Perceptual evaluation of blind source separation for robust speech recognition

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

[a] *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.*

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

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

Abstract

In a previous article, an evaluation of several objective quality measures as predictors of recognition rate after application of a blind source separation algorithm was reported. In this work, the experiments were repeated using some new measures, based on the perceptual evaluation of speech quality (PESQ), which is part of the ITU P862 standard for evaluation of communication systems. The raw PESQ and a nonlinearly transformed PESQ were evaluated, together with several composite measures. The results show that the PESQ-based measures outperformed all the measures reported in the previous work. Based on these results, we recommend the use of PESQ-based measures to evaluate blind source separation algorithms for automatic speech recognition.

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

* 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), d.milone@ieee.org (Diego Milone), lrufiner@bioingenieria.edu.ar (Hugo Leonardo Rufiner), myanagid@mail.doshisha.ac.jp (Masuzo Yanagida).

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

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

1 Introduction

Quality evaluation for speech processing is a very important step in the development of advanced algorithms. This is particularly important for the field of blind source separation (BSS) of speech, which has emerged in the last years and has been increasing in importance, with a large presence in the most important conferences and journals in the area. The increase in the number of papers published shows, however, a bias towards the presentation of new separation algorithms, while only a few examples can be found discussing effective ways to evaluate the quality of the algorithms [1]. As an example, one can cite the work by Vincent et al, where several measures are specifically designed to take into account the different factors that can affect the result of BSS algorithms [2]. As clearly stated in [3], the evaluation is a very complex task, and must be matched to the application for which the algorithm is produced.

For convolutive BSS, usually the quality of algorithms is reported using Signal to Interference Ratio, a measure that require knowledge of the mixing conditions to be estimated. In a recent review, more than 400 papers in the subject were compared, and some results from them using SIR are reported [4]. The authors of this review found that it is very difficult to determine if the reported results of the BSS algorithms would be applicable to real cases, due to the variety of evaluation conditions used. They report also that only about 10% of the references reported results on real recordings [4].

In our recent paper [5], we addressed the subject of quality evaluation of BSS algorithms for the specific task of automatic speech recognition (ASR). As it is clear, in this case the ultimate measure of quality would be the output of a speech recognition system. Nevertheless, there are two factors that make the evaluation using a recognizer undesirable. First of all, the method will have very little sensitivity with respect to the parameters of the algorithm. This means that perhaps a considerable improvement in quality would be needed before obtaining a difference in the recognizer output, and so two BSS algorithms would get the same recognition rate even if there are differences in quality among them. The second aspect is that obtaining statistically significant results from the output of a speech recognizer would imply the use of a large number of signals, making this impractical at least in the first stages of the research. These problems are very similar to the ones that are present in the case of subjective evaluation tests.

To overcome these problems, in [5] we proposed the use of objective quality measures that shown a good correlation with the recognition rates of the ASR system. From that study we found three measures that correlated well: weighted spectral slope (WSS), total relative distortion (TRD), and cepstral distortion (CD). Although there was not a clear winner for all cases, WSS

showed a very good performance in most cases. This measure was derived from experiments with speech perception [6] and has been reported to be very well correlated with subjective tests of quality. This makes it very interesting, because if some BSS algorithm gets a high score using this measure, it will probably produce good results in speech recognition tasks, and also will have good perceptual subjective quality.

After publication of that paper, we continued our research with other quality measures, particularly perceptually derived ones, and focused on the perceptual evaluation of speech quality (PESQ) measure. This measure is defined in the standard ITU P.862 as the mean for evaluating the quality of speech transmitted over communication channels [7]. It has been widely studied and shown a very high correlation with subjective quality for a wide variety of transmission channels, distortions, codification algorithms, and languages. Also, a recent study [8] has shown its good properties for evaluating speech enhancement algorithms in terms of subjective perceptual quality. This makes it a good candidate for our study, so we extended our previous work to this measure.

In [9], the PESQ measure was evaluated to predict recognition rate of an ASR system, showing good results. In that work, however, the correlation is evaluated with respect to the noisy speech (with additive and convolutive noise), without using any separation algorithm. On the contrary, in [10] the performance of three measures as predictors of recognition rate of an ASR system for speech with additive noise, after the application of single channel speech enhancement algorithms, is evaluated. PESQ is shown to have the higher correlation with recognition rate. Finally, in [11], the perceptual evaluation of audio quality (PEAQ) measure, which is a perceptual measure similar to PESQ but designed for general wideband audio sources and not only for speech, is evaluated as predictor of subjective quality. The work used music sources contaminated by convolutive noise and enhanced by BSS algorithms. The evaluation was done with respect to human listeners and subjective scores, comparing PEAQ with several other objective quality measures. In this article, we report the performance of PESQ after convolutive BSS for ASR tasks, which as far as we know, has not been reported in other works. In the next section, the measures compared will be introduced, and the experiments described. Next, the results will be presented. Finally a short discussion of the results and conclusions will end the article.

2 Methods

2.1 Selected measures

In this work we will compare the best three measures of our previous work with the PESQ measure. We will also explore a nonlinear mapping of the PESQ, and the use of composite measures based on multivariate regression to improve the correlation. For completeness we will re-write the definitions of the previously used measures, and will introduce the PESQ in more detail. 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.

(1) 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 [6]:

$$d_{WSS}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = K_{spl}(K - \widehat{K}) + \sum_{l=1}^{B} \bar{w}[l] \left(SL[l;m] - \widehat{SL}[l;m]\right)^2, \quad (1)$$

where K_{spl} is a constant that weights the 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:

$$w[l] = \left(\frac{C_{loc}}{C_{loc} + \Delta_{loc}[l]}\right)\left(\frac{C_{glob}}{C_{glob} + \Delta_{glob}[l]}\right) \quad (2)$$

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 [6,12]).

(2) 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 [3] [2]:

$$d_{TRD}(\mathbf{s}, \widehat{\mathbf{s}}; m) = \frac{\left\| \mathbf{e}_m^I + \mathbf{e}_m^N + \mathbf{e}_m^A \right\|^2}{\left\| \mathbf{s}_m^D \right\|^2}. \tag{3}$$

(3) 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 [13]:

$$d_{CD}(\mathbf{s}_m, \widehat{\mathbf{s}}_m) = \sum_{l=1}^{L} \left(c_m[l] - \widehat{c}_m[l] \right)^2. \tag{4}$$

(4) PESQ: This measure uses several levels of analysis in an attempt to mimic the human perception. Figure 1 presents a block diagram of PESQ calculation. The first stage is a gain/delay compensation. The gain for both

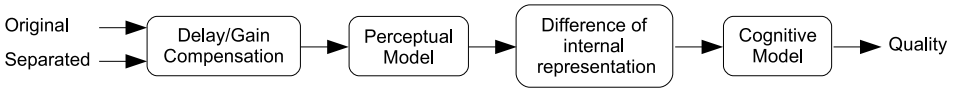

Figure 1. Scheme of the stages in PESQ calculation.

signals is adjusted to produce a constant prescribed power after band-pass filtering. The delay compensation follows at two levels. First at a general level, cross-correlation of envelopes is used to estimate a general delay between the original and the processed (separated) signal. Then a finer correlation/histogram-based algorithm is used to estimate delays for individual utterances. This produce a time-varying delay estimation for segments of the sentences.

The second stage is a transformation to a perceptual domain. This is made by a short-time Fourier transform, followed by a band integration using a Bark scale filterbank, to get a warped pitch-scaled representation. Then, a time-variant gain compensation is applied. The power densities of the original and the separated speech are transformed to the Sone loudness scale using the Zwicker's law.

In the third stage, the disturbance density is calculated by difference of the distorted and reference loudness density.

For the final stage, some cognitive models are applied. The disturbance density $D(f,t)$ is thresholded to account for masking thresholds of the auditory system. Also a second disturbance distribution is calculated, the asymmetrical disturbance density $DA(f,t)$, to take into account that some speech codec introduce time-frequency information in places were it was not present. To account for this, an asymmetry factor is calculated,

[3] Actually, in the original reference the measure defined is the Source to Distortion Ratio (SDR), which is related with TRD by $SDR = -10\log_{10}(TRD)$. SDR has a correlation of 0.77, compared to a correlation of 0.84 for TRD.

that has a large value if the separated signal is markedly different than the original one. The asymmetrical disturbance density is calculated as the product of the disturbance density with the asymmetry factor.

Both densities are integrated over frequencies using two different L_p norms.They are aggregate in segments of 20 frames using a L_6 norm and after that, again aggregated for the whole signal using a L_2 norm, thus producing two values, one for the disturbance D and other for the asymmetrical disturbance DA. The final score is calculated as $PESQ = 4.5 - \alpha D - \beta DA$, with $\alpha = 0.1$ and $\beta = 0.0309$. This produces a value that is between -0.5 and 4.5 and is called raw PESQ score [7].

The correlation of this measure with the subjective perceived quality measured using a MOS scale was evaluated over a wide range of distortions and speech codecs and in different languages, yielding correlation values larger than 0.90 in most cases [14].

(5) Nonlinear PESQ (NLP): In several works, a nonlinear mapping is used to improve the correlation of the PESQ with the target measure. The standard ITU P.862 [7] recommends a logistic function that maps the raw PESQ scores to another measure that correlates very well with subjective quality (as measured by MOS tests). Motivated by this, we propose here to use a nonlinear mapping of the form:

$$NLP = \alpha_1 + \frac{\alpha_2}{1 + \exp^{(\alpha_3 PESQ + \alpha_4)}}, \tag{5}$$

where the coefficient vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ is adjusted from the data to yield maximum correlation. This adjustment can be obtained by nonlinear regression methods.

(6) Composite measures: To improve the correlation even more, a linear combination of several measures can be constructed. This composite measure is defined as [8]:

$$CM_K = \omega_0 + \sum_{i=1}^{K} \omega_i M_i, \tag{6}$$

where M_i denotes each of the selected measures that will be used for the linear combination, K denotes the number of measures combined, and ω_i are coefficients adjusted to maximize the correlation. The values of the coefficients can be obtained by multivariate linear regression.

2.2 Experiments

We will shortly describe the experimental setup here, for more detailed description please refer to [5]. The experiment consisted in the determination of the Pearson's correlation coefficient for the quality measures as predictors of the recognition rate of the ASR system. For the recognition task, a database of Japanese speech was used. The source speech signals were replayed in a

real room, being simultaneously interfered by some noise source. Two loudspeakers were used and the resulting sound field was recorded with a pair of measurement omnidirectional microphones. The power of the desired and noise sources were adjusted to produce two different power ratios (0 dB and 6 dB). Three different kinds of noises were used (Computer noise, TV noise and Speech). The recordings were made in a room in which different amount of reflecting materials were added, to produce, with the same source-microphone locations, three different reverberation times.

The resulting recordings were separated using a standard frequency-domain BSS algorithm [5] using a combination of Jade and FastICA algorithms for each frequency bin. After separation, the signal more similar to the desired source was selected (by means of a correlation). This signal was feed to Julius, a standard ASR system for Japanese language. This system is a two-pass HMM recognizer, that uses acoustics models trained from around 20000 sentences uttered by 132 speakers of each gender. The word recognition rate of the system was evaluated for the resulting separated signals, in each experimental condition. The quality of the separated signal was also evaluated using the different quality measures. For WSS, CD and TRD, a frame size of 512 samples was used. For WSS, the values of constants used were $B = 36$, $K_{spl} = 0$, $C_{loc} = 1$ and $C_{glob} = 20$. For the CD measure, the cepstrum was truncated at $C = 50$ coefficients. For PESQ evaluation, the standard implementation available at the ITU web site was used [4]

3 Results and Discussion

The correlation was evaluated with the results grouped at different levels: for each reverberation condition, for each kind of noise (computer, TV and speech noise), and also considering all noises and reverberation conditions. Finally, a value of global correlation considering all reverberation conditions and all noise kinds was calculated, together with the standard deviation of the regression residual, σ_r.

The NLP measure and the composite measures depend on coefficients that must be adjusted from the data. For this task, we used the data corresponding to all combinations of noise and reverberation conditions. After obtaining the coefficients, the correlation of groups of data corresponding to each reverberation condition and each noise kind was evaluated.

For the NLP measure, the data were used to solve a nonlinear regression

[4] http://www.itu.int/rec/T-REC-P.862/en

problem using least squares. The best coefficient vector obtained was:

$$\boldsymbol{\alpha} = [-1.7558, 63.2960, -1.6962, 5.2361].$$

For the composite measures, we evaluated all possible combinations of the measures, taking them by pairs, by terns, by quartets, and all five (WSS, TRD, CD, PESQ and NLP), for a total of 26 composite measures. We selected the best composite measure for each case of $K = 2, 3, 4, 5$. The results are summarized in Table 1, which shows in the left side the results for each individual measure, and in the right side, those corresponding to the composite measures. The best single measure for each case was marked in bold letters.

Table 1
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. WSS, TRD and CD results adapted from [5]

Reverb.	Noise	WSS	TRD	CD	PESQ	NLP	CM_2	CM_3	CM_4	CM_5
Low	Comp	0.92	0.88	0.81	0.95	**0.96**	0.96	0.97	0.97	0.97
	TV	0.84	0.77	0.78	0.89	**0.90**	0.89	0.91	0.91	0.91
	Speech	**0.84**	0.62	0.79	0.77	0.76	0.77	0.78	0.78	0.79
	All	0.86	0.75	0.77	**0.87**	**0.87**	0.87	0.89	0.89	0.89
Medium	Comp	0.88	0.82	0.74	0.93	**0.94**	0.93	0.95	0.95	0.95
	TV	0.90	0.86	0.91	0.91	**0.96**	0.96	0.97	0.97	0.97
	Speech	0.66	**0.85**	0.76	0.71	0.77	0.76	0.72	0.73	0.73
	All	0.77	0.83	0.74	0.87	**0.91**	0.91	0.91	0.91	0.91
High	Comp	0.83	0.85	0.81	0.85	**0.88**	0.88	0.87	0.88	0.88
	TV	0.93	0.90	0.93	0.94	**0.97**	0.97	0.97	0.97	0.97
	Speech	0.77	0.72	0.79	0.79	**0.82**	0.82	0.82	0.83	0.82
	All	0.81	0.84	0.84	0.84	**0.87**	0.87	0.87	0.87	0.87
$\|\rho\|$	ALL	0.83	0.84	0.76	0.88	**0.90**	0.90	0.91	0.91	0.91
σ_r	ALL	4.98	5.74	6.42	5.02	**3.59**	3.57	3.53	3.51	3.50

Figure 2 shows a scatter graph for the case of all variables together, for the five single measures considered in this study, and the best composite measure CM_5.

As can be seen in the results, the NLP measure was the best single measure in almost all cases, having the lowest variance in the global evaluation. The global correlation is 0.90, which is a very good value for this measure. The next best measure is the raw PESQ score, which achieved a correlation of 0.88, although it has a larger variance than the nonlinear version.

Table 1 also shows that both PESQ and NLP measures have a very stable performance when varying the mixing conditions and with different noises, and

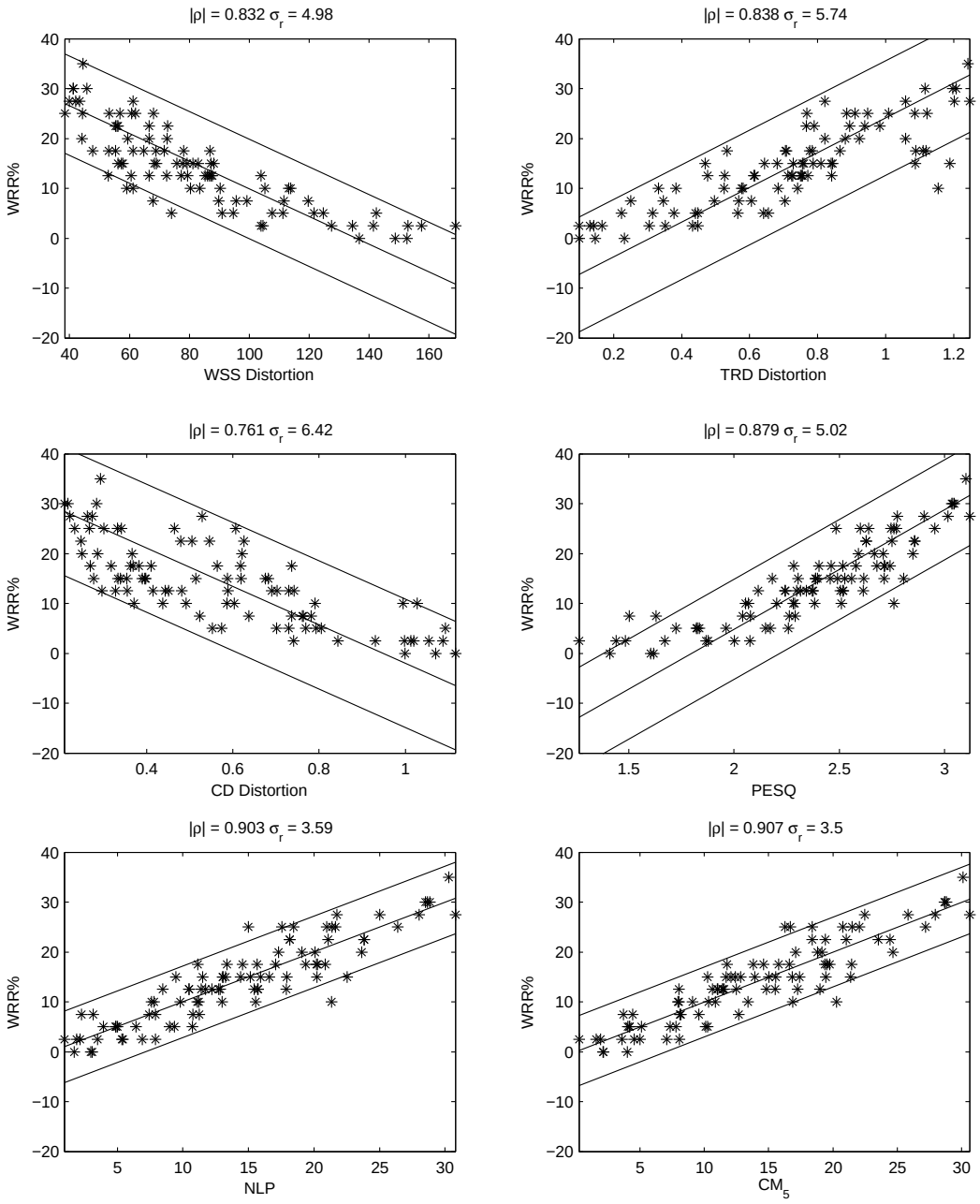

Figure 2. Regression analysis of quality measures for all experimental cases. WRR%, word recognition rate

so the measures seemed to be quite robust to variations of the experimental conditions.

It is interesting to note that the worst performance of the raw PESQ measure was obtained in evaluation the results with competing speech as noise. This can be due to the fact that eliminating a competing voice is probably one of the hardest problems [3]. As parts of the competing speech are present in the separated signal, the recognizer can confuse some phonemes. At the same time,

as the desired source is present in a good level, the output of the perceptual model will be not so different and the quality obtained would be good.

With respect to the composite measures, it can be seen that they do not provide a significant improvement in correlation with respect to NLP. NLP appears as a component in all the composite measures, and the addition of other measures provides only a marginal improvement. Given the increased cost related to the evaluation of several measures and combining them, it seems to be better to just use the NLP measure instead.

These results overcame a limitation of our previous work, in which we could not determine a single measure that performed well in all reverberation conditions. Being PESQ a perceptual measure designed to "mimic" the auditory process, it is highly correlated with subjective quality evaluations [7]. This shows that evaluating the results of a BSS algorithm using PESQ (either in raw PESQ mode or the nonlinear version NLP) will be very powerful, because a good score would suggest both, a high recognition rate and a high perceptual quality.

4 Conclusions

In this contribution, our previous work on objective quality measures for evaluation of BSS algorithms was extended to the ITU-P862 standard measure, PESQ. This measure is gaining strength for the evaluation of speech quality in many different tasks. We found that this measure outperformed all the previous studied objective quality measures for this specific task, which motivated the publication of the present update.

As said in our original paper, it must be noted that the present evaluation was done using a specific BSS algorithm and a specific speech recognition system, which strictly speaking means that the results are only valid for this case. Nevertheless, the BSS algorithm used was a standard frequency domain algorithm (the most widely used), and the speech recognizer was a standard MFCC-based HMM recognizer, so we expect these results to be applicable to other combinations of BSS algorithms/recognition systems with similar characteristics.

We strongly recommend the use of PESQ or NLP for the evaluation of BSS algorithms for speech recognition. These measures can be used alone or with other measures that evaluate other aspects of the separation algorithm as well, like SIR to account for the separation capabilities of the algorithms. Moreover, as in our previous paper, we discourage the use of SNR as it correlates poorly with recognition rates and is very sensitive to experimental conditions and very common artifacts in BSS like fractional delays between the signals to

evaluate.

References

[1] 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.

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

[3] D. P. Ellis, Evaluating speech separation systems, in: P. Divenyi (Ed.), Speech Separation by Humans and Machines, Kluwer Academic Press, 2005, Ch. 20, pp. 295–304.

[4] M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra, A survey of convolutive blind source separation methods, in: J. Benesty, M. M. Sondhi, Y. Huang (Eds.), Springer Handbook of Speech Processing, Springer-Verlag New York, Inc., 2007.

[5] L. Di Persia, M. Yanagida, H. L. Rufiner, D. Milone, Objective quality evaluation in blind source separation for speech recognition in a real room, Signal Processing 87 (8) (2007) 1951–1965.

[6] 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.

[7] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU-T Recommendation P.862.

[8] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on Audio, Speech, and Language Processing 16 (1) (2008) 229–238.

[9] H. Sun, L. Shue, J. Chen, Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. I, 2004, pp. 865–868.

[10] T. Yamada, M. Kumakura, N. Kitawaki, Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice, IEEE Transactions on Audio, Speech, and Language Processing 14 (6) (2006) 2006–2013.

[11] B. Fox, A. Sabin, B. Pardo, A. Zopf, Modeling perceptual similarity of audio signals for blind source separation evaluation, in: Proceedings of ICA2007, 2007, pp. 454–461.

[12] 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.

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

[14] A. Rix, J. Beerends, M. Hollier, A. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2001, pp. 749–752.