Source Localization based on Computational Auditory Scene Analysis

Václav Bouše
Siemens Audiologische Technik GmbH
91058 Erlangen, Germany

Rainer Martin
Ruhr-Universität Bochum
44801 Bochum, Germany

Introduction

- Motivation for **mCASA** – model-based Computational Auditory Scene Analysis [3]
  - To describe the acoustic scene (Fig. 2) in terms of spatial distribution of sources and their classification (as e.g. speech, music, noise, etc), using an a priori model of the detected signals
  - Usable for: Optimally activating and controlling of hearing aid (HA) algorithms (e.g. steering of a beamformer, HA program switching, ...)
- Objectives
  - **Speech localization of a single talker** (Binaural HA configuration)
  - Based on vowel detection – the characteristic components of human speech
  - Extendable system in order to detect other classes (e.g. music, noise…)

Basic principle

Processing steps

General approach	Implemented approach for speech localization
Signal representation	T-F domain
Detect signal fragments using a model of the desired signal class	Detect frames with obvious vowels
Isolate these fragments	Extract (T-F bins) with vowels based on their formant structure
Localize these fragments	GCC-PHAT
Calculate the image of the acoustical scene	Averaging localization results over time

System description

- Frame-by-frame processing in T-F domain
- Block diagram:

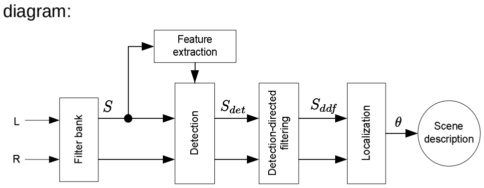

Fig. 1. System block diagram.

- Processing steps:

1. **Transformation of the input signals in the T-F domain** using a Short-time Fourier Transformation (STFT)

$$S_{\{l,r\}}(l,\Omega) = \mathrm{STFT}\left\{s_{\{l,r\}}(k)\right\}$$

2. **Feature extraction** for the vowel detection in each time-frame *l*.
   - Harmonicity (*R*) – estimated by the modified ACF method [2]
   - Positions of the two first formants (f_1 f_2) – estimated by the LPC analysis
   - Signal power (*P*)

$$\mathbf{X}(l) = \left[\ R(l) \quad f_1(l) \quad f_2(l)\ \right]$$

3. **Detection (DET) of time-frames with obvious vowels**
A hard-decision, based on the possible ranges $\mathcal{X}$ of harmonicity, signal power and the positions of the first two formants, i.e. only the frames with vowels are kept in the output signal of this block.

$$S_{det\{l,r\}}(l,\Omega) = M_{det\{l,r\}}(l) \cdot S_{\{l,r\}}(l,\Omega), \text{ where}$$

$$M_{det\{l,r\}}(l) = \begin{cases} 1 & \mathbf{X}(l) \subset \mathcal{X} \\ 0 & \text{otherwise} \end{cases}$$

4. **Detection-directed filtering (DDF)**
Only the T-F bins in a specific range B_i around the formants f_i are kept for the further processing, other bins are discarded.

$$S_{ddf\{l,r\}}(l,\Omega) = M_{ddf\{l,r\}}(l,\Omega) \cdot S_{det\{l,r\}}(l,\Omega), \text{ where}$$

$$M_{ddf\{l,r\}}(l,\Omega) = \begin{cases} 1 & \Omega \subset \bigcup_i f_i \pm B_i \\ 0 & \text{otherwise} \end{cases}$$

5. **Localization**
The GCC-PHAT localization method [2] is applied to the output signal from the DDF block.

$$\theta(l) = \mathrm{GCC}\left(S_{ddf\{l,r\}}(l,\Omega)\right)$$

6. **Scene description**
The position of the speech source
= averaging the estimated angle in each frame over a specified time interval (assuming spatial stationary of the speech source)

Evaluation

- **Test Configuration – Fig. 2**
  - Binaural configuration of behind-the-ear (BTE) hearing aids, simulated using the HRTF
  - Variable power of the speech source, constant power of the interferers
  - SNR measurement evaluates the same microphone signal as the detection algorithm

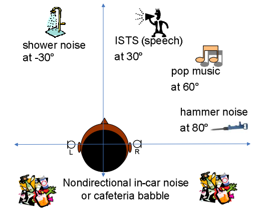

Fig. 2. Test scene with all test sources presented.

- **Function demonstration – Fig. 3**
  - Applying the localization method on the input signal, on the signal after the DET and after the DDF block
  - Results (example at SNR = -5dB):
    - Both blocks are beneficial
    - Speech source peak: ↑
    - Interferers' peaks: ↓
    - Global maximum after the DDF block corresponds to the speech source.

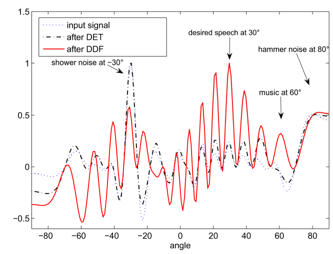

Fig. 3. Localization performed on the input signal, after the detection block (DET) and after the detection-directed filtering (DDF) block. SNR = -5dB.

- **Speech localization in omnidirectional noise – Fig. 4**
  - Detection performance without the side-effects of directional sources on the localization
  - Applying the localization method on the signal after the DET and after the DDF block.
  - Results:
    - Type of the noise (either in-car noise or a cafeteria babble) has a substantial influence on the system performance
    - DDF improves the localization in the SNR range between -11 and +9 dB

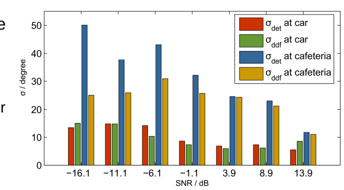

Fig. 4. Standard deviations of localization after the detection (DET) block and the detection-directed filtering (DDF) block.

- **Speech localization in a complex auditory scene – Fig. 5**
  - Combinations of test signals
  - Various SNR
  - Correct localization = the localization error is less than 5 degree
  - Results (correctness of the localization):
    - SNR > 10 dB:
      Both methods perform well
    - SNR ~ 0 dB:
      DDF still improves performance
    - SNR < -10 dB:
      Both methods fail

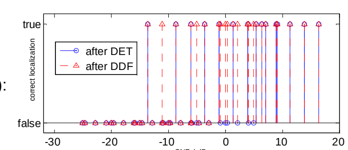

Fig. 5. Correctness of the localization after the detection (DET) block and the detection-directed filtering (DDF) block as a function of SNR.

Conclusion

- This work introduces a general framework for localization of acoustical sources
- Currently: speech localization based on vowel detection
- Reliable speech localization down to SNR = 0dB, method fails for SNR < -10dB
- Outlook:
  - Dropping the spatial stationarity and better ear assumption
  - System extension for other sound classes (e.g. music)

Acknowledgment

- This work was supported by the European Commission within the ITN AUDIS, grant agreement number PITNGA-2008-214699.

References

[1] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In IFA Proceedings, volume 17, pages 97–110, 1993.
[2] Nilesh Madhu and Rainer Martin. Acoustic Source Localization with Microphone Arrays. In Rainer Martin, Ulrich Heute, and Christiane Antweiler, editors, Advances in Digital Speech Transmission, chapter 6. Wiley, 2008.
[3] DeLiang Wang and Guy J. Brown. Computational Auditory Scene Analysis. John Wiley & Sons, 2006.

contact: mail@vbouse.com