two present a greater challenge. Yet, whether interference is present or not, reverberation is present in almost any situation, especially where the microphone is not close to the source. With the growing popularity of hands-free devices and more natural approaches to human-machine interaction, dereverberation becomes a crucial part of any speech enhancement process, in ASR, teleconferencing, and devices for hearing impaired listeners. In an increasing number of these cases, two microphone signals are available for use, which is the scenario we consider here (Li *et al.*, 2012).

It is well-established that information about the spatial location of a target sound source can be of great assistance for enhancement, including under reverberant conditions where the reflections come from random directions. Recent work has shown that the integration of spatial cues (binaural and statistical) can improve the state-of-the-art for separation of speech mixtures, especially under reverberant conditions (Alinaghi, Wang, Jackson, 2011). Here, we investigate the application of this method to dereverberation and its further combination with a model of precedence. The precedence effect (Litovsky *et al.*, 1999) describes the fact that the direct sound from a source arrives earlier at a sensor than along any reflected path, and relates to temporal masking behaviour in natural audition. This property means that any sharp rise in sound en-

ergy (an attack) is likely to have a much higher direct-to-reverberant ratio (DRR) than decaying segments, which the enhancement method exploits by labelling reliable and unreliable components of the signal accordingly.

Our evaluation considers the effect of the proposed enhancement method for two ASR systems: one designed for connected English digits, the other for isolated Polish words. Owing to linguistic differences, the ASR system for the Polish language comprises different algorithms than the more commonplace English language setup (Ziółko *et al.*, 2008). Therefore, it is necessary to investigate the performance of various dereverberation methods on both tasks to verify the generality of our proposed solution. The aim of our work is to test and optimise current speech dereverberation methods on Polish speech datasets and the Polish speech recognizer. We also consider conventional signal measures of enhancement, the segmental signal-to-reverberation ratio (SegSRR), and signal-to-distortion ratio (SDR).

This paper is organised as follows. Next in this section, the problem of reverberation is defined and current dereverberation solutions are reviewed. In Sec. 2, chosen algorithms are described. An experimental procedure and evaluation methods are presented in Sec. 3. The obtained results are shown in Sec. 4, then conclusions are drawn in the final section.

1.1. Reverberation

A sensor (microphone) located at some distance from the source receives a delayed and attenuated version of the desired speech $s(t)$. After the initial time delay gap (ITDG, 5–25 ms), early reflections arrive causing distortions. Late reflections are much weaker but because of their temporal longevity, they can smear the spectrum over the following phonemes (Jeub *et al.*, 2010). The spectrograms in Fig. 1 show how reverberation affects the speech spectrum over time.

We can describe reverberant speech $x(t)$ as a filtered version of clean speech $s(t)$:

$$x(t) = s(t) * h(t) \tag{1}$$

where $*$ denotes convolution and $h(t)$ is a room impulse response (RIR) describing the instantaneous state of an acoustical channel between a source and a sensor.

The RIR can be characterised by parameters dependent on the physical properties of the room and the location of both the source and sensor. These are the direct-to-reverberant energy ratio (DRR), and the reverberation time RT_{60}, by which sound reflections decay by 60 dB below the direct signal level. However, in real-life cases, the RIR has a more complex structure and a non-stationary tail, since the acoustic channel is altered with every small motion of the speech source and the air in the room (Gomez, Kawahara, 2010).

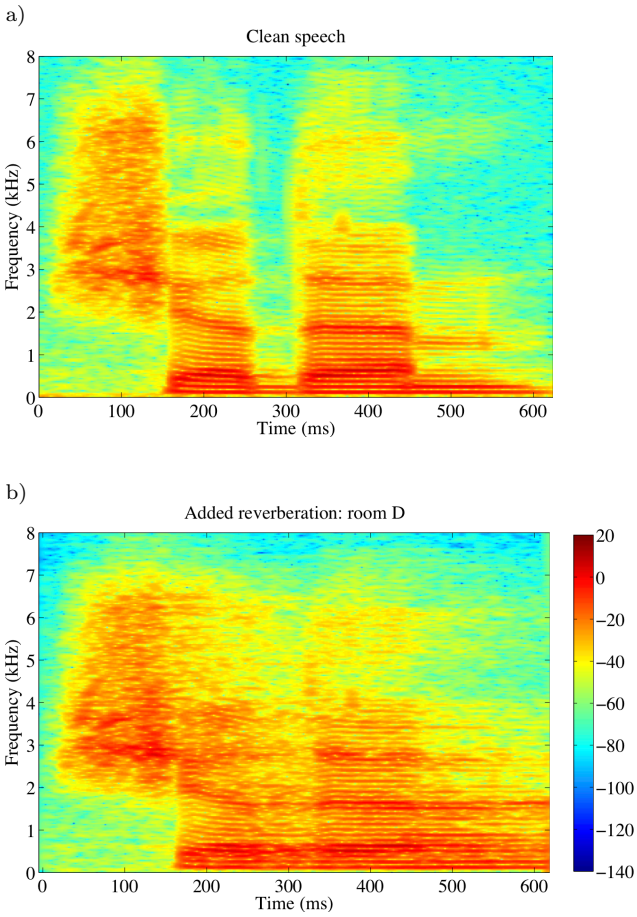

Fig. 1. Spectrograms of the Polish word "siedem" (IPA: ɕ ɛ dɛ m), a) close-microphone recording with anechoic BRIR, and b) with 0.89 s reverberation time and 6.12 dB DRR (room D).

For humans, reverberation is a natural property of speech and does not decrease intelligibility unless it is severe (Wu, Wang, 2006; Drgas *et al.*, 2008). Moreover, moderate reverberation provides reinforcement of the desired signal and spatial information needed for its separation from interference (Hartmann, 1999). On the contrary, for the ASR system, any amount of reverberation is harmful because of test and training conditions mismatch and the degradation of the acoustic features. When a sensor is placed directly by a source reverberation is insignificant. However, in both telecommunication and ASR, growing popularity of more convenient distant talking devices creates a demand for developing solutions for dereverberation.

1.2. Dereverberation methods

Dereverberation can be conducted before or during an ASR process. There are many successful designs focusing on the optimisation of dereverberation parameters based on feedback obtained from ASR results (Gomez, Kawahara, 2010; Seltzer *et al.*, 2004).

However, due to technical limitations we can implement only a standalone dereverberation system. Multi-microphone systems allow us to perform a spatial filtering of the signal, based primarily on the differences in its time of arrival to each sensor. Microphone arrays with an adaptive filter are a primary method for noise and reverberation removal in a car environment (Chien, Lai, 2005). A similar multi-microphone system supported by the neural network, designed and tested for robust ASR, was presented (Pearson *et al.*, 1996). Different methods have been designed to attenuate interference over a wide frequency range (Ward *et al.*, 2001), or comprise an additional speech enhancement method (Shi, Aarabi, 2003; Seltzer *et al.*, 2004). Their main disadvantage is that they comprise a sound capturing system, which makes them hardware dependent. In order to recover a clean signal from reverberant speech, an inverse filter can be estimated. A summary of methods employing linear prediction or blind estimation of RIRs may be found in (Naylor, Gaubitch, 2005). An interesting approach is to utilise the harmonicity of speech to provide more accurate RIR estimation for inverse filtering (Nakatani *et al.*, 2007). However, to provide accurate results, this approach requires a long recording of speech (at least 15 seconds of 16 kHz recording (Wu, Wang, 2006)), because of the high order of the RIR. Owing to the stochastic nature of the RIR tail, the inverse filtered speech still contains the late reverberation component and requires further processing.

To address precisely both the early and late parts of reverberation, a two-stage algorithm can be used. Solutions developed for background noise attenuation can be used to suppress the late component. Spectral subtraction is one of these methods used jointly with linear prediction (Wu, Wang, 2006; Krishnamoorthy, Prasanna, 2009) or Wiener adaptive filtering (Jeub *et al.*, 2010) to reverse the effect of early speech reflections. In our research we employed an implementation of spectral subtraction (Wu, Wang, 2006).

The present research is not only focusing on developing an effective dereverberation method but also on making it possible to integrate this method into a more complex system of speech enhancement. Therefore, it was deemed appropriate to test source separation methods for dereverberation.

The precedence effect is a phenomenon in the human auditory system that plays a role in speech source localisation. It is based on the perceptual emphasis of the first wave front and has been implemented to both dereverberate and separate speech signals (Palomaki *et al.*, 2004). We tested and implemented a precedence model developed by (Hummersone *et al.*, 2010).

The computationally complex, yet very promising method of source separation (Alinaghi *et al.*, 2011) has been suggested as a unification of two effective methods based on the statistical modelling of sources (Sawada *et al.*, 2007); Mandel *et al.*, 2010) and the Expectation-Maximization algorithm. This method was included in our tests, with reverberation being modelled as an additional interfering source (i.e., as a "garbage" source).

2. Enhancement methods

Because of a limitation in terms of the capturing system, integration with ASR, and length of test recordings many aforementioned methods are not suitable for our investigation. In our research, three dereverberation methods were applied to the time-frequency-domain input, $X(t,\omega)$, calculated by short-time Fourier transform (STFT). The first method, spectral subtraction (SpecSub), is a version of the well-known technique to suppress elements close to or below a spectral estimate of the noise level, which is adapted for reverberation noise arising from late reflections. The second method, the precedence mask (PrecMask) employs a similar criterion to identify reliable and noise-corrupted elements but applies it as a binary mask on the input. The tird, spatial and statistical cues, are combined in the binaural-BSS method (Alinaghi) which exploits the directional coherence of direct sound from a located source *versus* diffuse, incoherent reverberation. The two input signals are used to classify the dominant source in each time-frequency element, or cell, and thereby generate the mask. The second and the third method (PrecMask-Alinaghi) have potential to deal with early reflections and we also investigate their concatenation. The formulation of these methods is now described.

2.1. Spectral subtraction

In an enclosed acoustical environment, such as a room, studio, theatre, or transport station, the impulse response from source to receiver typically contains a pulse for direct sound, then multiple pulses for early reflections from the walls, floor, and other surfaces followed by reverberation, a dense congregation of late reflections from all directions. For enhancement, the reverberation can be modelled as an uncorrelated noise process, whose energy is subtracted from that of the corrupted speech. Traditionally (Boll, 1979), negative values are replaced by zero to ensure all energies are non-negative, and the noise spectrum is estimated during non-speech activity; this is unsuitable for reverberation which spills into the silences and is time-varying with the source signal.

The spectral-subtraction implementation we utilised (Wu, Wang, 2006) either attenuates or zeroes each time-frequency cell according to the criterion which we express in general form as:

$$|X(t,\omega)|^p \gtrless gh(t) * \left|\widetilde{X}(t,\omega)\right|^p, \tag{2}$$

where exponent $p = 2$, g denotes a gain applied to the reverberation, here the squared magnitude of the spectrogram elements constitute the input $\left|\widetilde{X}(t,\omega)\right|^2 = |X(t,\omega)|^2$, and $h(t)$ is a causal low-pass filter that smoothes the signal energy in each frequency bin:

$$h(t) = \frac{t-\tau+\alpha_S}{\alpha_S^2}\exp\left(-\frac{(t-\tau+\alpha_S)^2}{2\alpha_S^2}\right) \tag{3}$$

where the reverberation lag τ and time constant α_S give the form of a Rayleigh pdf. The output is produced

$$|S(t,\omega)|^p = \begin{cases} |X(t,\omega)|^p - gh(t) * \left|\widetilde{X}(t,\omega)\right|^p & \text{if (2)}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

A speech signal enhanced using spectral subtraction is shown in Fig. 2a.

2.2. Precedence mask

In human audition, the precedence effect is a temporal masking mechanism that plays a role in source localisation and signal enhancement in a reverberant environment. Localisation is performed by emphasising the first wavefront (assumed to be direct sound) and attenuating signals from different directions (Blauert, 1997).

In an auditory implementation (Hummersone *et al.*, 2010), the magnitude output is as in Eq. (4) with $p = 1$, time constant α_P, $h(t) = At\exp(-t/\alpha_P)$ with normalisation factor A, and $\left|\widetilde{X}(t,\omega)\right|$ being the Hilbert envelope of the Gammatone filterbank signals. The entire procedure is applied to each channel (left and right ear) separately. The enhanced left and right ear signals were used to compute the inter-aural correlation which was thresholded to form a mask. Here, we directly obtained the mask by applying criterion (2) with these definitions of p and $h(t)$, and $\left|\widetilde{X}(t,\omega)\right|$ being the Hilbert envelope of the frequency bin $X(t,\omega)$.

The STFT was used with a Hann window, frame size of 512 samples at 16 kHz (32 ms), and 50% overlap between frames, which slightly differs from the values for the binaural cues and the BSS method. One important feature of the precedence effect algorithm is its computational simplicity which translates into very short processing time needed in real-time applications. Figure 2b shows a spectrogram of speech enhanced using the precedence mask. A binary version of the estimated precedence mask was also considered,

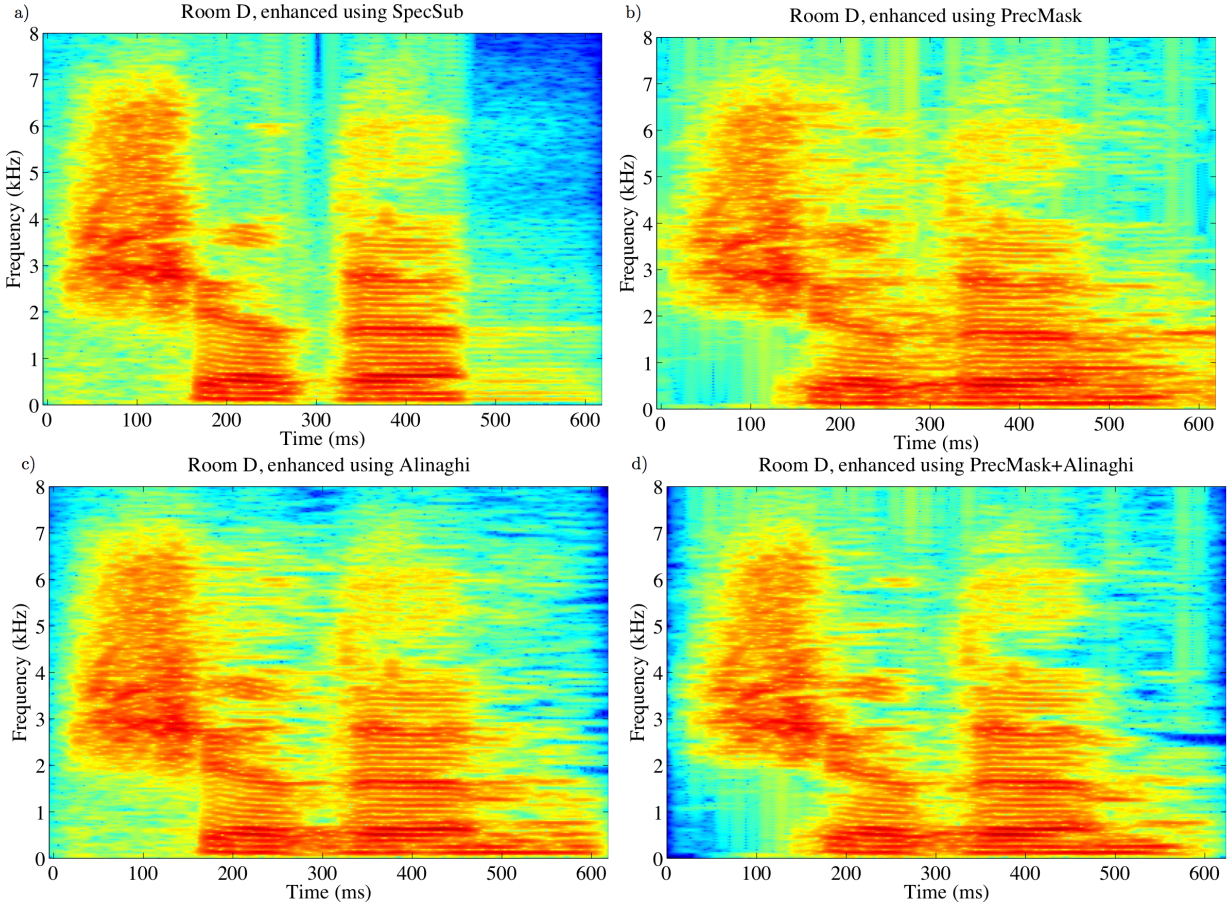

Fig. 2. Spectrogram of the Polish word "siedem" (IPA: ɕ ɛ dɛ m) with added reverberation (room D) after enhancement using: a) SpecSub and b) PrecMask, c) Alinaghi, d) PrecMask+Alinaghi. See Fig. 1 for clean speech and reverberant input to enhancement.

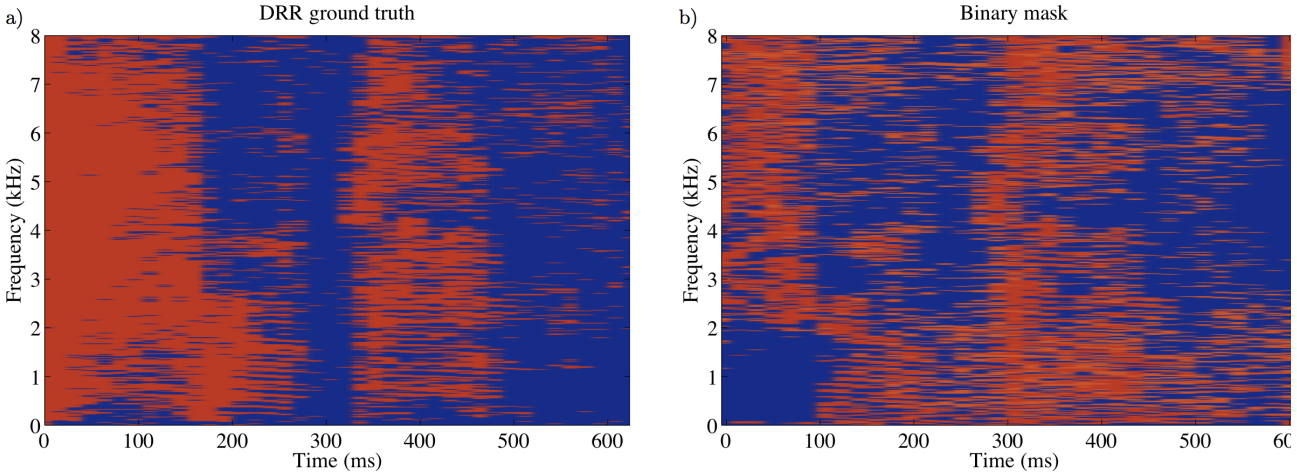

Fig. 3. Ground truth (a) from comparison of the direct and reverberant part of a signal and a binary mask (b) applied to reverberant speech (Room D), Polish word "siedem" (IPA: ɕ ɛ dɛ m). See Fig. 1 for clean speech and reverberant input to enhancement.

as illustrated in Fig. 3. Also shown is a mask based on the corresponding ground truth DRR, whose direct and reverberant signals were formed by convolving the source signal with the early and late parts of the RIR respectively (break point placed 5 ms after arrival of the first impulse). Time-frequency cells are marked positive where the direct signal energy is higher then reverberant energy.

We modified parameters g and α_P of the precedence model to achieve an optimal performance on the target system (Polish ASR) during the preliminary tests. It was found that fixed values other than those suggested by Hummersone (Hummersone *et al.*, 2010) can provide better results. In our optimisation both parameters were increased for the 16 kHz signals, gain factor by 6% to $g = 0.825$, and the time constant by 100% to $\alpha_P = 12.5$.

2.3. Binaural cues and blind source separation

Reverberation can be treated as an additional interfering source in the source separation method. We investigate a stereo source separation method (Alinaghi *et al.*, 2011). This method combines two different approaches, based on computational auditory scene analysis (CASA) and blind source separation (BSS), within a recursive optimisation procedure. Both of them perform the automatic assignment of time-frequency units of the speech mixture spectrogram to their corresponding sources and rely on the assumption of speech's sparseness in the STFT domain. The method is designed to work on signals from two microphones but can separate even more sources for application in undetermined cases. In our case, the input signals are due to a single speech source located in front of sensors and reverberation from binaural RIRs (BRIRs), which is modelled as an additional garbage source. Processing takes place in the time-frequency domain. STFT with a frame size of 1024 samples (64 ms) and 25% overlap is used to transform the 16 kHz signals.

The human auditory system, as well as CASA algorithms based on it, depends on interaural time and level differences (ITD and ILD) as primary cues to estimate the source location (Hartmann, 1999). These are binaural cues, and for the signal from the left $X_L(\omega, t)$ and right $X_R(\omega, t)$ channel in time-frequency domain they are described as

$$\frac{X_L(\omega,t)}{X_R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}, \tag{5}$$

where ω denotes frequency, t is time, and $\alpha(\omega, t)$ is time-and-frequency dependent ILD and $\phi(\omega, t)$ time-and-frequency dependent ITD. In the preliminary test, we found that grouping units across time and frequency (i.e., assuming frequency independent binaural cues) gave better results, therefore we used it in the following experiments.

The BSS method models the mixing of speech into tiles, each dominated by a single source, by 2 dimensional (number of sensors) vectors $\mathbf{h}_j = [h_{jL}, h_{jR}]^{\mathrm{T}}$:

$$\mathbf{X}(\omega,t) = \sum_{N \in h=1} \mathbf{h_j} S_j(\omega,t) \approx \mathbf{h_j} S_j(\omega,t), \tag{6}$$

where $\mathbf{X}(\omega, t) = [X_L(\omega, t), X_R(\omega, t)]^{\mathrm{T}}$. In this method, the time alignment of recovered sources in each frequency bin is not preserved. This ambiguity is resolved during a first iteration of the procedure, using only binaural cues.

We modelled all parameters (ILT, ITD, $\mathbf{h_j}$) as a Gaussian mixture for every time-frequency unit. The two-step Expectation-Maximization algorithm is used to find an optimal solution. In the expectation step (E-step) variance and mean of the mixture are estimated

and log likelihood of the observations is maximised in the second stage (M-stage).

In our experiments, we investigated the separate effects of the three methods, hereafter termed SpecSub, PrecMask, and Alinaghi respectively, and the combination of the precedence mask with the Alinaghi binaural-BSS method, PrecMask+Alinaghi, to see whether pre-screening of the input offered any advantage in the exploitation of spatial cues.

3. Experimental procedure

Figure 4 shows an overview of the speech enhancement front ends that were tested. To obtain results relevant for a wide range of acoustical conditions, binaural room impulse responses from five listening environments were utilised. The performance of each method was evaluated in four ways: using two signal measures, segmental signal-to-reverberation ratio (SegSRR) and signal-to-distortion ratio (SDR), and two recognition rates, on English connected digits and Polish isolated words. Details of those parts of the experimental method are now described in turn.

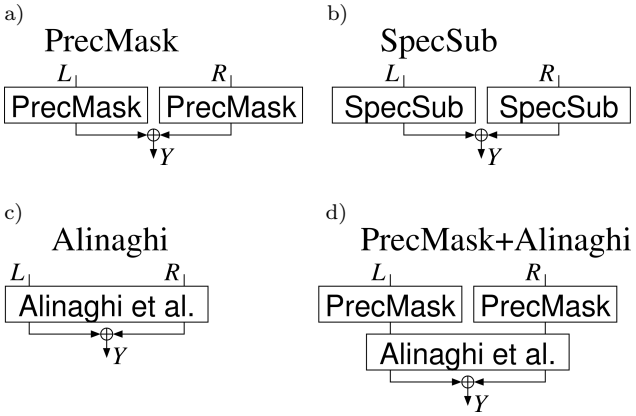

Fig. 4. Block schema of the speech enhancement methods applied prior to ASR.

3.1. Reverberant environment reproduction

To create reverberant datasets from clean speech recordings we convolved them with BRIRs from (Hummersone *et al.*, 2010). This method provides an imitation of a reverberant environment closer to reality than room simulation based on physical models. Table 1 shows the acoustic parameters of each room sorted by reverberation time, which ranges from that of a small office to what could be a lecture hall. Room X is an anechoic situation which comprises filtering introduced by sound capturing equipment only, and it serves as the reference signal. In our experiments, we used the zero azimuth BRIRs, such that the source was situated directly in front of the microphones. It can be assumed that at least for source separation methods this is the worst case scenario.

Table 1. Room acoustical properties (Hummersone *et al.*, 2010.)

Room	ITDG [ms]	DRR [dB]	RT_{60} [s]
A	8.72	6.09	0.32
B	9.66	5.31	0.47
C	11.9	8.82	0.68
D	21.6	6.12	0.89

3.2. Signal-related measurements

Two signal-related measures were used to evaluate the algorithms' performance. In both cases, the reference signal $s(n)$ was the clean speech recording convolved with the anechoic BRIR (room X). The analysed signal $\widetilde{s}(n)$ denotes the reverberant or enhanced speech. Before processing the input signals, they were aligned and normalised. The segmental signal-to-reverberation ratio (SegSRR) is designed to provide a summary statistic of the relative proportions of clean signal energy to that of the reverberant. It is based on (Krishnamoorthy, Prasanna, 2009) and is calculated for each of the 512-sample frames F_i, and averaged over all frames

$$\mathrm{SegSRR}(F_i) = 10\log_{10}\frac{\sum\limits_{n\in F_i} s(n)^2}{\sum\limits_{n\in F_i}(s(n)-\widetilde{s}(n))^2}\,. \tag{7}$$

The SegSRR value provides information about the amount of reverberant energy in a signal, suitable for quantitative comparison.

To estimate distortions, the modified signal-to-distortion ratio (SDR) method was used, as presented in (Vincent *et al.*, 2006). Its goal is to exclude all energy components that could have been introduced to $\widetilde{s}(n)$ by filtration of the reference signal $s(n)$ (e.g. by the RIR). By applying the 512-tap FIR Wiener filter to $\hat{s}(n)$, the signal $s_{\mathrm{wien}}(n)$ that best approximates $s(n)$ is obtained, and the SDR calculated:

$$\mathrm{SDR} = 10\log_{10}\frac{\mathrm{Var}(s_{\mathrm{wien}}(n))}{\mathrm{Var}(s_{\mathrm{wien}}(n)-\widetilde{s}(n))}\,. \tag{8}$$

To avoid the results being biased by the character of the speech signals used, a test was conducted on 20 varied recordings, sampled 16 kHz from both English (Garofolo *et al.*, 1993) and Polish corpora (Grocholewski, 1998). Results are presented for the rooms A, B, C, and D.

3.3. HTK based test

The Aurora digit recognition task (Pearce, Hirsch, 2000) has been used to evaluate dereverberation methods on English speech. The HTK software (version 3.4.1) (Young *et al.*, 2006) was employed

as a recognition tool. Each digit from the TIDigits database (Leonard, Doddington, 1993) is modelled as a whole word hidden Markov model (HMM) with 16 states and 3 gaussians per state. Feature vectors consist of 12 cepstral coefficients and the logarithm frame energy, with corresponding delta and acceleration values, a total of 39.

The test, primarily constructed to compare noise removal algorithms, consists of recordings from the TIDigits database. All recordings are sampled at 8 kHz and shaped for teletransmission with the G.712 characteristics. There are 8440 utterances spoken by 55 female and 55 male speakers used in the ASR training. A further 1001 utterances spoken by a distinct set of 52 female and 52 male speakers are employed for testing.

The system was trained on recordings convolved with X (anechoic) BRIR and processed by an evaluated method. It provides matched training and test conditions in terms of a common reference signal.

3.4. Polish ASR test

Our target system was the ASR system SARMATA, designed at the AGH in Krakow (Ziółko *et al.*, 2011), dedicated to the Polish language, incorporating acclaimed solutions invented for other languages, as well as novel methods, tailor-made for modelling Polish speech. Being under development, it is able to recognize precisely enunciated isolated utterances and is employed in commercial applications.

There are many differences between Polish and English speech which create the need for some differences in the methods (Ziółko *et al.*, 2008). The Polish language is strongly based on Latin. It has a complex grammar, and the exact meaning of the words generally depends on morphology. A single word may have up to several hundreds of derived forms topically correlated. In English, the position in the sentence is more important. Many combinations of different words may have a similar pronunciation, which is rare in Polish. Moreover, Polish speech contains a lot more plosive (e.g. [p], [k]) and fricative (e.g. [s], [z]) consonants and very high-frequency phones not existing in English, which may sound to non-Polish speakers almost like rustle or hum.

Speech parameterisation in SARMATA is based on discrete wavelet transforms, and multithreading is introduced to improve time efficiency – a key factor in real-time performance of a large vocabulary recognizer (LVR). Mayer wavelet decomposition is used with the designed perceptual tree to provide a psychoacoustic frequency characteristic. Parameter vectors are classified using a modified k-NN algorithm. The search for the closest recognition hypothesis is performed by the Viterbi algorithm with the n-gram model applied to discriminate unlikely connections between words.

Word recognition rate (WRR) has been used to evaluate the performance of the ASR on enhanced speech. Polish language recordings are 16 kHz isolated words from the CORPORA database (Grocholewski, 1998), containing mainly first names. In the training, 1830 words spoken by five speakers and convolved with X (anechoic) BRIR were used. We used the recordings of one other speaker including 217 words in the test.

4. Results

In the first stage of our experiment, we focused on the optimisation of the precedence effect method. The outcome is presented in Subsec. 4.1. The results of the comparison between the optimised precedence effect (PrecMask) and other methods are shown in Subsec. 4.2. The tables at the end of the section summarise the presented results.

4.1. Precedence effect optimization

The precedence effect method was tuned to optimize results of the Polish ASR. Starting from Hummersone's optimal parameters, different values of gain g and time α_p were tested on standard and binary precedence masks. In the optimal setup (PrecMask), the binary mask was applied and gain g was increased by 7% and time α_P by 50 %.

The increase of parameter g improved performance not only in environments with the highest DRR (room A and C), but also in room B with the lowest value. The longer time constant α_P clearly influenced effectiveness in room D (with the longest reverberation time). However, further increase of these parameters gave no overall improvement indicating the limitation of the method.

The binary mask invariably gave better recognition results in all tested environments in every test. Despite the fact that only Polish ASR results were taken into consideration when adjusting the parameters, improvement was present in all evaluation methods.

For all the methods except Polish ASR, there is apparent deterioration for all variations of the precedence effect for room B. The results pattern for the English language test is more similar to signal-related measurements than the Polish test, which indicates the importance of differences in the design.

In comparison to Hummersone's precedence effect method, PrecMask gave 0.5 dB improvement in SegSRR, 4.7 dB in SDR, 10.2% higher recognition rate on Aurora test and, 9.5% on Polish ASR.

4.2. Comparison between the different dereverberation methods

Three different dereverberation methods: spectral subtraction (SpecSub), the precedence effect (Prec-

Mask), and the separation method of binaural cues and blind source separation (Alinaghi), as well as a hybrid method consisting of PrecMask and Alinaghi (PrecMask+Alinaghi) were compared in the final test. Summarised results are presented in tables: for SegSRR (Table 2), SDR (Table 3), HTK based tests (Table 4), and Polish recognizer (Table 5).

Tables 2 and 3 show the signal related measurements of the algorithms' performance. For SegSRR SpecSub result is distinctly the best, and Alinaghi have the worst score, which reflects the fact that SpecSub is designed to cope with reverberation and background noise, while Alinaghi mainly attenuates interfering sources. In terms of SDR, the best result was obtained using Alinaghi, almost recovering the result of reverberant speech, which means that the distortions introduced by this method are minimal. The precedence effect PrecMask was the second best, but the hybrid method PrecMask+Alinaghi, scored even lower than SpecSub.

Table 2. SegSRR (dB) for all methods. The best scores in each acoustical condition and overall (mean across rooms A, B, C, and D) are shown in bold. The anechoic case (room X) is given for reference.

Method	Room					mean
	X	**A**	**B**	**C**	**D**	
None	∞	0.56	−1.48	−0.63	−0.69	−0.56
SpecSub	9.04	**2.38**	**1.18**	**1.65**	**2.06**	**1.82**
PrecMask	8.48	0.67	−0.97	−0.04	−0.52	−0.22
Alinaghi	**13.97**	0.89	−1.16	−0.21	0.23	−0.07
PrecMask +Alinaghi	7.65	0.82	−0.77	0.25	0.19	0.13

Table 3. SDR (dB) for all methods. The best scores in each acoustical condition and overall (mean across rooms A, B, C, and D) are shown in bold. The anechoic case (room X) is given for reference.

Method	Room					mean
	X	**A**	**B**	**C**	**D**	
None	∞	13.30	7.24	11.30	5.21	–
SpecSub	10.60	9.09	7.19	8.85	**7.16**	8.58
PrecMask	11.99	10.61	7.02	9.77	5.42	8.96
Alinaghi	**17.65**	**12.68**	**7.34**	**11.11**	6.53	**11.06**
PrecMask +Alinaghi	11.15	9.41	6.62	8.74	5.97	8.38

Table 4 shows the recognition ratio on the HTK recognizer. Again, the shape of the chart is similar to the SDR results, but the ranking of the methods is different. All methods score very similarly except for Alinaghi with a recognition even worse than the reverberant speech. The leading method PrecMask+Alinaghi provides an overall 20% increase in recognition over reverberant speech.

Table 4. SRR (%) on English connected digits. The best scores in each acoustical condition and overall (mean across rooms A, B, C, and D) are shown in bold. The anechoic case (room X) is given for reference.

Method	Room					mean
	X	**A**	**B**	**C**	**D**	
None	**97.2**	79.2	47.2	68.9	28.7	64.2
SpecSub	96.7	85.8	72.2	87.4	67.7	82.0
PrecMask	96.0	87.3	76.2	89.7	58.0	81.4
Alinaghi	97.1	63.3	39.9	76.6	39.6	63.3
PrecMask +Alinaghi	95.2	**87.4**	**78.0**	**90.6**	**74.9**	**85.2**

On the Polish recognizer (Table 5), SpecSub has an average recognition of 6.5% higher than reverberant speech, but in small rooms (X and A) its performance is worse than the reverberant speech. PrecMask gives better results, a further 2.7% increase over the baseline, scoring under SpecSub only in room D, with the highest RT_{60}. The source separation method Alinaghi gives a 10.7% improvement in recognition, and by applying preprocessing by the precedence effect (PrecMask+Alinaghi) a further 3.3% growth is observed, especially in rooms C and D.

Table 5. WRR (%) on Polish isolated words. The best scores in each acoustical condition and overall (mean across rooms A, B, C, and D) are shown in bold. The anechoic case (room X) is given for reference.

Method	Room					mean
	X	**A**	**B**	**C**	**D**	
None	**82.5**	65.4	63.2	44.9	41.9	59.6
SpecSub	67.3	63.6	73.3	64.5	61.8	66.1
PrecMask	78.3	**74.2**	**77.0**	68.2	46.1	68.8
Bina+BSS	78.8	**74.2**	75.2	64.5	59.0	70.3
PrecMask +Alinaghi	78.8	72.4	76.1	**72.8**	**68.2**	**73.6**

The hybrid method PrecMask+Alinaghi improves ASR by 14%, from 59.6% on reverberant speech to 73.6%. The most notable enhancement, at the rate of 26.3%, is achieved for room D. This combination proved to be better than other methods in high RT_{60} rooms C (5% better than the next method), and D (9%).

5. Conclusions

This report presents our work on testing and developing dereverberation techniques suitable for Polish ASR. It has been shown that the Polish recog-

nizer's performance under reverberant conditions differs from that observed for the HTK based recognizer. Especially, in the room with the lowest DRR, the Polish ASR performance is much better than when using the English solution. Moreover, we identified that an alternative approach is needed in the case of both preparing training datasets and speech enhancement. The distortion of signal introduced by dereverberation methods (especially SpecSub) is substantial for Polish ASR and a bypass of the enhancement method for clean signal should be introduced in any working solution. Finally, using our hybrid method consisting of the precedence effect and source separation algorithms (PrecMask+Alinaghi) we acquired a significant improvement in ASR, including most challenging cases with severe reverberation.

By testing with different evaluation methods, we learned that neither of the signal related measurements, by themselves, provides information about how the method would perform in ASR. Though, in the case of comparing the variation of one method (PrecMask), the upswing in the signal related metrics also appeared in the ASR based test result. It leads to the conclusion that, in the evaluation of algorithms targeting ASR, different testing methods should be used, but only test on the target platform guarantees desired enhancement.

Adding a background noise attenuation method to the developed dereverberation system is the next step to produce a complete speech enhancement front end to the Polish ASR. Further research can be done by incorporating the precedence effect inside the Alinaghi recursive method to improve source separation and dereverberation.

Acknowledgments

This work was supported by LIDER/37/69/L-3/11/NCBR/2012 grant.

References

1. Alinaghi A., Wang W., Jackson P.J.B. (2011), *Integrating binaural cues and blind source separation method for separating reverberant speech mixtures*, [in:] Proc. of ICASSP, Prague, pp. 209–212.
2. Blauert J. (1997), *Spatial Hearing: The Psychophysics of Human Sound Localization*, 2nd Edition, MIT Press.
3. Boll S.F. (1979), *Suppression of acoustic noise in speech using spectral subtraction*, Acoustics Speech and Signal Processing, IEEE Trans., **27**, 2, 113–120.
4. Chien J.T., Lai P.Y. (2005), *Car speech enhancement using a microphone array*, Int. Journal of Speech Technology, **8**, 1, 79–91.
5. Drgas S., Kociński J., Sek A. (2008), *Logatom articulation index evaluation of speech enhanced by blind source separation and single-channel noise reduction*, Archives of Acoustics, **33**, 4, 455–474.
6. Fukumori T., Nakayama M., Nishiura T., Yamashita Y. (2013), *Estimation of speech recognition performance in noisy and reverberant environments using pesq score and acoustic parameters*, [in:] Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, pp. 1–4.
7. Garofolo J.S., Lamel L.F., Fisher W.M., Fiscus J.G., Pallett D.S., Dahlgren N.L., Zue V. (1993), *Timit acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, Philadelphia.
8. Gomez R., Kawahara T. (2010), *Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood*, Audio, Speech and Language Processing, IEEE Trans., **18**, 7, 1708–1716.
9. Grocholewski S. (1998), *First database for spoken polish*, [in:] Proc. of International Conference on Language Resources and Evaluation, Grenada, pp. 1059–1062.
10. Hartmann W.M. (1999), *How we localize sound*, Physics Today, **52**, 11, 24–29.
11. Hinton G., Deng L., Yu D., Dahl G., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Kingsbury B. (2012), *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal Processing Magazine, **29**, 6, 82.
12. Hummersone C., Mason R., Brookes T. (2010), *Dynamic precedence effect modeling for source separation in reverberant environments*, Audio, Speech, and Language Processing, IEEE Trans., **18**, 7, 1867–1871.
13. Jeub M., Schafer M., Esch T., Vary P. (2010), *Model-based dereverberation preserving binaural cues*, Audio, Speech, and Language Processing, IEEE Trans., **18**, 7, 1732–1745.
14. Krishnamoorthy P., Prasanna S. (2009), *Reverberant speech enhancement by temporal and spectral processing*, Audio, Speech, and Language Processing, IEEE Trans., **17**, 2, 253–266.
15. Leonard R.G., Doddington G. (1993), *Tidigits*, Linguistic Data Consortium, Philadelphia.
16. Li K., Guo Y., Fu Q., Yan Y. (2012), *A two microphone-based approach for speech enhancement in adverse environments*, [in:] Consumer Electronics (ICCE), 2012 IEEE International Conference, pp. 41–42.
17. Litovsky R.Y., Colburn H.S., Yost W.A., Guzman S.J. (1999), *The precedence effect*, J. Acoust. Soc. Am., **106**, 1633–1654.
18. Mandel M.I., Weiss R.J., Ellis D. (2010), *Model-based expectation-maximization source separation and localization*, Audio, Speech, and Language Processing, IEEE Trans., **18**, 2, 382–394.
19. Nakatani T., Kinoshita K., Miyoshi M. (2007), *Harmonicity-based blind dereverberation for single-channel speech signals*, Audio, Speech, and Language Processing, IEEE Trans., **15**, 1, 80–95.

20. Naylor P.A., Gaubitch N.D. (2005), *Speech dereverberation*, [in:] Proc. of Int. Workshop Acoust. Echo Noise Control, Eindhoven.

21. Palomaki K.J., Brown G.J., Wang D. (2004), *A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation*, Speech Communication, **43**, 4, 361–378.

22. Pearce D., Hirsch H. (2000), *The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions*, [in:] ISCA ITRW ASR., pp. 29–32.

23. Pearson J., Lin Q., Che C., Yuk D.S., Jin L., de Vries B., Flanagan J. (1996), *Robust distant-talking speech recognition*, [in:] Proc. of ICASSP, Atlanta, **1**, 21–24.

24. Sawada H., Araki S., Makino S. (2007), *A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures*, [in:] Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 139–142.

25. Seltzer M.L., Raj B., Stern R.M. (2004), *Likelihood-mazimizing beamforming for robust hands-free speech recognition*, Speech and Audio Processing, IEEE Trans., **12**, 5, 489–498.

26. Shi G., Aarabi P. (2003), *Robust digit recognition using phase-dependent time-frequency masking*, [in:] Proc. of ICASSP, Hong Kong, pp. 684–687.

27. Vincent E., Gribonval R., Fevotte C. (2006), *Performance measurement in blind audio source separation*, Audio, Speech, and Language Processing, IEEE Trans., **14**, 4, 1462–1469.

28. Ward D.B., Kennedy R.A., Williamson R.C. (2001), **Constant directivity beamforming**, [in:] Microphone Arrays, Springer-Verlag.

29. Wu M., Wang D. (2006), *A two-stage algorithm for one-microphone reverberant speech enhancement*, Audio, Speech, and Language Processing, IEEE Trans., **14**, 774–784.

30. Young S. J., Kershaw D., Odell J., Ollason D., Valtchev V., Woodland P. (2006), *The HTK Book Version 3.4*, Cambridge University Press.

31. Ziółko B., Manandhar S., Wilson R.C., Ziółko M., Gałka J. (2008), *Application of htk to the Polish language*, [in:] Proc. of International Conference on Audio, Language and Image Processing, Shanghai.

32. Ziółko M., Gałka J., Ziółko B., Jadczyk T., Skurzok D., Masior M. (2011), *Automatic speech recognition system dedicated for Polish*, [in:] Proc. of Interspeech, Florence.