Improving noisy speech recognition with blind source separation methods: validation with artificial mixtures

Benoit Albouy and Yannick Deville*

Laboratoire d'Acoustique Métrologie Instrumentation, Université Toulouse III
Bât. 3R1B2, 118 Route de Narbonne, 31062 Toulouse Cedex, France

albouy@cict.fr,ydeville@cict.fr

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems are progressively appearing in everyday-life applications [3], but are limited to noiseless environments. On the other hand, emerging Blind Source Separation (BSS) methods provide improved signal denoising. Therefore, we here introduce a new two-stage approach, where BSS is used as a front-end to increase the performance of a subsequent ASR system.

2. CONSIDERED BSS/ASR METHODS

The considered BSS configuration [1-2] involves two signals $Y_1(z)$ and $Y_2(z)$ which are convolutive mixtures of two acoustic sources $X_1(z)$ and $X_2(z)$. In a real setup, $Y_1(z)$ and $Y_2(z)$ are provided by two microphones. We here consider a simplified model of this situation where $Y_1(z)$ and $Y_2(z)$ are artificially derived from $X_1(z)$ and $X_2(z)$, based on:

$$\left\{ \begin{array}{lcl} Y_1(z) & = & X_1(z) + \alpha A_{12}(z) X_2(z) \\ Y_2(z) & = & A_{21}(z) X_1(z) + \alpha X_2(z) \end{array} \right. \qquad (1)$$

where $A_{ij}(z)$ are mixing filters and α controls the ratio of the source powers. To separate these signals, we use two alternative methods [1-2], resp. based on normalized decorrelation and (sub-)optimum separating functions. The two outputs of the BSS front-end here resp. yield speech and noise signals. This speech signal is then applied to an ASR software, i.e. HTK [4]. This software uses Hidden Markov Models (HMM), which are first estimated by the Baum-Welch algorithm during a training phase and then used by the Viterbi algorithm during the recognition phase [3].

3. EXPERIMENTAL RESULTS

The considered sources are: i) speech for voice dialing, containing command words and a digit sequence and ii) in-car recorded noise. The α coefficient is varied in the experiments, whereas we use the two fixed 13th-order strictly causal MA filters $A_{12}(z)$ and $A_{21}(z)$ defined in [1-2].

*The authors would like to thank Jérôme Farinas and Régine André-Obrecht for their support concerning the use of the HTK software.

HTK is here operated so as to model words with HMM and to compute the word recognition rate (WRR). The latter rate is defined as: $WRR = (N - D - S - I)/N$ where N is the real number of words in the sequence, and D, S, I are resp. the numbers of deletion, substitution and insertion errors. Silence insertions are ignored in this parameter.

The WRR in the noiseless case ($\alpha = 0$) defines optimum performance and is equal to 94.74%. The WRR achieved for various noise levels (defined by α) with or without BSS front-ends is summarized in the following table:

α	WRR (%)		
	without BSS	decorrelation BSS	optimum BSS
0.1	84.21	94.74	94.74
0.25	63.16	94.74	94.74
0.5	68.42	94.74	94.74
0.7	68.42	94.74	94.74
1	36.84	94.74	94.74
3	47.37	89.47	94.74
5	36.84	89.47	89.47

4. DISCUSSION AND CONCLUSIONS

The above table shows that the WRR of a plain ASR system tends to decrease very rapidly when the noise level is increased. On the contrary the overall system containing a BSS front-end is completely insensitive to noise up to quite high noise levels (esp. the optimum BSS approach). This demonstrates the usefulness of BSS methods in noisy ASR applications. To validate this approach, we plan to apply it to real mixed signals and more complex ASR tasks, such as large vocabulary continuous speech recognition.

5. REFERENCES

[1-2] N. Charkani, Y. Deville, "Self-adaptive separation of convolutively mixed signals with a recursive structure", Signal Processing, vol. 73, no. 3, pp. 225-254, 1999 and vol. 75, no. 2, pp. 117-140, 1999.
[3] R. De Mori, "Spoken dialogues with computers", Academic Press, London, 1998.
[4] S. Young et al, "The HTK book (for HTK Version 3.0)", Microsoft Corporation, 1999.