QUALITY ASSESSMENT FOR SINGLE CHANNEL SOURCE SEPARATION

Vladimir Sedlak[1], Daniela Durackova[1], Tomas Kovacik[1] and Roman Zalusky[1]

[1] Institute of Electronics and Photonics, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava

E-mail: vladimir.sedlak@stuba.sk

Received 30 April 2013; accepted 14 May 2013

1. Introduction

There are many signal processing applications where the desired signal is corrupted by highly correlated noise sources. Separating such signals from their mixture has often been considered as one of the most challenging research topics in the area of signal enhancement. Source separation methods are divided into two grounds: blind and non-blind. The first group, usually called Blind Source Separation (BSS) covers methods for separation of completely unknown sources without using additional information about character of source or mixing procedure. These methods are typically based on the assumption that the sources are non-redundant and very often used statistical independence, de-correlation and others. Non-blind source separation means the separation of sources for which further information is available.

Quality assessment in this area is very important step because there are many algorithms, techniques and procedures suitable for solving the problem of source separation. And so, in our article are shown and compared the most often used metrics (mostly objective metrics) for source separation.

The rest of paper is structured as follows: In the next section, we present brief introduction to single channel source separation. In the Section 3 we present metrics for quality assessment. In the Section 4 are shown our result and Section 5 concludes work.

2. Single channel source separation

The separation procedure is depicted graphically in figure 1 and is divided into two main blocks. Top block shows process of data acquisition, in other words the creation of mixtures and signal capture. Sources or original signals produce input signal $s(t)$ which can be written in vector notation as $\mathbf{s} = g(x_1,\ldots,x_K)$, where g is some possibly non-linear and stochastic mixing process. The bottom block represents data processing to achieve or estimate original signals from mixture and usually is based on data filtering, data decomposition and grouping or on source modeling.

The approaches appropriate for solving this issue (single-channel separation) can be divided into two groups: 1) model based method, and 2) source driven or computational auditory scene analysis (CASA)-based method. The first group, model-based separation system is based on statistical models including vector quantization (VQ) [1] or Gaussian mixture models (GMMs) [2]. The CASA-based methods search auditory scenes in the time-frequency domain which are probably to come from the same sources of speech signals by exploiting the characteristics of human auditory system [3]. The CASA-based methods rely on extracting psychoacoustics cues from the given mixed signals and work in two stages: segmentation and grounding.

In our experiments we used separation procedure based on estimation of ideal (binary) mask. This technique belongs to source driven approaches which were mentioned earlier and can be used for speech enhancement or separation of two speaker's signals. The binary mask (in the real condition) needs to be estimated from noisy input signal, and that is a challenging task, particularly in adverse noisy conditions. The main issue is how accurate do we need to estimate the binary mask without affecting speech intelligibility. Other factors that may influence intelligibility of speech synthesized by the ideal mask include the choice of local SNR (Signal-to-Noise Ratio), the masker type, speech materials and SNR of input signal.

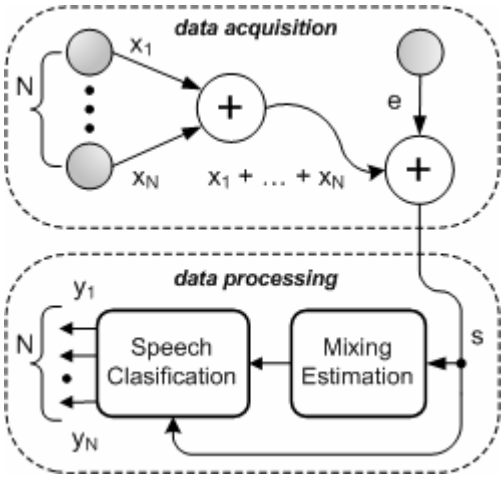

Fig.1: *Flow chart of single-channel source separation*

For estimation of ideal mask we used technique described in [4]. At beginning is input signal divided into frames of constant length (20 ms) with 50% overlap between segments. These segments are corrected by window function (Hamming function) in the next step and finally a fast Fourier transformation (FFT) is applied to segments. Segmentation and FFT produce time-frequency (T-F) representation of the input signal which is used to compute local SNR. Each local SNR is compared with a threshold value to determine whether to retain the T-F (binary mask value is 1) or to eliminate it (binary mask value is 0). Then is this mask applied to the FFT magnitude spectrum of mixture signal and finally inverse FFT together with overlap-and-add method are used to get desired signal from mixture.

3. Quality assessment

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

Objective methods can be classified into intrusive (reference) measures and non-intrusive (non-reference) measures. The intrusive measures compare the output signal (distorted signal) with the original signal, which is usually called the reference signal. The non-intrusive methods do not require a reference signal because the speech quality is determined only by the output speech signal. In general, objective speech quality measures can be categorized into three domains: time domain, spectral domain or perceptual domain.

Based on literature review and articles [5, 6] we made decision for using the following metrics in our experiments:

- ***Segmental SNR (SNRseg)*:** segmental version of SNR.
- ***Log-Likelihood Radio (LLR)***: statistical test used to compare the fit of two signals.
- ***Weighted Spectral Slope (WSS)***: measures the weighted differences of spectral slope over 25 critical frequency bands between the two corresponding signal frames.
- ***Perceptual Evaluation of Speech Quality (PESQ)*** [7]*:* metric is recommended by ITU-T P. 862 for speech quality assessment.
- ***BBS EVAL metrics*** [6]*:* consist of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artificial Ratio (SAR).

4. Experiments

For experiments we used the same signal as in our previous work [8]. Speech signals were obtained from IEEE sentence database and were originally sampled at 24 kHz and down-sampled to 8 kHz. Noise signals were taken from the AURORA database (Hirsh, 2000) which includes different types of noise (car, exhibition hall, restaurant, etc.).

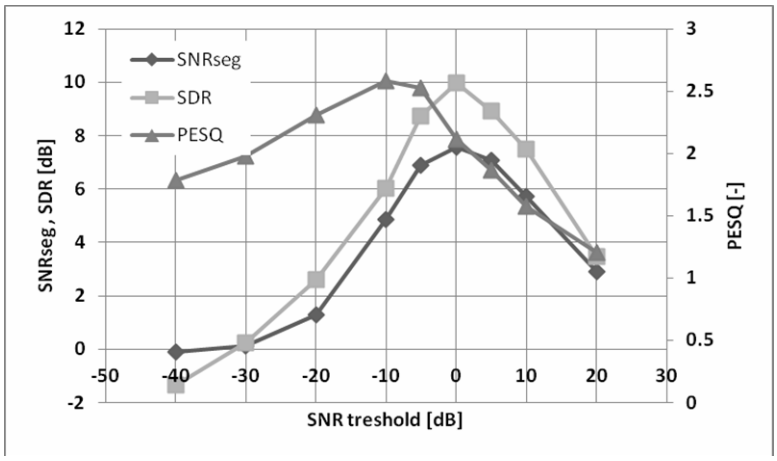

Fig.2*: Ideal binary mask performance versus input sources ratio*

Because chosen separation algorithm is based on a comparison of local SNR with selected threshold and its value has significantly impact on performance, as we can see in [4], first step was verifying result under different values of this threshold (figure 2).

Tab. 1. *Achieved result under different types and levels of input noise*

Noise level	Noise type	SNRseg	LLR	WSS	PESQ	SIR	SAR	SDR
0 dB	train	-0.64	0.16	31.22	1.21	10.14	-3.04	-6.63
	car	-1.03	0.11	27.44	1.01	8.25	-3.65	-4.49
	street	0.31	0.12	27.35	1.45	7.74	-2.68	-3.68
5 dB	train	-0.06	0.14	28.81	1.67	10.67	0.27	-0.43
	car	-0.05	0.15	28.02	1.77	10.28	0.34	-0.44
	street	0.46	0.19	34.72	1.67	10.7	1.08	0.31
10 dB	train	1.52	0.22	32.42	1.78	12.24	3.97	7.88
	car	1.25	0.18	33.52	1.89	12.1	3.71	8.34
	street	2.01	0.24	35.32	2.29	11.25	4.33	7.65

Results depicted in figure 2 reveal excepted trend (threshold value cannot be very large and neither very small) of performance, so based on this results was selected suitable value for other experiments. In our case in was -10dB.

Input or mixed signals for all experiments were created artificially with constant mixing ration. We decided to use parameter Signal-to-Signal Ratio (SSR) which measures differences between signals. Mixing procedure consist with selection of primary (target) signal to which is added masker signal. SSR between these signals was set to 5dB.

In table 1 are presented result obtained under different types and levels of input noise. Separated or output signals were compared with original clean signals (without additive noise) and local SNR threshold was set to -10 dB. Based on these results we can say that the best performance was obtained when the input signal was corrupted by street noise. The unit of metrics SNSseg, SIR, SAR, and SDR is dB, others metrics are without unit.

5. Conclusion

In this work are presented metrics suitable for quality assessment of algorithms for separation of signals. For experiments the input signal was created as mixture of multiple signals with known level of noise. The aim was not only review of metrics but also show expected values of these metrics in case, where input signals are corrupted by noise. Based on our subjective assessment we can say that the best match between our quality decision and presented objective methods was achieved in case of PESQ. The disadvantage of this method is its complexity and computational demands.

Acknowledgement

This work is resulting from the project VEGA 1/0987/12 sponsored by Ministry of Education, Slovak Republic.

References:

[1] M. Asgari, M. Fallah, E. A. Mehrizi and A. Mostafavi: A VQ-based Single Channel Audio Separation for Music/Speech Mixtures, In: *International Conference on Computer Modelling and Simulation*, Brno, Czech Republic, 89 (2009).

[2] H. Wang, Y. Wang, W. Wang, B. Zhu and S. Ma: Single Channel Polyphonic Music Signal Separation Based on Bayesian Harmonic Model, In: *International Conference on Image and Signal Processing*, Beijing, China, 2784 (2011).

[3] Y. K. Lee, I. S. Lee and O. W. Kwon: Single Channel Speech Separation Using Phase-Based Methods, IEEE Transaction on Consumer Electronics, vol. **4**, 16 (2010).

[4] N. Li and P. Liozou: Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction: Journal of Acoustical Society of America, **123**, 1673 (2008).

[5] K. Kokkinakis and P. C. Loizou: Evaluation of Objective Measures For Quality Assessment of Reverberant Speech, In: *ICASSP 2011*, Prague, Czech Republic, 2420 (2011).

[6] P. Mowlaee, R. Saeidi, M. G. Christensen and R. Martin : Subjective and Objective Assessment of Single-Channel Speech Separation Algorithms, In: *ICASSP 2012*, Kyoto, Japan, 69 (2012).

[7] S. Paulsen and T. Uhl: Quantifying the Suitability of Reference for the PESQ Algorithm, In: *Internal Conference on Communication Theory, Reliability and Quality of Service*, Athens, Greece, 110 (2010).

[8] V. Sedlak, D. Durackova and R. Zalusky : Investigation Impact of Environment for Performance of ICA for Speech Separation, In: *Electro 2012*, Rajecke Teplice, Slovakia, 69 (2012)