Small Microphone Array Design and Processing for Speech Recognition of Vocal Orders in Restaurants

Patrick Marmaroli¹, Philippe Martin², Xavier Falourd¹, Hervé Lissek²
1 : Laboratoire d’Electromagnétisme et d’Acoustique (LEMA), EPFL, Switzerland
2 : AER Sàrl, Lausanne, Switzerland

Summary

Veovox is a project led by a Swiss company Veovox in collaboration with Swiss research institutes whose purpose is to market an order-taking device, enabling a waiter in a restaurant to take orders by voice. With this device, the waiter only needs to pronounce the order to a personal digital assistant (PDA), which directly translates the voice order into text through an integrated voice recognition engine. The text is then submitted to the restaurant management software solution for further processing. We present here the developed audio processing system of such a device developed by the LEMA. In the context of a more or less noisy restaurant composed of distributed parasitic speakers and diffuse noise (cocktail noise), the developed audio processing consists in enhancing the speech of the waiter from the recorded audio signal in order to provide a high-quality signal to the recognition engine.

Array Design and Signal Processing

1. Array Design

Eight microphones which steer in four different directions: forward, backward, left, and right based on differential beamforming theory.

The global array consists in two perpendicular and orientable linear arrays of 4 microphones each.

Three sub-arrays per linear array:
- 3 mic. d = 2 cm, 1.2 - 6 kHz
- 3 mic. d = 4 cm, 600 - 1200 Hz
- 2 mic. d = 8 cm, 300 - 600 Hz

For each sub-array:
- If 2 mic. : 1st order array (cardioid)
- If 3 mic.: 2nd order differential array (cardioid + hypercardioid)

2. Noise Reduction

In order to clean the signal, a statistical noise reduction is done on each channel using the spectral subtraction technique. In the frequency (0Ω) domain and assuming the observation Y is composed of the speech X and an additive noise N, the magnitude of speech can be estimated by the formula:

$$|X(\omega)| = |Y(\omega)| - \hat{N}(\omega)$$

Where $$\hat{N}(\omega)$$ is an estimation of N(\omega).

3. Spatial Masking

This step consists in enhancing the separation between punctual sources distributed in a room (like parasitic speakers).

$$X_{\text{prod}} = \begin{cases} X_{\text{prod}} & \text{if } X_{\text{prod}}^2 \geq n \times \text{max}(X_{\text{prod}}^2, X_{\text{prod},\text{frwd}}, X_{\text{prod},bckwd}, X_{\text{prod},\text{left}}, X_{\text{prod},\text{right}}) \\ 0 & \text{else} \end{cases}$$

The comparison of the four channels in the frequency domain allows the design of adaptive filters, one per channel, which, by spectral subtraction, decrease the signals from other directions.

4. Cube Coherence Masking

In order to reduce distortion due to previous processes, a cube coherence masking which consists in only keeping coherent frequency bins through all the processes applied. $$X_{\text{cube}}$$: spectrum of the raw signal taken on one arbitrary microphone.

$$C_1 = \frac{X_{\text{raw}} - X_{\text{raw}}}{C_1^2}$$
$$C_2 = \frac{X_{\text{prod}} - X_{\text{prod}}}{C_2^2}$$
$$X_{\text{prod}} = X_{\text{prod}} \left( \frac{C_1^2}{C_1^2 + C_2^2} \right)$$

5. VAD

Automatic speech recognition systems generally need an input signal where all what is non speech is eliminated. This is the goal of Voice Activity Detection (VAD) algorithm.

$$D = 10 \log \frac{\sigma_n^2}{\sigma_s^2}$$

H : a boolean variable which is equal to 1 if speech is present and 0 otherwise.

$$H = \begin{cases} 1 & \text{if } D > \lambda \\ 0 & \text{else} \end{cases}$$

Where $$\lambda$$ is a threshold automatically determined or empirically chosen.

Conclusion

This algorithm has been developed in the context of real time audio processing. A first version has been developed on Matlab which has been transposed on Simulink and then implemented on DSP. Sensibility studies are in progress but first results of our experimental tests in different real situations are very promising:

- we can see on both opposite figures the comparison between the raw and the processed signal (in a Leq and in a temporal waveform point of view). An attenuation of 15 dB is obtained with a very low distortion on the speech signal giving a very good score in speech recognition. Further work consists in reducing both the number of microphones and the needed time to well estimate noise spectral properties (approximately 500 ms at 16 kHz today). A patent is pending.

Acknowledgments

This project has been financially supported by the Swiss Innovation Promotion Agency, under grant agreements 77662.1 ESPS-ES and 9628.1 PFES-ES. Our scientific partners in this project are Idiap Research Institute (Martigny, Switzerland) and the HEVS (Sion, Switzerland). Thanks to the Ecole Hôtelière de Lausanne (EHL) for their help for real situation recordings.

3rd Workshop on Speech in Noise: Intelligibility and Quality, 6-7 January 2011, Lyon, France