Distant Speech Recognition
Using Multiple Microphones
in Noisy and Reverberant Environments

Hanna Runer
Institution Electrical and Information Technology (EIT)
Faculty of Engineering (LTH)
Lund University, Sweden.

Abstract

Does a speech recognition device work better, when being spoken to a couple of meters away, if a beamforming technique is used? This article is based on a Master's thesis with the same name.

Introduction

Speech is the most natural and primary way of communication for human beings. An increasing number of speech controlled, wireless, and hands-free devices and applications are appearing on the market, such as home automation systems. As the market expands, it becomes more competitive and the demands on the performance is increasing. One property that increases mobility for the user is to be able to use the application in a larger perimeter, without performance being compromised. But the distance between speech source and microphone is a vulnerable path on which several types of disturbances can conjoin with the speech signal. Typical disturbances are noise and reverberations, and a large challenge is to distinguish speech from noise signals.

The thesis addresses the issue of decreasing performance when the user speaks to an application from different distances, in environments with different noise levels and reverberation. The focus of the thesis lies on evaluating whether beamforming can increase, or at least keep, the performance when the speaker is located a couple of meters away from the microphones. This problem could be solved by adding multiple microphones to remove disturbances.

Theory

Beamforming uses the time it takes a sound wave to propagate between microphones placed at different locations and, in this case a set of four mics, can be combined to emphasize a speech signal.

To be able to recognize a recorded spoken word the uttermost important characteristics of the spoken word must be extracted. These characteristics are then matched against a library of multiple words. A decision is made of what word, if any, was most likely spoken.

Tests

The main tests were performed in an offline manner in Matlab, using pre-recorded test words and noises. The purpose of these tests was to analyze different environmental combinations of noise, reverberations, speaker distance, and microphone set-ups. In total, 732.000 words were tested.

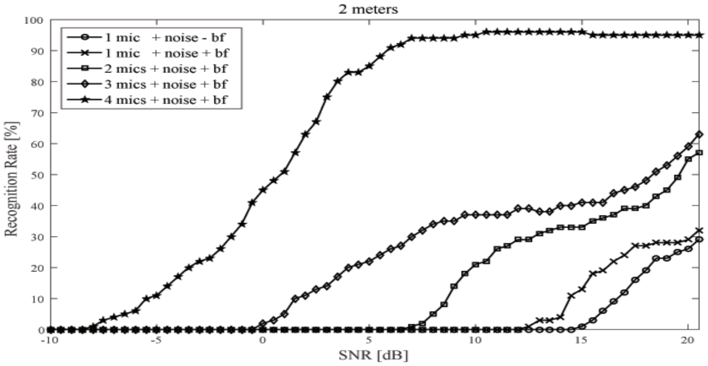

Figure 1: Results: speech + white noise at 2m distance, on y-axis the recognition rate [0-100 %] and on x-axis the Signal to Noise Ratio

Results and Conclusion

The results show that noise and reverberation severely damage the performance. Most results also show that beamforming, in most environments, is the best choice, and that the performance rate gets increasingly better the more microphone are utilized, which can be seen in Figure 1. Thus, using multiple microphones and beamforming is superior to one microphone and no beamforming. But there is definitely room for improvements, to further increase the performance rate. For example, to be able to introduce flexibility in user environments, one needs to take reverberations into account in the algorithms. Perhaps also introduce an adaptive beamforming algorithm which adapts the noise canceling depending on changes in the environment.