Abstract

This thesis is related to the field of acoustic signal processing and its applications to emerging communication environments. Acoustic signal processing is a very wide research area covering the design of signal processing algorithms involving one or several acoustic signals to perform a given task, such as locating the sound source that originated the acquired signals, improving their signal to noise ratio, separating signals of interest from a set of interfering sources or recognizing the type of source and the content of the message. Among the above tasks, *Sound Source localization* (SSL) and *Automatic Speech Recognition* (ASR) have been specially addressed in this thesis. In fact, the localization of sound sources in a room has received a lot of attention in the last decades. Most real-word microphone array applications require the localization of one or more active sound sources in adverse environments (low signal-to-noise ratio and high reverberation). Some of these applications are teleconferencing systems, video-gaming, autonomous robots, remote surveillance, hands-free speech acquisition, etc. Indeed, performing robust sound source localization under high noise and reverberation is a very challenging task. One of the most well-known algorithms for source localization in noisy and reverberant environments is the *Steered Response Power - Phase Transform* (SRP-PHAT) algorithm, which constitutes the baseline framework for the contributions proposed in this thesis. Another challenge in the design of SSL algorithms is to achieve real-time performance and high localization accuracy with a reasonable number of microphones and limited computational resources. Although the SRP-PHAT algorithm has been shown to be an effective localization algorithm for real-world environments, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue. In this context, several modifications and optimizations have been proposed to improve its performance and applicability. An effective strategy that extends the conventional SRP-PHAT functional is presented in this thesis. This approach performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid that reduces the computational cost required in a practical implementation with a small hardware cost (reduced number of microphones). This strategy allows to implement real-time applications based on location information, such as automatic camera steering or the detection of speech/non-speech fragments in advanced videoconferencing systems.

As stated before, besides the contributions related to SSL, this thesis is also related to the field of ASR. This technology allows a computer or electronic device to identify the words spoken by a person so that the message can be stored or processed in a useful way. ASR is used on a day-to-day basis in a number of applications and services such as natural human-machine interfaces, dictation systems, electronic translators and automatic information desks. However, there are still some challenges to be solved. A major problem in ASR is to recognize people speaking in a room by using distant microphones. In distant-speech recognition, the microphone does not only receive the direct path signal, but also delayed replicas as a result of multi-path propagation. Moreover, there are multiple situations in teleconferencing meetings when multiple speakers talk simultaneously. In this context, when multiple speaker signals are present, *Sound Source Separation* (SSS) methods can be successfully employed to improve ASR performance in multi-source scenarios. This is the motivation behind the training method for multiple talk situations proposed in this thesis. This training, which is based on a robust transformed model constructed from separated speech in diverse acoustic environments, makes use of a SSS method as a speech enhancement stage that suppresses the unwanted interferences. The combination of source separation and this specific training has been explored and evaluated under different acoustical conditions, leading to improvements of up to a 35% in ASR performance.

Keywords: Sound source localization, sound source separation, SRP-PHAT, microphone array, speaker detection, automatic speech recognition.