Ideal Binary Mask as the computational goal of auditory scene analysis

Author's Name

March 29, 2019

Abstract

What is the computational goal of auditory scene analysis? This is a key issue to address in the Marrian infomation-processing framework. It is also an important question for researchers in computatinal auditory scene analysis(CASA) because it bears directly on how a CASA system should be evaluated. In this chapter I discuss different objectives used in CASA. I suggest as a main CASA goal the use of the ideal time-frequency(T-F) binary mask whose value is one for a T-F unit where the target energy is greater than the interference energy and is zero otherwise. The notion of the ideal binary mask is motivated by the auditory masking phenomenon. Properties of the ideal binary mask are disucssed, include their relationship to automatic speech recognition and human speech intelligibility. This CASA goal has led to algorithms that directly estimate the ideal binary mask in monaural and binaural conditions, and these algorithms have substantially advanced the state-of-the-art performance in speech separation.

1 Introduction

In a natural environment, a target sound, such as speech, is usually mixed with acoustic interference. A sound separation system that removes or attenuates acoustic interference has many important applications, such as automatic speech recongition (ASR) and speaker identification in acoustic environments, audio information retrieval, sound-based human computer interaction, and intelligent hearing aids design.

Because of its importance, the sound separation problem has been extensively studied in signal processing and related fields. Three main approaches are speech enhancement (Lim, 1983; O'Shaughnessy, 2000), spatial filtering with a microphone array, and blind source separation using independent component analysis (ICA). Speech enhancement typically assumes certain prior knowledge of interference; for example, the standard spectral subtraction technique is easy to apply and works well when the background noise is stationary. However, the enhancement approach has difficulty in dealing with the unpredictable nature of general environments where a variety of intrusions, including nonstationary ones such as competing talkers, may occur. The objective of spatial filtering, or beamforming, is to estimate the signal that arrives from a specific direction through proper array configuration, hence filtering out interfering signals from other directions. With a large array spatial filtering can produce high-fidelity separation, and at the same time attenuate much signal reverberation. A main limitation of spatial filtering is what I call **configuration stationarity**: It has trouble tracking a target that removes around or switches between different sound sources. Closely related to spatial filtering is ICA-based blind source separation, which assumes statistical independence of sound sources and formulates the separation problem as that of estimating a demixing matrix. To make standard ICA formulation work requires a number of assumptions on the mixing process and the number of microphones. ICA gives impressive separation results when its assumptions are met. On the other hand, the assumptions also limit the scope of the applicability. For example, the stationarity assumption on the mixing process - similar to the configuration stationarity in spatial filtering - is hard to satisfy when speakers turn their heads or move around.

While machine separation remains a challenge, the auditory system shows a remarkable capicity for sound separation, even monaurally. According to Bregman, the auditory organizes the acoustic input into perceptual streams, corresponding to different sources, in a process called auditory scene analysis. Bregman further asserts that ASA take place in two stages in the auditory system: the first stage decomposes the acoustic mixture into a collection of sensory elements or segments, and the second stage selectively groups segments into streams. This two stage conception corresponds in essence to an analysis-synthesis strategy. Major ASA cues include proximity in frequency and time, harmonicity, smooth transition, onset synchrony, common location, common amplitude and frequency modulation, and prior knowledge.

Research in ASA has inspired a series computational studies to model auditory scene analysis. Mirroring the above two-stage conception, computational auditory scene analysis generally approaches sound seraration in two main stages: segmentation and grouping. In segmentation, the acoustic input is decomposed into sensory segments, each of which likely originates from a single souce, by analyzing harmonicity, onset, frequency transition, and amplitude modulation. In grouping, the segments that likely origniates from the same source are grouped, based mostly on periodicity analysis. In comparision with other separation approcahes, the main CASA success has been in monaural separation with minimal assumptions. It also creates a new set of challenges and demands, such as reliable multipitch tracking and special handling of unvoiced speech.

$$\alpha = \sqrt{\beta} \tag{1}$$

1.1 Subsection Heading Here

Write your subsection text here.

2 Conclusion

Write your conclusion here.