The Cocktail Party Problem: Solutions and Applications

Karl Wiklund[1] and Simon Haykin[1]
[1]Dept. of Electrical and Computer Engineering, McMaster University, Ontario, Canada, L8S 4K1

1. INTRODUCTION

If a genuine, real-time solution to the cocktail party problem[1] could be found, it would have many potential applications. These uses would include for example: interference suppression for hearing aids, improving environmental awareness for wearers of active noise control headsets, as well as improving the audio quality of telephone conversations. Surveying the literature however, one finds that while there has been much work done on the subject of cocktail party processing and computational auditory scene analysis[2], essentially none of it has dealt with the problems of real-time and embedded processing[3]. This is an important challenge given that the previously mentioned applications must run in real-time while using minimal computational resources.

2. Computational Auditory Scene Analysis (CASA)

As the name implies, Computational Auditory Scene Analysis, is the machine-based counterpart to Bregman's Auditory Scene Analysis[4]. While there are many different variants of the computational process, they all share the same essential feature: given some discrete time-frequency decomposition, assign an individual unit $s(t,f)$ a gain $m(t,f)$ such that the desired target signal is preserved, and the unwanted interference is suppressed.

The classification procedure is ultimately based on simple approximations of the four basic cues outlined by Bregman: the interaural time-difference (ITD), interaural intensity difference (IID), pitch and temporal onset. All four of these cues have simple signal processing analogues. The ITD and pitch, for example, can be computed via the cross-correlation and auto-correlation functions respectively, while IID and onset are based on comparisons involving the signal's power envelope. These basic computations are all well-established in the CASA literature.

3. Data Fusion in CASA Systems

Data fusion in CASA systems is not a trivial matter given the highly variable nature of the acoustic environment. This variability exists from moment-to-moment as a particular environment changes, as well as on a room-to-room basis, given that any CASA unit must be able to operate in a wide variety of different general environments. Attempts at producing statistically optimal fusion rules therefore have not been successful given the grave difficulties in actually forming the real-world probabilities for the range of scenarios that can exist. This is further complicated by the computational complexity of such estimation algorithms, which severely limits their practicality in real-time and embedded systems.

3.1 Hierarchical Cue Fusion

Consideration must therefore be given to the behaviour of these cues in realistic environments, and their respective robustness to noise and reverberation.

We based our approach to cue fusion on the following two principles:

1)The most acoustically robust cues are the most important in terms of grouping. Less robust cues should be used in a supplementary role in order to constrain the association of the primary cues.

2)The variability of the cue distributions means that the interpretation of the cues must be in terms of the mean and variance over several channels, and not in terms of individual time-frequency units.

In practice, these principles can be realized given the knowledge that both the speech onset period and periods of approximately constant pitch[5] are relatively robust to noise and reverberation when compared to both the ITD and IID. These last two cues, although less robust, supply the spatial information necessary for speech separation, and must therefore be aggregated in order to resolve a stream's identity.

3.2 Fuzzy Logic Data Fusion

Owing to the previously mentioned limitations of probabilistic methods, as well as the rule-based nature of Bregman's outline, the use of fuzzy logic for data fusion is a natural choice. Cue fusion therefore, can be described in terms of a series of *IF-THEN* rules that make use of somewhat vague definitions. For example, the combination of the ITD and IID can indicate the presence of a target speaker at some pre-defined spatial location. In other words, *IF* "most" of the ITDs *AND* "most" of the IIDs indicate the presence of a target, *THEN* a target is "likely" to be present. The time-frequency mask for CASA segregation can be readily formed using the truth-values of the applicable fuzzy rules, thus forming a "softmask" approach to segregation.

3. Control and Adaptation

The reliability of the auditory cues, and as a consequence, the reliability of the fusion mechanisms, depends on the acoustic environment in ways that are difficult to quantify.

On the whole however, it can be said that increasing levels of noise and reverberation reduce the quality of the filtered signal. Two combat these different sources of interference, two adaptation methods were incorporated into the FCPP.

3.1 Recursive Smoothing

In the FCPP, this scheme takes the form of the double-sided single-pole recursion[6] shown in equation (2):

$$\hat{\rho}(t,j) = \beta(t) \cdot \rho(t,j) + (1-\beta(t)) \cdot \hat{\rho}(t,j-1) \qquad (2)$$

where $\rho(t,j)$ is the truth value of the fuzzy rule relevant to the tth time-step, and jth frequency bin. The time-varying smoothing parameter is $\beta(t)$, and the smoothed gain estimate is $\hat{\rho}(t,j)$.

The smoothing parameter depends on the presence or absence of an onset period. As these segments are relatively free of reverberation, their associated spatial cues are more reliable, and the relevant frames should therefore be given greater emphasis in terms of the level of certainty (gain) attached to them. This is accomplished by letting the smoothing value take on two different values as shown below

$$\beta(t) = \begin{cases} 0.3 & \text{if } onset = TRUE \\ 0.1 & \text{if } onset = FALSE \end{cases} \qquad (3)$$

3.2 The Gain Floor

The second aspect of the control problem performs the task of controlling the level of musical noise by adapting to changing levels of background noise. This problem was addressed not by smoothing, but by selectively adding in the unprocessed background noise. Specifically, the final gain calculation for the controller is expressed as

$$g(t,j) = \hat{\rho}(t,j) + \neg\hat{\rho}(t,j) \cdot FLOOR \qquad (4)$$

where $g(t,j)$ is the gain for the jth frequency bin at time t, $\hat{\rho}(t,j)$ is the smoothed gain estimate from (2), $\neg\hat{\rho}(t,j)$ is its complement, and the value of $FLOOR$ is dependent on the currently estimated SNR.

4. Post-processing Using Spectral Subtraction

The cue fusion and estimation routines that have been described so far are unfortunately ambiguous with respect to noise sources located behind the listener. That is, the directional cues are unable to distinguish between sources that are in front of, or behind, the listener. In order to overcome this problem, the use of an additional pair of directional microphones is proposed. A very simple variation of the standard spectral subtraction algorithm can then be used to distinguish between frames dominated by the target, situated in front of the listener, and frames dominated by interfering sources located somewhere behind the listener.

5. Results

In the experiment shown below, a target signal in a reverberant room was positioned straight ahead of a KEMAR dummy is embedded in a scenario with three competing talkers positioned at azimuths of 67°, 180°, and 270°, with the respective time positions of the signals being randomized.

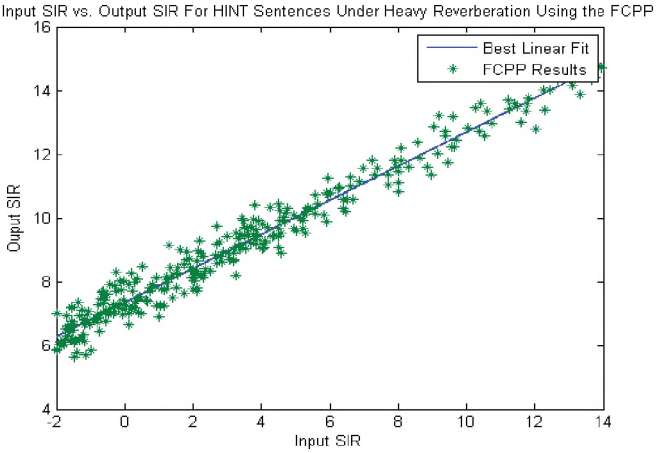

Figure 5. Output segmental SIR for a given input SIR in a reverberant room.

REFERENCES

[1] Rong, Dong. Perceptual Binaural Speech Enhancement in Noisy Environments. M.A.Sc thesis, McMaster University, 2004.

[2] Wang, Deilang and Brown, Guy J. "Fundamentals of Computational Auditory Scene Analysis" Ed. Deliang Wang and Guy J. Brown. Computational Auditory Scene Anlaysis. Piscataway, NJ: IEEE Press, 2006. 1-44.

[3] Wang, Deliang. Computational Auditory Scene Analysis and its Application to Hearing Aids. IHCON 2008. August 13-17, 2008. Lake Tahoe, CA.

[4] Bregman, Albert. Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press, 1990.

[5] Brown, Guy J. and Palomaki, Kalle J. "Reverberation". Ed. Deliang Wang and Guy J. Brown. Computational Auditory Scene Analysis. Piscataway, NJ: IEEE Press, 2006. 209-250.

[6] Diethorn, Eric J. "Subband Noise Reduction Methods for Speech Enhancement". Ed. Yiteng Han and Jacob Benesty. Audio Signal Processing For Next Generation Multimedia Communication Systems. Norwell, MA: Kluwer, 2004. 91-118.