ROBUST SPEAKER LOCALIZATION USING A MICROPHONE ARRAY

Norbert Strobel and Rudolf Rabenstein
Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
e-mail: {strobel, rabe}@LNT.de

ABSTRACT

This paper presents a speaker localization system using a microphone array. The array is operated as a steered filter-and-sum beamformer implemented as a summed correlator. In particular, we emphasize the use of a speech pause detector to improve the robustness of the speaker localization system by avoiding erroneous position estimates when no speech signal is present. Simulation results and measurements show that speech pause detection improves the overall system performance considerably.

1 Introduction

For acoustic speaker localization, we use a number of spatially distributed microphone sensors to capture incoming sound waves. If the amplitude gradient across the microphone array is negligible, the time delays between different microphone signals can be expressed in terms of the unknown source location parameters.

Speaker localization is related to the classical sonar target detection problem already described in [1, 2]. However, these early references are not directly applicable to our problem, since (1) they offer little advise about how to efficiently implement a steered beamformer without special-purpose hardware, and (2) they do not take into account nonstationary signals such as speech.

We cover solutions to both problems. To this end, we first concentrate on how to efficiently implement the traditional ML estimator (steered beamformer) for a source position such that its output energy can be obtained by sampling cross-correlation functions and summing the results. This way, we no longer need to actively manipulate a delay chain when recording input signals. Then, we briefly discuss a hierarchical search strategy to reduce the complexity even further. In addition, we introduce a speech pause detector to avoid erroneous source position estimates when no speech signal is present. Finally, experimental results are presented and some conclusions are drawn.

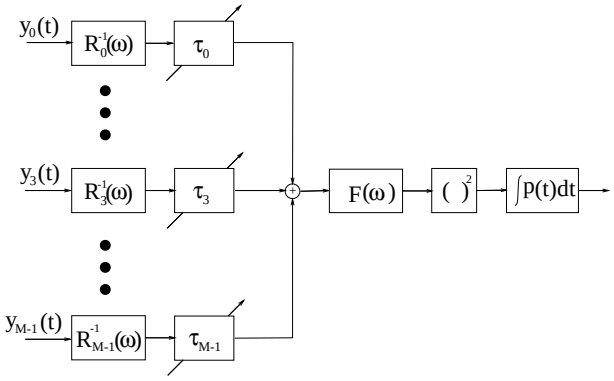

Figure 1: Maximum likelihood estimator for a speaker position. A time-domain implementation of a steered filter-and-sum beamformer is shown. A "beam" is formed by the array and "swept" over the area that is of interest to the observer. The ML estimate for the source position is obtained for that set of steering delays that yields the maximum output energy.

2 Summed-Correlator Beamformer

In Fig. 1, the classical ML estimator for a source position is shown. Assuming that the noise components at the M microphones are mutually uncorrelated, a ML estimate for some source position can be obtained by forming a "beam" by the array and "sweeping" it over all potential speaker positions by suitably adjusting the time delays. The ML estimate for the source position follows for that set of associated steering delays that yields the maximum output energy.

Although the arrangement in Fig. 1 provides a rather intuitive interpretation of the ML estimator, it is not very attractive from an implementation point of view. Even to efficiently steer a microphone array into multiple directions requires considerable ingenuity [3].

To circumvent the problem of how to quickly and efficiently delay the multiple microphone signals, we recommend a summed-correlator implementation of the

steered beamformer. This eliminates the need for a sophisticated input delay chain. Apart from implementation advantages, the proposed structure also offers the opportunity to consider acoustic multipath propagation. A detailed description of this method can be found in [4].

3 Sequential Source Localization Using Sub-Arrays

A potential drawback to any steered beamformer approach is the fact that we have to focus at all potential speaker positions. Depending on the spatial accuracy desired, the search complexity may be considerable. We found it advantageous to decompose a microphone array into subarrays operating in a sequential search mode. This strategy maximizes the overall summed-correlator output by sequentially maximizing one short-time cross correlation function at a time. Assume, for example, that there are two subarrays only. Then the subarray with the smallest aperture is used first to estimate the source bearing. Afterwards the subarray whose microphones are spaced furthest apart may be applied to determine the source range. The bearing and range estimates now provide an initial location for the third step during which the overall beamformer is used to find the final position estimate within a small area around the initial location.

4 Speech Pause Detector

When a speaker pauses, no speech signal is available. Then a steered beamformer cannot produce a correct estimate for the speaker position. It is thus necessary to detect speech pauses to reduce the number of erroneous position estimates. To this end, a the use of a speech pause detector (SPD) is proposed. We chose a simplified version of the voice activity detector initially developed for the GSM network [5]. It is based on two hypotheses:

- H_0: speaker not active (speech pause)
- H_1: speaker active (no speech pause)

In the first case, the maximum of the short-time cross-correlation function (ccf) between two microphone signals reflects the noise energy, whereas in the second case, it comprises the sum of speech signal energy and noise energy. Thus, for each frame of the input signal, the maximum of the short-time ccf can be compared with a threshold to distinguish between H_0 and H_1. If the maximum exceeds the threshold, then H_1 is detected, otherwise H_0 is assumed to be true. Figure 2 shows the block diagram of the SPD.

To cope with nonstationary background noise, the threshold is adapted whenever no speaker is active. As a consequence, we have to distinguish between noise and speech. Their main difference is that noise is stationary while speech is nonstationary in general. Only vowels are stationary, but they are usually louder than other speech segments. As a result, their short-time signal energy can be used to distinguish them from noise. Thus, to decide if a signal frame represents speech or noise, we first check whether if is stationary by observing the differences between ccf maxima of successive frames. If those do not exceed a threshold, then the signal is assumed to be stationary. If, in addition, the maxima of the ccfs are lower than those of the average speech signal, then we decide that the input signal is indeed noise and not a vowel sound. Figure 3 shows a example of the adaptation process. The upper diagram shows a speech signal sampled at 48 kHz. The lower diagram depicts the maxima of the cross-correlation function (ccf) and the associated SPD threshold for each signal frame. We see that the SPD threshold mainly changes when there is noise. It is kept at its current level when speech is present.

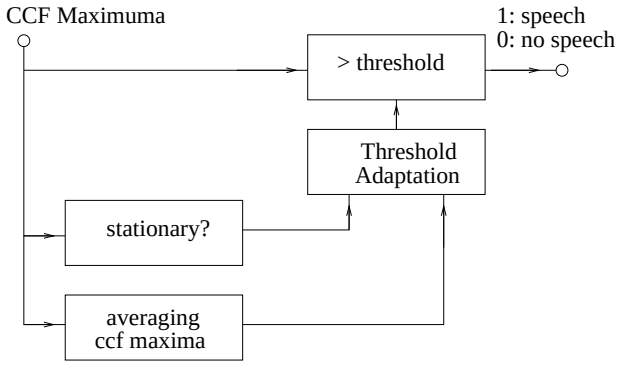

Figure 2: SPD Block Diagram: For each signal frame the maximum of its associated short-time ccf is compared to a threshold. If the maximum exceeds a threshold, then speech is detected.

The performance of a detector can be analyzed by comparing the probability of detection

$$P_D = \Pr\{H_0|H_0\} \tag{1}$$

with the probability of false alarm

$$P_F = \Pr\{H_0|H_1\}. \tag{2}$$

The usual way these quantities are discussed is through a parametric plot of P_D versus P_F: the receiver operating characteristic (ROC).

Since the probabilities P_D and P_F are not easy to compute for arbitrary real speech signals due to the lack of a probabilistic model, an indirect approach was used to estimate them. Placing a loudspeaker at a known position, it was assumed that it was producing speech,

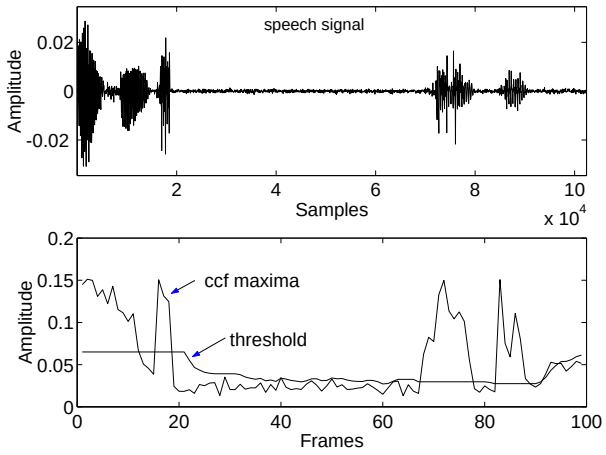

Figure 3: Threshold adaptation: The upper diagram shows a speech signal sampled at 48 kHz. The lower diagram depicts the maxima of the cross-correlation function (ccf) and the associated SPD threshold for each signal frame. We see that the SPD threshold mainly changes when there is no speech signal.

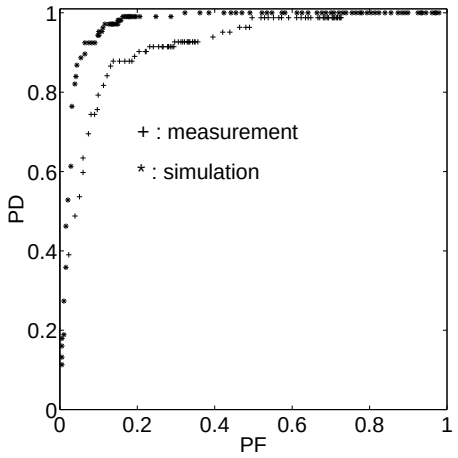

Figure 4: ROC of SPD: The higher the threshold, the more pauses are detected and the larger P_D. A rising threshold, on the other hand, suppresses an increasing number of speech frames as well, and P_F also increases albeit much more slowly than P_D.

if the steered beamformer estimated its position correctly. Wrong position estimates, on the other hand, were contributed to speech pauses. This way speech sequences could be generated comprising segments labeled H_0 and H_1. These annotated speech sequences were subsequently used to evaluate the decisions made by the SPD. Figure 4 shows ROC curves for both simulations and measurements. The higher the threshold, the more pauses are detected, i.e., P_D increases. Unfortunately, it also happens that a higher threshold suppresses more and more speech frames who are not spoken loudly enough, i.e., P_F also increases with the threshold. Fortunately P_D rises much faster than P_F. This leaves us with enough room to set a threshold such that almost all pauses are detected. The price we pay are a few speech segments which are misclassified as pauses, although they could have been used to reliably estimate a speaker position. However, for most applications including ours, a small value for P_F is tolerable.

In our beamformer setting, the detector is applied to each microphone pair. If only one detector indicates a speech pause, the associated signal frame will not be used to estimate a speaker position.

5 Experimental Results

The performance of the speaker localization algorithm is examined based on both simulations and real measurements.

5.1 Experimental Setup

The measurements took place in an anechoic chamber which is 2.54 m long, 2.74 m wide, and 2.36 m high. The input signal was sampled with 48 kHz and subdivided into frames with 2048 samples. Successive frames were overlapped by 50%. The source signal level for the simulations was set such that average sensor SNR was 20 dB. For the measurements we estimated a SNR of 30 dB. Both simulations and measurements used a microphone array with six microphones as shown in Fig. 5. The aperture of the three small microphone pairs was 0.15 m, and the center-to-center distance between the small subarrays was 0.6 m. Thus, the overall array aperture was 1.35 m.

Since the accuracy with which a speaker can be localized depends on the speaker position, the search space in front of the linear microphone array was divided into three regions as shown in Fig. 5. An arc with radius 0.30 m divided R_0 and R_1, and every location further away from the array center than 0.90 m was considered to be part of R_2.

5.2 Simulations and Measurements

The performance of our speaker localization system should be primarily inferred from the simulation results. In that case, position estimates were obtained by placing a speaker active for 10 seconds at randomly selected positions within R_0, R_1, and R_2. The proposed speaker localization method was used to estimate a position for each speech frame. The algorithm was considered to

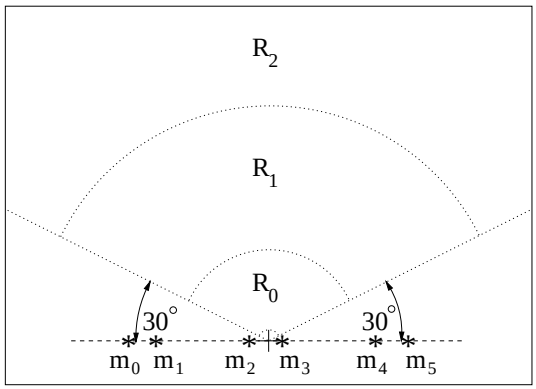

Figure 5: Microphone array configuration and search regions. To show that the the accuracy with which a speaker can be localized depends on the speaker position, the search space in front of the microphone array was divided into three regions.

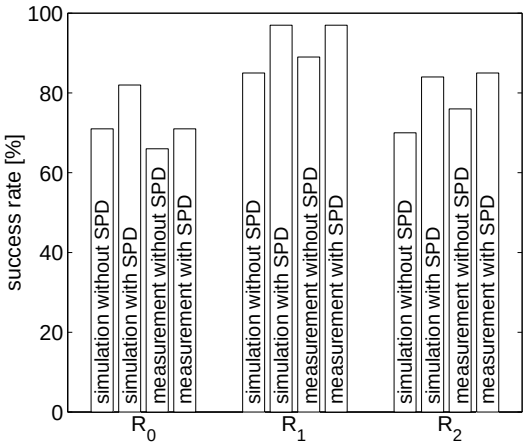

Figure 6: Success rates obtained via simulations and with measurements. Both simulations and measurements were carried out with and without the SPD.

have found the "correct" speaker location if the position estimate was within a radius of 15 cm around the actual speaker position. Since the overall speech signal yields more than 400 individual position estimates, we can compute a ratio of correctly estimated speaker positions to the total number of position estimates for each speaker position. This ratio is termed *success rate*. Since three test region R_i, $i = 0, 1, 2$, exist, we obtain three success rates for simulations without the SPD and three for simulations with the SPD. They are summarized in Fig. 6.

In addition, measurements were taken to verify the simulation results. They comprised ten speaker positions within each of the three test regions. A speech signal, 54 seconds in length, was played at each test position, and sound recording equipment was used to acquire six tracks. The sound tracks were fed into the same localization algorithm as used for the simulations. Position estimates were computed with and without the SPD. The resulting success rates are also part of Fig. 6.

6 Discussion and Conclusions

From Fig. 6, we conclude that the SPD always improves the overall performance of the proposed speaker localization system. This is not surprising, since signal frames identified as pauses are discarded.

Figure 6 also indicates that speakers very near to the array, i.e., in R_0, are hardest to find. Nevertheless, there the success rate is still 82% (71% without SPD). Speaker positions further away but still within R_1 and R_2 can be more reliably estimated. The associated success rates are 97% (85% without SPD) and 84% (70% without SPD), respectively.

Comparing the success rates with and without SPD, we see that the former are significantly more accurate than the latter. In fact, the current level of performance suggests that our overall speaker localization method works well for single speakers within a few meters of a microphone array.

Acknowledgements This work is part of the ongoing Sonderforschungsbereich (SFB) No. 603 being carried out at the University of Erlangen-Nürnberg. It is supported by the Deutsche Forschungsgemeinschaft (DFG).

References

[1] W. R. Hahn. Optimum signal processing for passive sonar range and bearing estimation. *Journal of the Acoustical Society of America*, 58(1):201 – 207, 1975.

[2] G. C. Carter. Time delay estimation for passive sonar signal processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):463 – 470, 1981.

[3] W. Kellermann. A self-steering digital microphone array. In *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3581 – 3584, 1991.

[4] N. Strobel, T. Meier, and R. Rabenstein. Speaker localization using steered filtered-and-sum beamformers. In B. Girod, H. Niemann, and H.-P. Seidel, editors, *Proceedings Vision, Modeling, and Visualization '99*, pages 195–202, Erlangen, 1999.

[5] European Telecommunications Standards Institute. Digital cellular telecommunication system; voice activity detector for enhanced full rate speech traffic channel. http://www.etsi.fr/, March 1997. GSM 06.82.