Objective quality evaluation in blind source separation for speech recognition in a real room

Leandro Di Persia [a,b,1,*] Masuzo Yanagida [c]
Hugo Leonardo Rufiner [a,b,1] Diego Milone [b,a,1,2]

[a] *Laboratorio de Cibernética. Facultad de Ingeniería, Universidad Nacional de Entre Ríos, C.C. 47 Suc. 3 - 3100 Paraná, Argentina.*

[b] *Grupo de Investigación en Señales e Inteligencia Computacional. Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Ciudad Universitaria, C.C. 217 - 3000 Santa Fe, Argentina.*

[c] *Department of Knowldge Engineering, Doshisha University, 1-3, Tatara-Miyakodani, Kyo-Tanabe, 610-0321, Japan.*

Abstract

The determination of quality of the signals obtained by blind source separation is a very important subject for development and evaluation of such algorithms. When this approach is used as a pre-processing stage for automatic speech recognition, the quality measure of separation applied for assessment should be related to the recognition rates of the system. Many measures have been used for quality evaluation, but in general these have been applied without prior research of their capabilities as quality measures in the context of blind source separation, and often they require experimentation in unrealistic conditions. Moreover, these measures just try to evaluate the amount of separation, and this value could not be directly related to recognition rates. Presented in this work is a study of several objective quality measures evaluated as predictors of recognition rate of a continuous speech recognizer. Correlation between quality measures and recognition rates is analyzed for a separation algorithm applied to signals recorded in a real room with different reverberation times and different kinds and levels of noise. A very good correlation between weighted spectral slope measure and the recognition rate has been verified from the results of this analysis. Furthermore, a good performance of total relative distortion and cepstral measures for rooms with relatively long reverberation time has been observed.

Key words: Quality Measures, Blind Source Separation, Robust Speech Recognition, Reverberation.

1 Introduction

Blind source separation (BSS) of sound sources is a technique aiming at recover the signals emitted by some sound sources, from records obtained by remote sensors, without using any information about transfer characteristics or geometrical location of sources and sensors [1]. BSS is a complex process because the ambient may change the sound field due to sensors being remotely located with respect to sources. As a consequence, the received signals at sensors are not only mixed, but also modified in a way that can be assimilated to processing by a linear time-invariant (LTI) system [2]. In free field conditions, impulse response from each source to each sensor would be a delayed impulse, with the amplitude related to energy decay of sound, and delay related to transmission time in the source-sensor path. In a closed environment, however, sound is reflected in all free surfaces, returning to sensors from different directions. So impulse responses have a complex structure, with many impulses located at different delays, corresponding to echoes arriving from different directions. This reverberation phenomenon produces echoes and spectral distortion affecting the spatial perception of sound [2,3,4], intelligibility [5,6] and degrades recognition rates in case of automatic speech recognition (ASR) systems [7], even if the system is trained with reverberant signals recorded in the same room [8].

Quality evaluation of the resulting separated signals is a complex problem that depends on the application field. In some cases, the main interest is not recovering the original signal but preserving some characteristics that are required for the task concerned. For example, when retrieval of a voice to be used in a hearing aid device is desired, perfect reconstruction of the original waveform is not as important as a good perceptual quality. In the same way, for ASR systems, auditory perception is not as important as preserving some acoustic cues that are used by the system to perform the recognition. On the contrary, in other situations the aim is to recover the original signal as exactly as possible, such as a waveform coder. So far, few works have been presented with specific proposals for quality evaluation in the field of BSS.

[*] Corresponding author. Facultad de Ingeniería y Ciencias Hídricas (UNL): Ciudad Universitaria (CC 217), Ruta Nacional N 168 - Km. 472.4, Santa Fe (CP3000), Argentina. Tel.:+54-342-4575245 ext. 145. Fax:+54-342-4575224.

Email addresses: `ldpersia@ciudad.com.ar` (Leandro Di Persia), `myanagid@mail.doshisha.ac.jp` (Masuzo Yanagida), `lrufiner@bioingenieria.edu.ar` (Hugo Leonardo Rufiner), `d.milone@ieee.org` (Diego Milone).

[1] This work is supported by ANPCyT-UNER, under Project PICT 11-12700 and UNL-CAID 012-72

[2] This work is supported by CONICET and ANPCyT-UNL, under Project PICT 11-25984

Particularly, in the context of automatic speech recognition, the only available way to evaluate the performance of some blind source separation algorithm is through a speech recognition test.

The objective of the present work is to find objective quality measures that correlates well with automatic speech recognition rate, when using blind source separation as a mean to introduce robustness into the recognizer. To fulfill this objective, first some set of potentially good measures need to be selected. In the next section a brief review on quality evaluation in the context of BSS and speech processing will be given. Based on this review, in Section 3 specific quality measures will be selected for the evaluation in our experimental framework. Next, a detailed description of the experimental design for determining the relation between speech recognition rates and the obtained measures results will be given. Results and discussion will be presented in Sections 5 and 6 respectively, followed by conclusions in Section 7.

2 Brief review of quality evaluation

2.1 Quality evaluation for BSS

In the particular case of evaluating BSS algorithms, many different alternatives have been used, generally derived from other areas of signal processing. Those methods can be classified into two main areas: *subjective assessment*, where some appreciation is used regarding subjective perceived quality of resulting sound [9,10], or visual differences between waveforms of separated signal and original ones [10,11,12], or visual differences of spectrograms of separated signals and original ones [9]; and *objective evaluation*, where some numerical quantity directly associated to separation quality is used, permitting an objective comparison between different algorithms.

Regarding objective measures that have been applied to BSS problem, these can be divided into three kinds:

(1) Measures that require knowledge about transmission channels: These measures use information about impulse responses between each sound source and each microphone, or require knowledge of individual signals arriving at each microphone. These kinds of measures are hard to apply to realistic environments as they depend on factors that may vary from one experiment to another. Among them, it can be mentioned: Multichannel inter symbol interference (MISI) [1]; Signal to interference ratio (SIR) [13,14,15] and Distortion - Separation [16].
(2) Measures that use information about sound sources: In this case some

measures of discrepancy between the separated signal and the original source signal is used. One drawback of these measures is that, by comparing with original sources, the algorithms that perform separation but not reverberation reduction will yield poorer results as the resulting signal will be always distorted, even for perfect separation. Some measures of this kind commonly used in BSS are: Total relative distortion (TRD), proposed in [17,18], and segmental Signal to noise ratio (segSNR) [15,19].

(3) Indirect measures: In this case the processed signal is used as input to another system with which the result can be evaluated in an objective way. The most typical example of this is an ASR system with which evaluation is made on recognition rate obtained after separation [3] [20,21].

All of these show an important lack of experimentation in the area of quality evaluation for algorithms of BSS in realistic environments. Problems for quality measure proposals came mainly from two aspects that must be taken into account for a correct evaluation of such algorithms: *reality level* required in experiments, which is necessary for the results to be directly extrapolated to practical situations, and *task complexity*, as for example, some BSS algorithms search only for separated signals, while others try to eliminate reverberation effects too. These aspects also need to be considered carefully for choosing a suitable kind of evaluation.

2.2 Quality measures applied to other areas

In applications where the final result will be listened by humans, the ideal way for quality assessment is by means of subjective evaluation of perceptual quality [22]. Many standardized tests allow the evaluation with subjective measures. For example, the composite acceptability (CA) of diagnostic acceptability measure (DAM) [23] consists of a parametric test where the listener has to evaluate acceptability of sound based on 16 categories of quality. Other widely used subjective measure is the mean opinion score (MOS), a measure where each subject has to evaluate the perceived quality in a scale of 1 to 5. This kind of tests has high cost both in time and resources.

Several objective quality measures have been proposed to overcome this drawback [24,25,26]. In [27] the correlation between a large number of objective measures and the subjective measure CA-DAM is studied, evaluated over several speech alterations, contamination noises, filters and coding algorithms. In that work, weighted spectral slope measure (WSS) was the best predictor of subjective quality.

[3] This case is opposite to the objective of the present work, where we are not evaluating the separation itself, but its impact on an ASR system.

Table 1
Performance of quality measures for different tasks (see full explanation in text). Second column: $|\rho|$ DAM, correlation as predictor of CA-DAM. Third column: WER%, word error rate in isolated word recognition. Fourth column: r_{err}, prediction error as predictor of recognition rate in robust continuous ASR

Measure	$\|\rho\|$ DAM	WER%	r_{err}
segSNR	0.77	–	4.71
IS	–	11.35	0.41
LAR	0.62	–	0.88
LLR	0.59	8.45	–
LR	–	8.63	–
WLR	–	9.15	–
WSS	0.74	8.45	1.72
CD	–	8.88	–
LSD	0.60	–	–

In the last years, some objective measures that use perceptual models have been introduced. The first widely adopted was perceptual speech quality measure (PSQM) [28] and more recently, the audio distance (AD) based on measuring normalizing blocks (MNB) was proposed [29]. This measure present a correlation with MOS of more than 0.9 for a wide variety of coding algorithms and languages [30].

Regarding ASR, these measures have been used at two different levels. In template-based isolated word recognizers, a measure of distance between the test signal and the stored templates is needed [31]. Several objective measures originally proposed for speech enhancement or coding have been successfully used within this context [32,33,34]. On the other hand, the capability of those measures to predict recognition rate of a robust speech recognizer has been studied [35].

As a summary of the background available for our work, Table 1 shows a comparison of results obtained by various researchers in different tasks[4]. First column lists the objective quality measures used: segmental signal to noise ratio (segSNR), Itakura-Saito distance (IS), log-area ratio (LAR), log-likelihood ratio or Itakura distance (LLR), likelihood ratio (LR), weighted likelihood

[4] As the application fields and contexts are different respect to this work, these results are not directly applicable to our research, but can give some cues on potentially interesting measures to evaluate.

ratio (WLR), weighted spectral slope (WSS), cepstral distortion (CD), and log-spectral distortion (LSD) [24,25,26,32,36].

The second column presents the absolute value of correlation coefficient between quality measures and subjective test CA-DAM [22], evaluated over a set of speech modified with 322 different distortions. It should be noted that correlation for segSNR was calculated using only a subset of 66 distortions produced by waveform coders, for whom a measure based in waveform similarity has sense. Excluding this case, a high correlation between subjective quality level and WSS can be noted.

The third column presents the percentage of word error rate (WER%) for an isolated word recognition system, in which the measures where applied as selection criteria for classification, for a set of 39 word of a telephone recording database [34]. A good performance for recognizers based on LLR and WSS measures can be noted.

Finally, the fourth column shows the performance of measures as predictors of recognition rate in a continuous speech recognition system using a robust set of features, on speech contaminated with additive noise [35]. The presented value (r_{err}) is the mean squared prediction error, averaged over all sentences in the database of processed speech. In this case the best performance is obtained by LAR and IS measures.

3 Selected measures

As mentioned in Section 1, only objective measures that make use of sound source information will be used in this work. This kind of measures attempt to evaluate some “distance” or “distortion” of separated signal with respect to original signal and have been selected for three reasons. First, by using this approach experiments can be performed with mixtures recorded in real rooms (this gives the experiment a high level of realism) and there is no need to know any information about transmission channels between sources and sensors. Second, as the sources must be available, the experiments could be extended to other mixing conditions. Third, as in general the ASR systems are trained with clean speech, using a method that permits to compare algorithm output with the “ideal” clean one is reasonable.

Based on the analysis presented in Section 2 of previous works, a set of 9 objective quality measures was selected for this study [5]:

[5] Detailed equations and parameters used are listed with unified notation in Appendix A

(1) Segmental signal to noise ratio (segSNR): This measure is included because it is widely used due to its simplicity. Besides this, it has been used in the context of BSS to evaluate separation algorithms, as mentioned in Section 2 [22].
(2) Itakura-Saito distortion (IS): This measure is derived from linear prediction (LP) analysis [37,31]. Its good performance as predictor of recognition rate for signals with additive noise in continuous speech recognition systems makes this measure a good candidate for the present research.
(3) Log-area ratio distortion (LAR): It is also derived from LP coefficients [22,37]. This measure has been selected given its good performance as predictor of recognition rate in continuous speech recognition systems, as can be seen in Table 1.
(4) Log-likelihood ratio distortion (LLR): This measure is calculated similarly to IS distortion [37,32]. Its good performance as a dissimilarity measure in isolated word recognition systems, makes interesting its application in the context of this research.
(5) Weighted spectral slope distortion (WSS): This measure is mainly related to differences in formant locations [24], and was selected because of its relative good performance in all cases presented in Table 1.
(6) Total relative distortion (TRD): It is based on an orthogonal projection of the separated signal on the original signal [18]. As this measure is specific for performance evaluation of BSS algorithms, it was considered appropriate to include it in this work.
(7) Cepstral distortion (CD): This measure is also known as truncated cepstral distance [25]. As ASR systems for continuous speech make use of cepstral-based feature vectors, it is reasonable to include some measures using distances calculated in the cepstral domain.
(8) Mel cepstral distortion (MCD): This measure is calculated in a similar way as CD, but the energy output of a filter bank in mel scale is used instead of spectrum of signals. Also, as many ASR systems use mel cepstral coefficients, it is reasonable to use a distance measure based on them as a predictor of recognition rate.
(9) Measuring normalizing blocks (MNB): This technique applies a simple model of auditory processing but then evaluates the distortion at multiple time and frequency scales with a more sophisticated judgment model [29]. It is a modern approach including perceptual information and, due to his high correlation with MOS scores, was considered as a good candidate for this study.

With the exception of MNB, all the selected measures are frame based, therefore each of them yields a vector (see Appendix A). However, for the evaluation a unique value for each sentence is needed. To achieve this, median value has been employed, as suggested in [37], because in general the measures are affected by outliers corresponding to silence segments at the beginning and the end of each original sentence, in which only noise is observed.

4 Experimental setup

In order to evaluate the performance of selected measures as predictors of recognition rate, an experimental setup was designed. This consists of the reproduction in a room of pre-recorded clean speech sentences and noise, to obtain the mixtures to be used in the evaluation. Reproduction was made through loudspeakers with frequency range from 20 Hz to 20 kHz. In all experiments, two sources where used and the resulting sound field was picked-up at some selected points by two Ono Sokki MI 1233 omnidirectional measurement microphones, with flat frequency response from 20 Hz to 20 kHz and with preamplifiers Ono Sokki MI 3110. In the following sections, brief descriptions of the speech database, spatial location of sources and microphones, separation algorithm and speech recognizer employed in this work will be given.

4.1 Speech Database

In this study a subset of a database generated by the authors is used. It consists of recordings of 20 subjects, 10 male and 10 female, each pronouncing 20 sentences selected for a specific task (remote controlling of a TV set using voice commands). These sentences, in Japanese language, were recorded in an acoustically isolated chamber using a close contact microphone with sampling frequency of 44 kHz, later downsampled to 16 kHz with 16 bit quantization. From this database, one male and one female speakers were selected for this study. In consequence, the original sources consist of 40 utterances, 20 from a male speaker and 20 from a female speaker. The corpus contains an average of 1.4 words/sentence, with average duration of 1.12 s.

Three kinds of interfering signals were selected. One is a signal obtained from recording the noise in a room with a large number of computers working. Spectral and statistical characteristics of this noise source can be seen in Fig. 1. The second kind of noise is a speech signal, pronouncing a sentence different from those used as desired sources. In the case of sentences spoken by female speakers, utterances from male speakers were used as noise and vice versa. The third noise employed is a recording of sound emitted by a TV set. This noise includes speech simultaneously with music. The same TV sound was used to interfere with spoken sentences of both speakers.

4.2 Spatial Setup

All the mixtures were performed in an acoustically isolated chamber as shown in Fig. 2. This setup includes two loudspeakers and two microphones with or

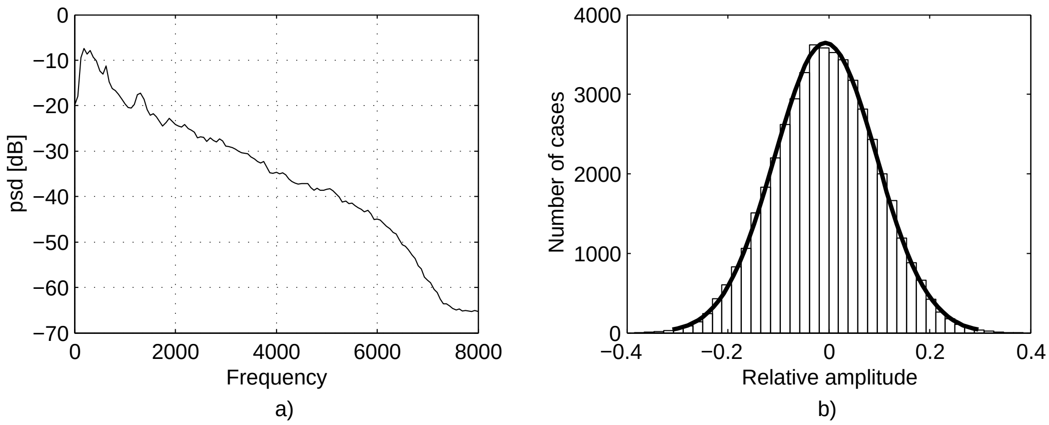

Figure 1. Computer noise characteristics. a) shows power spectral density (psd) estimated by Welch method, and b) shows an estimation of probability density function (pdf)(the line is a normal fit for histogram)

without two reflection boards (used to modify reverberation time). As can be seen in the figure, there are three locations for microphones, a, b and c. In addition, the speech source and noise can be reproduced by loudspeakers in the way shown in Fig. 2, named “position 1”, or they can be exchanged to “position 2” (that is, playing the source in the speakerphone labeled “noise” and vice versa). Powers of reproduced signals were adjusted in such a way to get a power ratio of speech and noise at loudspeakers output of 0 dB or 6 dB. Each of these spatial-power (SP) combinations will be referred as an “SP-Case” in the following. Table 2 shows the codes assigned to each of the SP-Cases, with explanation of the parameters used in each case.

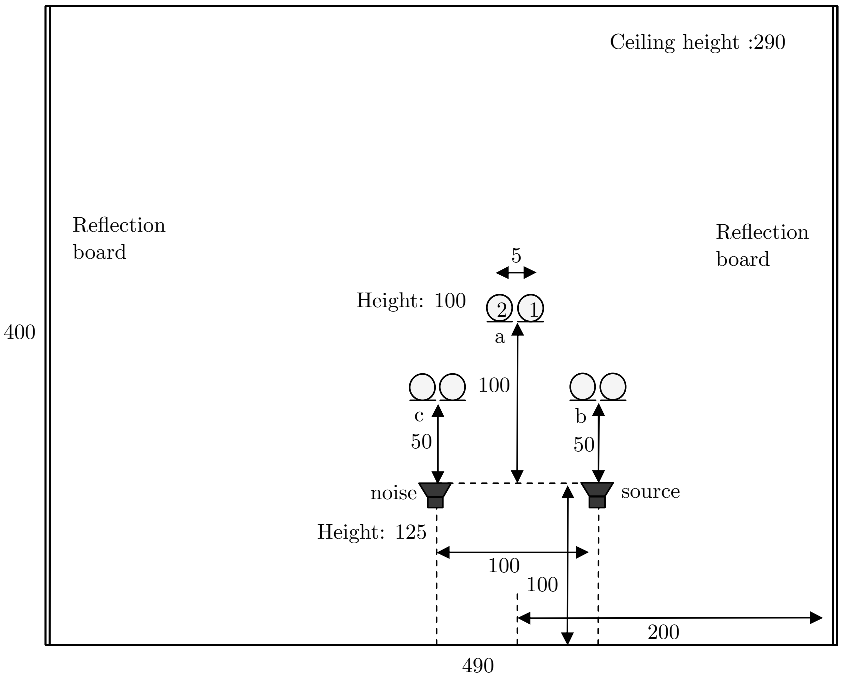

Figure 2. Room used for all recordings. All dimensions are in cm.

Table 2
Signal-Power experimental case codes and their meanings

SP-Case	Microphones	Source-Noise	Power ratio
a10	a	position 1	0 db
a16	a	position 1	6 db
a20	a	position 2	0 db
a26	a	position 2	6 db
b10	b	position 1	0 db
b16	b	position 1	6 db
b20	b	position 2	0 db
c10	c	position 1	0 db
c20	c	position 2	0 db
c26	c	position 2	6 db

In order to analyze the changes in reverberation properties of the room, with the same positions explained before, there was one or two reflection boards added. Without reflection boards, measured reverberation time[6] was $\tau_{60} =$ 130 ms, whereas with one reflection board this time increased to $\tau_{60} = 150$ ms, and with two reflection boards the time was $\tau_{60} = 330$ ms. The same 10 SP-Cases previously mentioned were repeated in each of the reverberation conditions, giving three sets of experiments which would be referred from now on as Low ($\tau_{60} = 130$ ms) , Medium ($\tau_{60} = 150$ ms) and High ($\tau_{60} = 330$ ms)[7] .

Briefly, there are three reverberation conditions. For each of them, 10 SP-Cases were performed, with different combinations of microphone and source locations and different power ratios. Each of these cases consists of 20 utterances from a male and 20 from a female speaker, mixed with each of the three kinds of noise employed, adding to a total of 120 utterances for each SP-Case. In total, separation and recognition over 3600 experimental conditions was evaluated.

[6] The reverberation time τ_{60} is the time interval in which the sound pressure level of a decaying sound field drops by 60 dB, that is to one millionth of its initial value [38].

[7] This naming convention is just to distinguish relative duration of reverberation times in this set of experiments, but this does not imply that the case named "High" actually corresponds to very long reverberation time.

4.3 Separation algorithm

The BSS algorithm is based on independent component analysis (ICA) in the frequency domain [1,39]. Given a number M of active sources and a number N of sensors (with $N \geq M$), and assuming that the environment effect can be modeled as the output of an LTI system, the measured signals at each microphone can be modeled as a convolutive mixture model [1]:

$$x_j\left(t\right) = \sum_{i=1}^{M} h_{ji}\left(t\right) * s_i\left(t\right) \tag{1}$$

where x_j is the j-th microphone signal, s_i is the i-th source, h_{ji} is the impulse response of the room from source i to microphone j, and $*$ stands for convolution. This equation can be written in compact form as $\mathbf{x}\left(t\right) = \mathbf{H}\left(t\right) * \mathbf{s}\left(t\right)$.

Taking a short-time Fourier transform (STFT) of the previous equation, the convolution becomes a multiplication, and assuming that the mixture filters are constant over time (that is, impulse responses does not vary in time), this can be written as:

$$\mathbf{x}(\omega, \tau) = \mathbf{H}(\omega)\mathbf{s}(\omega, \tau). \tag{2}$$

Thus, for a fixed frequency bin ω this means that a simpler instantaneous mixture model can be applied. Under the assumption of statistical independence of the sources over the STFT time τ, the separation model for each frequency bin can be solved using one of the methods for independent component analysis (ICA) [39]. In this context, for each frequency bin ω a matrix $\mathbf{W}\left(\omega\right)$ is searched such as $\mathbf{y}(\omega, \tau) = \mathbf{W}(\omega)\mathbf{x}(\omega, \tau)$, where resulting separated bins $\mathbf{y}\left(\omega, \tau\right)$ should be approximately equal to the original $\mathbf{s}\left(\omega, \tau\right)$. This frequency domain algorithm is a standard formulation for convolutive mixtures, that is known to produce good results for short reverberation times [8].

We have used a STFT with a Hamming window of 256 samples. In order to have enough training data to perform ICA on each frequency bin, a window step of 10 samples was used. For each frequency band, a combination of JADE [40] and FastICA [41] algorithms are used to achieve separation. FastICA is sensitive to initial conditions, because it is a Newton-like algorithm. For this reason, JADE was applied to find an initial approximation to separation matrix, and then FastICA was employed to improve the results (with the JADE guess as initial condition). For FastICA we have used the nonlinear function $G(y) = log(a + y)$ with its derivative $g(y) = \frac{1}{a+y}$. Both complex versions of JADE and FastICA were obtained from the websites of their authors.

One problem of this approach is that the ICA algorithms can give arbitrary permutations and scalings for each frequency bins. So in two successive frequency bins, extracted source i can correspond to different original sources, with arbitrary scaling in amplitude. Permutation and amplitude indeterminacies are solved by the algorithm proposed in [19]. Permutation is solved using the amplitude modulation properties for speech: at two near frequency bins, the envelope of the signal in that band should be similar for bins originated by the same source. Using correlations with accumulated envelopes of already separated bins, one can classify new frequency bands. To estimate the envelopes, we used a 20 milliseconds averaging lowpass filter. The amplitude indetermination is solved by applying the obtained mixing matrix to only one of the separated sources. After the separation and the solution of the indeterminacies, the overlap-and-add method of reconstruction was used to obtain the time-domain signals [42].

For each of the reverberation conditions, mixture signals captured by microphone in each combination of sentences and noises were processed with this algorithm. From each pair of separated signals, the signal more likely to represent the desired source was selected by means of a correlation. In this way, a database with separated signals corresponding to each of the sentences in each experimental condition was generated. Before the application of quality measures, correlation was used to compensate any possible delay between separated and original signal, and to detect possible signal inversions (if maximum correlation is negative, the signal is multiplied by -1). Also all signals were normalized to minimize the effect of magnitude indeterminacies. This was done by dividing both separated and original by their respective energy.

4.4 Recognizer

The recognizer used was the large vocabulary continuous speech recognition system Julius [43], based on hidden Markov models (HMM). This is a standard recognition system widely used for Japanese language. The decoder performs a two-pass search, the first with a bi-gram and the second with a tri-gram language model. This system was used with acoustic models for continuous density HMM in HTK [44] format. The models were trained with two databases provided by the Acoustic Society of Japan (ASJ): a set of phonetically balanced sentences (ASJ-PB) and newspaper article texts (ASJ-JNAS). Around 20000 sentences uttered by 132 speaker of each gender were used.

The recognizer use 12 mel frequency cepstral coefficients (MFCC) computed each 10 milliseconds, with temporal differences of coefficients ($\Delta MFCC$) and energy (ΔE) for a total of 25 feature coefficients. Also cepstral mean normalization was applied to each utterance. Phonetic tied-mixture triphones are

Table 3
Word recognition rates (WRR%) with only one source in the real room. In this case, there is only reverberation effect. This can be considered an upper limit of the obtainable recognition rate in each case.

Mic.	Low Rev.	Med. Rev.	High Rev.
1cm	83	80	79
a	80	75	53
b	75	66	66
c	56	49	44

used as acoustic models. The full acoustic model consists of 3000 states tying 64 gaussian from a base of 129 phonemes with different weights depending of the context. For the language model, both bi-gram and tri-gram models were generated from 118 million words from 75 months newspaper articles, which were also used to generate the lexicon [45].

Word recognition rate was evaluated as:

$$WRR\% = \frac{T - D - S}{T} 100\% \tag{3}$$

where T is the number of words in the reference transcription, D is the number of deletion errors (words present in the reference transcription that are not present in the system transcription) and S is the number of substitution errors (words that were substituted by others in the system transcription) [44].

To compare with obtained results, WRR by this system was evaluated on the source sentences reproduced in the room but without any interfering noise, with microphones in location a, b and c, and also with microphones located at 1 cm from source. This permits to evaluate the degradation effect caused on the recognizer by reverberation, even without interfering noise. These results are shown in Table 3. As the algorithm generally does not reduce –in great amount– the reverberation effect, these values can be taken as a baseline limit for obtainable recognition rate in each case [8].

Word recognition rates for mixtures and for BSS separated signals was also evaluated, as shown in Fig. 3. For this figure, the SP-Cases where grouped according to location of microphones relative to sources, as equal distance (a10+a20 and a16+a26), nearer to desired source (b10+c20 and b16+c26), and nearer to noise source (b20+c10).

[8] It must be noted that feeding the ASR system with the original clean sentences, yielded a word recognition rate of 100%.

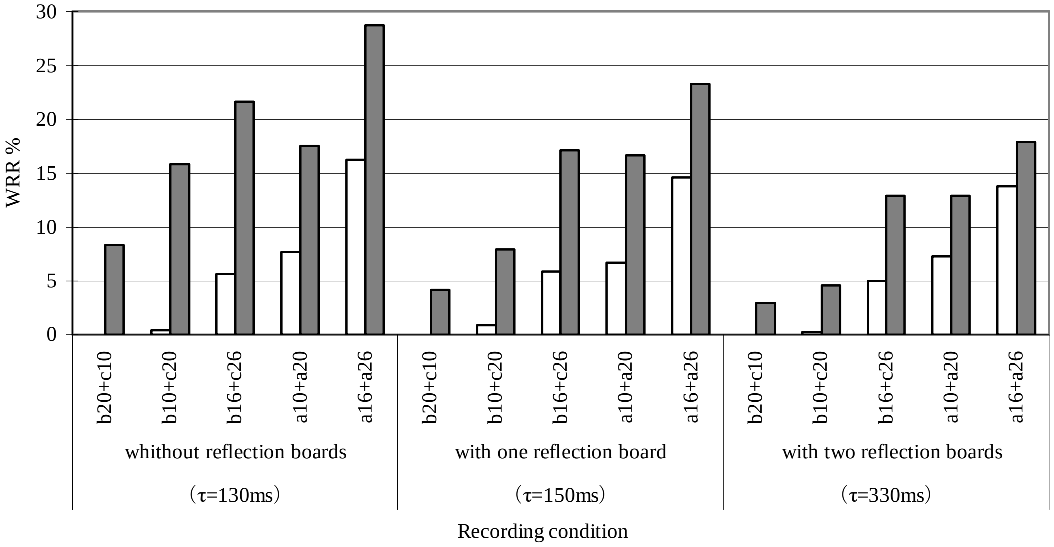

Figure 3. Word recognition rates (WRR%) using mixtures and separated sources for different reverberation conditions. White bars: mixed (recorded) sound; Gray bars: sound separated by BSS. Label b20+c10 mean average of results of these SP-Cases (with similar meaning for the other labels).

5 Results

For each of the reverberation conditions and for each SP-Case, quality measures have been calculated for all sentences and all noise types. Then, an average of each measure for all utterances has been taken, grouping them for noise kind. In this way, for each combination of reverberation condition/SP-Case, three values of quality were generated corresponding to average quality in each noise kind. In the same way, for each combination of reverberation condition and SP-Case, recognition rate was also evaluated, separated by the kind of noise, obtaining three values of recognition rate for each case.

With these data three analyses were made. Table 4 presents Pearson correlation coefficient (absolute value) for the analyses, defined as [46]:

$$\rho_{xy} = \frac{\sum_i \left[(x_i - \bar{x})(y_i - \bar{y}) \right]}{\left[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2 \right]^{1/2}} \tag{4}$$

where x_i represents the quality measure to be used as predictor, y_i the WRR%, and $\bar{x}$, $\bar{y}$ the corresponding estimated mean values. First, for each reverberation condition, correlation of quality measures as predictors of recognition rate was evaluated, discriminated for each kind of noise, in such a way that the sample consist of 10 pairs of quality measures/recognition rates (each pair is a SP-Case). Second, the same analysis was performed considering all kind of noises, that is taking in the sample the 30 pairs of quality measures/recognition rates, considering all kinds of noise for a reverberation condition, giving as a result

Table 4
Correlation coefficient $|\rho|$ for all experiments. Best value for each case has been marked in boldface. "All" includes in the sample all noise kinds for a given reverberation condition, and "ALL" includes all noise kinds and all reverberation conditions. Last row shows the standard deviation of the regression residual.

Reverb.	Noise	segSNR	IS	LAR	LLR	WSS	TRD	CD	MCD	MNB
Low	Comp.	0.80	0.44	0.70	0.70	**0.92**	0.88	0.81	0.87	0.90
	TV	0.68	0.13	0.76	0.72	**0.84**	0.77	0.78	0.80	0.78
	Speech	0.64	0.77	0.74	0.65	**0.84**	0.62	0.79	0.84	0.56
	All	0.61	0.39	0.73	0.71	**0.86**	0.75	0.77	0.80	0.62
Medium	Comp.	0.78	0.43	0.58	0.62	**0.88**	0.82	0.74	0.85	0.85
	TV	0.76	0.31	**0.92**	0.91	0.90	0.86	0.91	0.85	0.85
	Speech	0.78	0.76	0.82	**0.85**	0.66	**0.85**	0.76	0.62	0.64
	All	0.76	0.46	0.75	0.74	0.77	**0.83**	0.74	0.72	0.78
High	Comp.	0.77	0.53	0.74	0.75	0.83	**0.85**	0.81	0.80	0.83
	TV	0.81	0.71	0.92	0.92	**0.93**	0.90	**0.93**	0.90	0.87
	Speech	0.74	0.33	0.75	0.74	0.77	0.72	**0.79**	0.75	0.66
	All	0.75	0.50	0.78	0.79	0.81	**0.84**	**0.84**	0.79	0.75
$\lvert\rho\rvert$	ALL	0.74	0.43	0.73	0.71	0.83	0.84	0.76	0.77	0.75
σ_r	ALL	5.94	9.70	6.68	7.36	4.98	5.74	6.42	5.77	7.86

one value of general correlation for each reverberation condition. Third, the correlation was evaluated without data segregation, that is, including in the sample all the kinds of noise and all the reverberation conditions.

Also, dispersion graphic for all data was made in Fig. 4 for the case of all noise kinds and all reverberation conditions (last rows in Table 4). A dispersion graphic was drawn for each measure including a total least squares regression line and two lines that mark regression value plus/minus two times the standard deviation of residual. This standard deviation was estimated according to $\sigma_r^2 = \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2 / (N-2)$ where y_i is the true WRR% and $\widehat{y}_i$ the predicted value by regression [46]. Also, this figure shows the values of $|\rho|$ and σ_r.

6 Discussion

Many interesting findings can be extracted from the analysis of Table 4. In general, it can be said that the measure showing the maximum correlation as predictor of recognition rate is WSS. This is because it is the best in 6 of 12 cases. In the cases where it has not been the one with largest correlation, it can be seen that it is close to the maximum value. Regarding the global

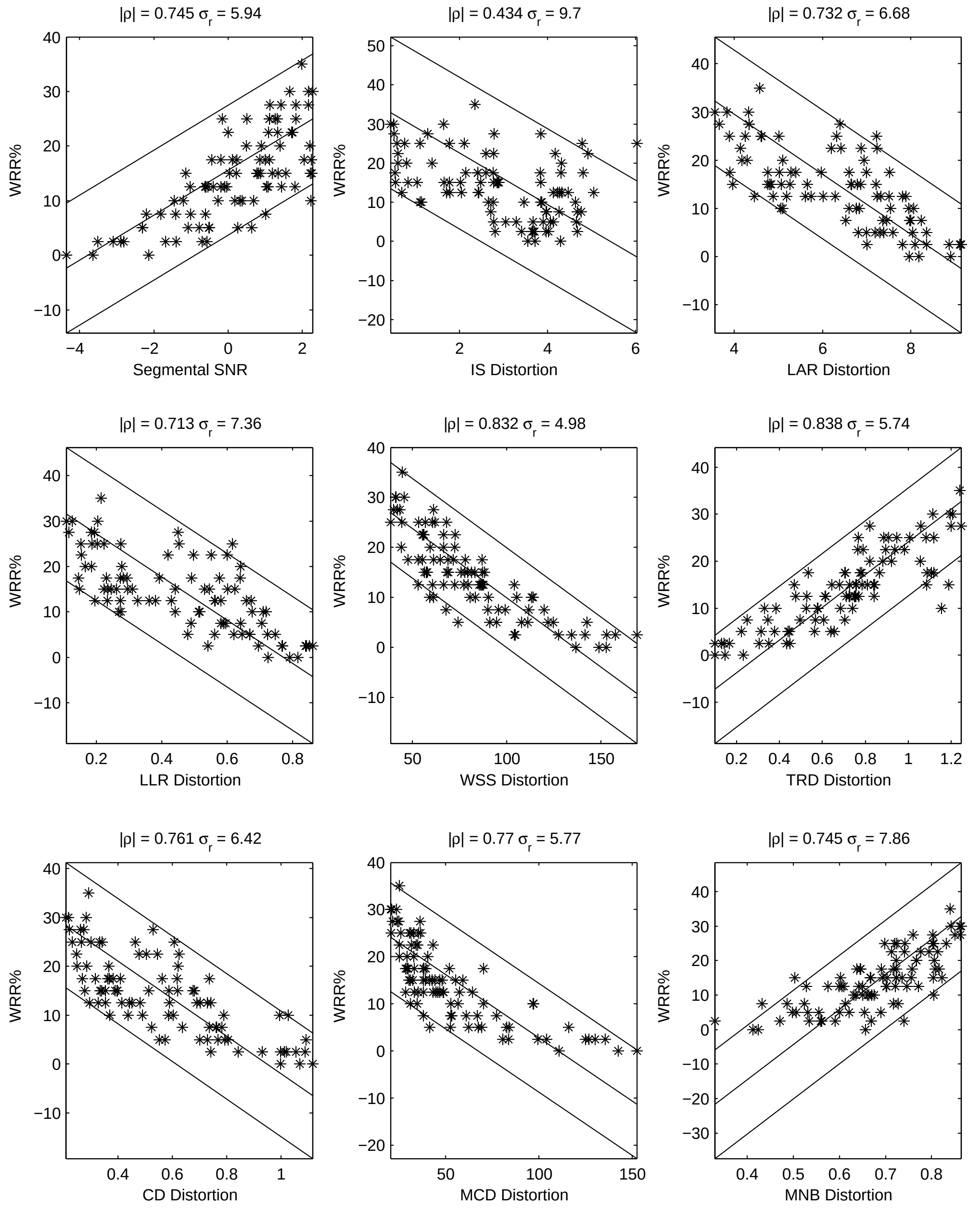

Figure 4. Regression analysis of quality measures for all experimental cases. WRR%, word recognition rate.

value of correlation ($|\rho|$ ALL), WSS is relegated to the second position, but the difference (0.0063) is not significative[9] . Furthermore, in the global case it is the measure that has lower residual variance.

For the lowest reverberation time, WSS is clearly superior to the other measures. In the intermediate reverberation case, the best performance is for TRD

[9] This difference is not seen in Table 4 due to the two decimal precision used.

but sharing the success for different noises with LAR, LLR and WSS. In the High reverberation case, CD measure seem to behave better, but closely followed by TRD.

One possible explanation for the lowering of correlation of WSS measure is the following: as reverberation time increases, performance of separation algorithm decreases, and so resulting separated signals will have increasing amount of interfering signal. In this case, as the original signal is present at a high level, measures that take into account preservation of special features (like formants in WSS) would give good values, although the interfering level would be high enough for the recognizer to fail. Conversely, those measures that have more relation to whole spectral distances between signals would behave closer to recognition rates.

Regarding the effect of different kinds of noise, in the case of computer noise, the best measure is WSS showing the highest correlation for low and medium reverberation times, while TRD is better for long reverberation times. The algorithm used for separation can perform very well in this case of quasi-stationary noise. Therefore, separated signal will have very similar spectral contents to the original one, being mainly distorted due to reverberation effects. For TV noise the results show that, at low reverberation, WSS is a better measure, at medium reverberation the best measure is LAR (although all LLR, WSS and CD are very near), and for high reverberation WSS and CD have a better performance. This can be also explained by a good performance for the separation algorithm which can also manage this non-stationary noise. In the case of speech noise, at low reverberations the best measure is WSS, for medium reverberation TRD and LLR are the best, and at high reverberation, CD is the best one. Speech noise is the hardest condition for the separation algorithm, and this lowering of performance of WSS can be explained in a similar way as before (related to degradation of WSS for long reverberation times).

TRD is the second measure in global performance, particularly well correlated at medium reverberation times. It also has the second lower global residual variance. This can be related to the fact that this measure was designed specifically to evaluate blind source separation algorithms. This could show an important relation (that was not necessarily obvious a priori) between the evaluation of the separation algorithm itself and its performance for speech recognition.

The relative good results of cepstral measures is not a surprise. Their quite uniform performance for all levels of reverberation can be related to the internal representation of the recognizer based in cepstral coefficients. So, changes in these coefficients are reflected directly in recognition rate, giving some uniform behavior. Although one could expect a general better performance for MCD than for CD, the results not always agree. It would be interesting to perform the same experiments with a recognizer based in different feature vectors, like Rasta-PLP [47,48,49], to check whether the good performance is recognizer-related or in fact can be generalized. Comparing only the cepstral measures, MCD is very stable with reverberation, keeping high correlation levels in all cases. CD presents better correlation than MCD at higher reverberation times.

In opposition to the expected results, MNB performance was rather poor, compared to the previous ones. This measure was specially designed for speech vocoders, and maybe that distortions are very different to the ones presented in blind source separation.

It is also interesting to verify that segmental SNR has been outperformed in all cases by almost all measures. This must be considered in all works where improvement of algorithms is reported using this measure. According to these results, improvements reported in terms of SNR will not be reflected in recognition rates.

The results presented here were obtained for this separation algorithm and this specific recognizer, so strictly speaking they are only applicable for these cases. Nevertheless, we consider that as long as the separation algorithm uses similar processing (i.e. frequency domain BSS) and the speech recognizer uses the same paradigm (HMM with MFCC features) the results should not change qualitatively. On the other hand, all the experiments here were made with Japanese language. However, there are some studies, like [29] where it is shown that the results of objective quality measures for different languages are quite similar, and so we expect an objective measure not to change significantly when applied to different languages (specially in the case of WSS, where both good recognition rate and good perceptual quality are achieved).

7 Conclusions

From the analysis of the obtained results, the measure presenting more correlation with word recognition rates is WSS. When reverberation time increases, it has been proved that the performance of this measure degrades gradually, meanwhile TRD and cepstral-based measures perform better than WSS. This is an important guide at the time of choosing a suitable separation quality measure for speech recognition.

On the other hand, remembering that WSS is a highly correlated measure with subjective evaluation of quality (Table 1), one additional advantage of using this measure becomes evident. If the algorithm under evaluation is designed not only for the front-end of ASR, but also as an enhancement part of the system that would present their result to human listeners, it can be expected that using WSS as a quality measure will allow to achieve both objectives: a good recognition rate together with good perceptual quality of speech.

One of the possible practical applications of these results in the field of BSS for ASR is in algorithm selection/tuning. In early research stages, where a particular separation algorithm and its parameters should be selected, direct evaluation by means of recognition rate would be prohibitive as a result of the large amount of test over complete databases. The alternative is to use one of the objective quality measures to select some candidate algorithms and their parameters, and then perform a fine tuning with the complete ASR system.

A Quality measures details

The following notation will be used: let the original signal be $\mathbf{s}$ and separated signal $\widehat{\mathbf{s}}$, both of M samples. Frame m of length N of original signal is defined as $\mathbf{s}_m = [s[mQ], \ldots, s[mQ+N-1]]$, where Q is the step size of the window in a short-time analysis, and with analogous definition for corresponding frame of the separated signal. In the case of measures derived from linear prediction, a system order P is assumed. Using this notation, the evaluated measures are:

(1) SegSNR: Given a frame of original signal and corresponding frame of separated signal, segSNR is defined as [22]:

$$d_{SNR}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = 10 \log_{10} \frac{\|\mathbf{s}_m\|^2}{\|\widehat{\mathbf{s}}_m - \mathbf{s}_m\|^2} \ , \tag{A.1}$$

where $\|\cdot\|$ is the 2-norm defined as usual, $\|\mathbf{x}\| = \left(\sum_{n=1}^{N} x[n]^2\right)^{1/2}$.

(2) IS distortion: Given LP coefficients vector of original (clean) signal, $\mathbf{a_m}$, and LP coefficient vector for the corresponding frame of separated signal, $\widehat{\mathbf{a}}_\mathbf{m}$, IS distortion is defined as [31,37]:

$$d_{IS}(\mathbf{a}_m, \widehat{\mathbf{a}}_m) = \frac{\sigma_m^2}{\widehat{\sigma}_m^2} \frac{\widehat{\mathbf{a}}_m^T \mathbf{R} \widehat{\mathbf{a}}_m}{\mathbf{a}_m^T \mathbf{R} \mathbf{a}_m} + \log\left(\frac{\widehat{\sigma}_m^2}{\sigma_m^2}\right) - 1, \tag{A.2}$$

where $\mathbf{R}$ is the autocorrelation matrix, and σ^2, $\widehat{\sigma}^2$ are the all-pole system gains.

(3) LAR distortion: Given reflection coefficient vector for an LP model of a signal, $\mathbf{k}_m = [\kappa(1;m), \ldots, \kappa(P;m)]^T$, the Area Ratio vector is defined as $\mathbf{g}_m = [g(1;m), \ldots, g(P;m)]^T$, where $g(l;m) = \frac{1+\kappa(l;m)}{1-\kappa(l;m)}$. These coefficients are related to the transversal areas of a variable section tubular model for the vocal tract. Using these coefficients, for a frame of original signal, and corresponding frame of separated signal, LAR distortion is defined

as [22,37]:

$$d_{LAR}(\mathbf{g}_m, \widehat{\mathbf{g}}_m) = \left\{ \frac{1}{P} \left\| \log \mathbf{g}_m - \log \widehat{\mathbf{g}}_m \right\|^2 \right\}^{\frac{1}{2}}. \qquad \text{(A.3)}$$

(4) LLR distortion: Given LP coefficient vector of a frame of original and separated signal, $\mathbf{a_m}$ and $\widehat{\mathbf{a}}_\mathbf{m}$ respectively, LLR distortion is given by [32,37]:

$$d_{LLR}(\mathbf{a}_m, \widehat{\mathbf{a}}_m) = \log \frac{\widehat{\mathbf{a}}_m^T \mathbf{R} \widehat{\mathbf{a}}_m}{\mathbf{a}_m^T \mathbf{R} \mathbf{a}_m}, \qquad \text{(A.4)}$$

where $\mathbf{R}$ is the autocorrelation matrix.

(5) WSS distortion: Given a frame of signal, the spectral slope is defined as $SL[l;m] = S[l+1;m] - S[l;m]$, where $S[l;m]$ is a spectral representation (in dB), obtained from a filter bank using B critical bands in Bark scale (with index l referring to position of filter in filter bank). Using this, WSS between original signal and separated one is defined as [24,37]:

$$\begin{aligned} d_{WSS}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = {} & K_{spl}(K - \widehat{K}) + \\ & \sum_{l=1}^{B} \bar{w}[l] \left(SL[l;m] - \widehat{SL}[l;m] \right)^2, \end{aligned} \qquad \text{(A.5)}$$

where K_{slp} is a constant weighting global sound pressure level, K and $\widehat{K}$ are sound pressure level in dB, and weights $w[l]$ are related to the proximity of band l to a local maximum (formant) and global maximum of spectrum, as $\bar{w}[l] = (w[l] + \widehat{w}[l])/2$, with:

$$w[l] = \left(\frac{C_{loc}}{C_{loc} + \Delta_{loc}[l]} \right) \left(\frac{C_{glob}}{C_{glob} + \Delta_{glob}[l]} \right) \qquad \text{(A.6)}$$

with a similar definition for $\widehat{w}[l]$, where C_{glob} and C_{loc} are constants and Δ_{glob}, Δ_{loc} are the log spectral differences between the energy in band l and the global or nearest local maximum, respectively. This weighting will have larger value at spectral peaks, especially at the global maximum, and so it will give more importance to distances in spectral slopes near formant peaks (for more details, see [24,31,34]).

(6) TRD: The separated source can be decomposed as $\widehat{\mathbf{s}} = \mathbf{s}^D + \mathbf{e}^I + \mathbf{e}^N + \mathbf{e}^A$, where $\mathbf{s}^D = \langle \widehat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s} / \|\mathbf{s}\|^2$ is the part of $\widehat{\mathbf{s}}$ perceived as coming from the desired source, and $\mathbf{e}^I$, $\mathbf{e}^N$ and $\mathbf{e}^A$ the error parts coming from the other sources, sensors noises and artifacts of the algorithm. For each frame m of these components, TRD is defined as [17,18]:

$$d_{TRD}(\mathbf{s}, \widehat{\mathbf{s}}; m) = \frac{\left\| \mathbf{e}_m^I + \mathbf{e}_m^N + \mathbf{e}_m^A \right\|^2}{\left\| \mathbf{s}_m^D \right\|^2}. \qquad \text{(A.7)}$$

(7) CD: Given the vectors of cepstral coefficients $\mathbf{c}_m$ and $\widehat{\mathbf{c}}_m$, corresponding to a frame of original signal and corresponding separation result, CD for

the first L coefficients is defined as [31]:

$$d_{CD}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = \sum_{l=1}^{L} \left(c_m[l] - \widehat{c}_m[l]\right)^2 . \tag{A.8}$$

(8) MCD: Given mel cepstral coefficients $\mathbf{c}_m^{mel}$ and $\widehat{\mathbf{c}}_m^{mel}$ corresponding to original and resulting separated signal respectively, calculated using a filter bank of B filters in mel scale, MCD for the first L coefficients is defined as [22,31]:

$$d_{MCD}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = \sum_{l=1}^{L} \left(c_m^{mel}[l] - \widehat{c}_m^{mel}[l]\right)^2 . \tag{A.9}$$

(9) MNB: This measure is more complex than the previous ones, so it will be only outlined here. It includes first a time-frequency representation, that is transformed to Bark scale to obtain a representation more closed to the auditory mapping. After this transformation the auditory time-frequency representations of the reference $S(t,f)$ and test $\widehat{S}(t,f)$ are analyzed by a hierarchical decomposition of measuring normalizing blocks in time (tMNB) and frequency (fMNB). Each MNB produces a series of measures and a normalized output $\widehat{S}'(f,t)$. For a tMNB, the normalization is done by:

$$\begin{aligned} e(t,f_0) &= \frac{1}{\Delta f}\int_{f_0}^{f_0+\Delta f} \widehat{S}(t,f)df - \frac{1}{\Delta f}\int_{f_0}^{f_0+\Delta f} S(t,f)df \\ \widehat{S}'(f,t) &= \widehat{S}(t,f) - e(t,f_0) \end{aligned} \tag{A.10}$$

where f_0 and Δf define a frequency band for the integration. By integration of $e(t,f_0)$ over time intervals, a group of measures for this tMNB is obtained. The same is used for each fMNB, with the roles of t and f interchanged. So, the hierarchical decomposition proceeds from larger to smaller scales, for frequency and time, calculating distances and removing the information of each scale. After this process, a vector of measures $\boldsymbol{\mu}$ is obtained. Then a global auditory distance (AD) is built by using appropriate weights $AD = \sum_{i=1}^{J} w_i \mu_i$. Finally, a logistic map is applied to compress the measure and adjust it to a finite interval, given by $L(AD) = \frac{1}{1+e^{aAD+b}}$. The authors have proposed two different hierarchical decompositions, called structure 1 and 2, that use different tMNB and fMNB decompositions. For more details, refer to [29].

For the analysis, the following parameters were used:

- Frame length $N = 512$ samples (32 ms of signal).
- Step size for analysis window $Q = 128$ samples (8 ms of signal).
- Order for LP models $P = 10$.
- WSS: $B = 36$, $K_{spl} = 0$, $C_{loc} = 1$ and $C_{glob} = 20$ as recommended by author in [24].

- CD: truncation at $L = 50$ coefficients.
- MCD: number of filters $B = 36$, number of coefficients $L = 18$.
- MNB (structure 1): $a = 1.0000$ and $b = -4.6877$ as suggested in [29]. The signals were subsampled to 8000 Hz before applying this measure.

References

[1] A. Cichocki, S. Amari, Adaptive Blind Signal and Image Processing. Learning Algorithms and applications., John Wiley & Sons, 2002.

[2] M. Kahrs, K. Brandenburg (Eds.), Applications of Digital Signal Processing to Audio and Acoustics, The Kluwer International Series In Engineering And Computer Science, Kluwer Academic Publishers, 2002.

[3] R. Gilkey, T. Anderson (Eds.), Binaural and Spatial Hearing in Real and Virtual Environments, Lawrence Erlbaum Associates, 1997.

[4] J. Blauert, Spatial Hearing - The Psychophysics of human sound localization, MIT Press, 1997.

[5] T. Finitzo-Hieber, T. Tillmann, Room acoustics effects on monosyllabic word discrimination ability by normal and hearing impaired children, Journal of Speech and Hearing Research 21 (1978) 440–458.

[6] C. Crandell, J. Smaldino, Classroom acoustics for children with normal hearing and with hearing impairment, Language, Speech, and Hearing Services in Schools 31 (4) (2000) 362–370.

[7] B. Kinsbury, N. Morgan, Recognizing reverberant speech with RASTA-PLP, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, pp. 1259–1262.

[8] J. Benesty, S. Makino, J. Chen (Eds.), Speech Enhancement, Signals and Communication Technology, Springer, 2005.

[9] N. Mitianoudis, M. Davies, New fixed-point ica algorithms for convolved mixtures, in: Proceedins of the Third International Conference on Independent Component Analysis and Source Separation, 2001, pp. 633–638.

[10] S. Ikeda, N. Murata, An approach to blind source separation of speech signals, in: Proceedings of the 8th International Conference on Artificial Neural Networks, Vol. 2, 1998, pp. 761–766.

[11] H. Gotanda, K. Nobu, T. Koya, K. Kaneda, T. Ishibashi, N. Haratani, Permutation correction and speech extraction based on split spectrum through fastica, in: Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, 2003, pp. 379–384.

[12] T. Lee, A. Bell, R. Orglmeister, Blind source separation of real world signals, in: Proceedings of IEEE International Conference Neural Networks, 1997, pp. 2129–2135.

[13] L. Parra, C. Spence, Convolutive blind separation of non-stationary sources, IEEE Transactions on Speech and Audio Processing 8 (3) (2000) 320–327.

[14] L. Parra, C. Alvino, Geometric source separation: merging convolutive source separation with geometric beamforming, IEEE Transactions on Speech and Audio Processing 10 (6) (2002) 352–362.

[15] S. C. Douglas, X. Sun, Convolutive blind separation of speech mixtures using the natural gradient, Speech Communication 39 (1-2) (2003) 65–78.

[16] D. Schobben, K. Torkkola, P. Smaragdis, Evaluation of blind signal separation methods, in: Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation, 1999, pp. 261–266.

[17] R. Gribonval, L. Benaroya, E. Vincent, C. Févotte, Proposals for performance measurement in source separation, in: Proceedings of the 4th Symposium on Independent Component Analysis and Blind Source Separation, 2003, pp. 763–768.

[18] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE Trans. on Audio, Speech and Languiage Processing 14 (4) (2006) 1462– 1469.

[19] N. Murata, S. Ikeda, A. Ziehe, An approach to blind source separation based on temporal structure of speech signals, Neurocomputing 41 (2001) 1–24.

[20] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, N. Kitawaki, Blind source separation in reflective sound fields, in: International Workshop on Hands-Free Speech Communication, 2001, pp. 51–54.

[21] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, N. Kitawaki, Combined approach of array processing and independent component analysis for blind separation of acoustic signals, IEEE Transactions on Speech and Audio Processing 11 (3) (2003) 204–215.

[22] J. Deller, J. Proakis, J. Hansen, Discrete Time Processing of Speech Signals, Macmillan Publishing, New York, 1993.

[23] W. D. Voiers, Diagnostic acceptability measure for speech communication systems, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1977, pp. 204–207.

[24] D. H. Klatt, Prediction of perceived phonetic distance from critical-band spectra: a first step, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (1982) 1278–1281.

[25] A. Gray, J. Markel, Distance measures for speech processing, IEEE Transactions on Acoustics, Speech, and Signal Processing 24 (5) (1976) 380–391.

[26] R. M. Gray, A. Buzo, A. H. Gray, Y. Matswama, Distortion measures for speech processing, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (4) (1980) 367–376.

[27] T. P. Barnwell, Correlation analysis of subjective and objective measures for speech quality, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1980, pp. 706–709.

[28] J. G. Beerends, J. A. Stemerdink, A perceptual speech quality measure based on a psychoacoustic sound representation, Journal of the Audio Engineering Society 42 (3) (1994) 115–123.

[29] S. Voran, Objective estimation of perceived speech quality. i. development of the measuring normalizing block technique, IEEE Trans. on Speech and Audio Processing 7 (4) (1999) 371–382.

[30] S. Voran, Objective estimation of perceived speech quality .ii. evaluation of the measuring normalizing block technique, IEEE Trans. on Speech and Audio Processing 7 (4) (1999) 383–390.

[31] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[32] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 23 (1) (1975) 67–72.

[33] D. Mansour, B. H. Juang, A family of distortion measures based upon projection operation for robust speech recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 37 (2) (1989) 1659–1671.

[34] N. Nocerino, F. K. Soong, L. R. Rabiner, D. H. Klatt, Comparative study of several distortion measures for speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, 1985, pp. 25–28.

[35] J. H. Hansen, L. M. Arslan, Robust feature estimation and objective quality assessment for noisy speech recognition using the credit card corpus, IEEE Transactions on Speech and Audio Processing 3 (3) (1995) 169–184.

[36] M. Basseville, Distance measures for signal processing and pattern recognition, Signal Processing 18 (1989) 349–369.

[37] J. H. Hansen, B. Pellom, An effective quality evaluation protocol for speech enhancement algorithms, in: Proceedings of the Interternational Conference on Spoken Language Processing, Vol. 7, 1998, pp. 2819–2822.

[38] H. Kuttruff, Room Acoustics, 4th Edition, Taylor & Francis, 2000.

[39] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, Inc., 2001.

[40] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non Gaussian signals, IEE Proceedings-F 140 (1993) 362–370.

[41] E. Bingham, A. Hyvärinen, A fast fixed-point algorithm for independent component analysis of complex valued signals, International journal of Neural Systems 10 (1) (2000) 1–8.

[42] J. B. Allen, L. R. Rabiner, A unified approach to short-time Fourier analysis and synthesis, Proceedings of the IEEE 65 (11) (1977) 1558–1564.

[43] A. Lee, T. Kawahara, K. Shikano, Julius — an open source real-time large vocabulary recognition engine, in: Proceedings of the European Conference on Speech Communication and Technology, 2001, pp. 1691–1694.
URL `http://julius.sourceforge.jp/en/julius.html`

[44] S. Yung, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK book (for HTK Version 3.3), Cambridge University Engineering Department, Cambridge (2005).

[45] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, K. Shikano, Free software toolkit for Japanese large vocabulary continuous speech recognition, in: Proceedings of the International Conference on Spoken Language Processing, Vol. 4, 2000, pp. 476–479.

[46] D. C. Montgomery, G. C. Runger, Applied Statistics and Probability for Engineers, 3rd Edition, Third Edition, 2003.

[47] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, Journal Acoust. Soc. Am. 87 (4) (1990) 1738–1752.

[48] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. on Speech and Audio Processing 2 (4) (1994) 578–589.

[49] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, , G. Tong, Integrating RASTA-PLP into speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, 1994, pp. 421–424.