In P. Divenyi (Ed.), "*Speech Separation by Humans and Machines*," Chapter 12, pp. 181-197, Kluwer Academic, Norwell MA, 2005

Chapter 12

On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*

DeLiang Wang

Department of Computer Science and Engineering and Center for~~of~~ Cognitive Science
The Ohio State University, Columbus, OH
dwang@cis.ohio-state.edu

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 recognition (ASR) and speaker identification in real 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 (van Veen and Buckley, 1988; Krim and Viberg, 1996), and blind source separation using independent component analysis (ICA) (Lee, 1998; Hyvärinen et al., 2001). 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 moves 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 (van der Kouwe et al., 2001). 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 capacity for sound separation, even monaurally (i.e. with one microphone). According to Bregman (1990), the auditory system organizes the acoustic input into perceptual streams, corresponding to different sources, in a process called auditory scene analysis (ASA). Bregman further asserts that ASA takes 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 of computational studies to model auditory scene analysis (Weintraub, 1985; Cooke, 1993; Brown and Cooke, 1994; Ellis, 1996; Wang and Brown, 1999). Mirroring the above two-stage conception, computational auditory scene analysis (CASA) generally approaches sound separation 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 source, by analyzing harmonicity, onset, frequency transition, and amplitude modulation. In grouping, the segments that likely originate from the same source are grouped, based mostly on periodicity analysis. In comparison with other separation approaches, the main CASA success has been in monaural separation with minimal assumptions.[1] It also creates a new set of challenges and demands, such as reliable multipitch tracking and special handling of unvoiced speech.

In comparison with other well-established separation approaches, CASA faces a somewhat distinct issue: there is no consensus on how to quantitatively evaluate a CASA system (Rosenthal and Okuno, 1998). Almost every study adopts its own evaluation criteria. This is partly due to the fact that CASA is still in its infancy, but it may reflect deeper confusion on the computational goal of auditory scene analysis. The lack of common

[1] More accurately, CASA also makes a number of assumptions, but such assumptions tend to conform to the constraints under which the auditory system operates.

evaluation criteria makes it difficult to document and communicate the progress made in the field. Sensible evaluation criteria can also serve as the guiding principle for model development.

This chapter intends to examine the goal of CASA. After analyzing the advantages and disadvantages of different computational objectives, I suggest ideal time-frequency (T-F) mask as the computational goal of auditory scene analysis. The remainder of the chapter is organized as follows. The next section reviews different CASA evaluation criteria. Section 3 is devoted to a general discussion of the CASA goal, including an analysis of several alternative CASA objectives. Section 4 introduces the ideal binary mask, analyzes its~~their~~ properties, and argues for its~~their~~ use as the CASA goal. Section 5 describes two models that explicitly estimate the ideal binary mask. Finally, Section 6 concludes the chapter.

2 CASA EVALUATION CRITERIA

CASA criteria that have been suggested can be divided into the following four categories: Direct comparisons between segregated target and premixing target, changes in automatic speech recognition (ASR) score, evaluation with human listening, and fit with biological data. Each is described below.

- **Comparison with premixing target**. Obviously this assumes the availability of premixing sound sources, which is not an unreasonable assumption for system evaluation. The evaluation criterion employed in Cooke's study (1993) is the match between a model-generated group of target elements and the group of elements in clean target speech. Brown and Cooke (1994) use a segregated target stream, which is a binary T-F mask, to resynthesize target speech and noise intrusion, and then calculate a normalized ratio between resynthesized speech and resynthesized noise. Subsequently, Wang and Brown (1999) use conventional signal-to-noise ratio (SNR), measured in decibels, between resynthesized speech and resynthesized noise. More tailored for speech, Nakatani and Okuno (1999) calculate~~s~~ spectral distortion by comparing the short-term spectra of segregated speech and those of clean speech. Bodden (1993) in his binaural model of speech segregation estimates a time-varying Wiener filter for each sound mixture, which consists of energy ratios between the target speech and the mixture within critical bands.
- **ASR measure**. A main motivation behind research on speech separation is to improve ASR performance in the presence of acoustic interference. So it is natural to evaluate a CASA model in

terms of ASR score. This measure is used in the Weintraub model (1985) - probably the earliest CASA study. The evaluation metric is straightforward: Measuring changes in the recognition score using a standard ASR system before and after sound mixtures are segregated by the CASA model in question. Early ASR evaluations produce ambiguous results, partly because processing stages in CASA tend to distort the target signal, creating a mismatch between segregated signal and clean signal used ASR training. More recent attempts have yielded better outcomes (see, for example, Glotin, 2001).

- **Human listening**. Human listeners can be involved in evaluating a computational model in terms of speech intelligibility on original mixtures and on segregated speech (Stubbs and Summerfield, 1988; 1990). An improvement in speech intelligibility would lend support to the value of the model. However, human listeners are very good at segregating a sound mixture, and this creates a potential confound for using listeners to test model output. One practical difficulty is that in order to give room for a model to improve on intelligibility, interference must be very strong, which could be exceedingly hard for models to perform. Listeners with hearing loss may be better suited for such evaluation as it is well-known that people with sensorineural impairment have greater difficulty in segregating target speech in a noisy environment (Moore, 1998). Of course, if the objective of the model is to improve hearing of abnormal listeners or that of normal listeners in highly noisy environments, this evaluation methodology is the best choice. Ellis (1996) made a different use of human listening in evaluation: His listeners were used to score the resemblance of segregated sounds to component sounds in the mixture.

- **Fit with biological data**. Some CASA researchers are interested in modeling the human ASA process, while some others are interested in elucidating neurobiological mechanisms underlying ASA. For such models, the main evaluation criterion is how well the models account for known perceptual or neurobiological data. Wang (1996) sought to model a number of ASA phenomena on the basis of a neural oscillator network (see also Norris, 2003). McCabe and Denham (1997) proposed a different neural network that simulates psychophysical results on auditory streaming. Recently, Wrigley and Brown (2004) put forward a neural oscillator model of auditory attention and used it to quantitatively simulate a set of psychological data.

In P. Divenyi (Ed.), "*Speech Separation by Humans and Machines*," pp. 181-197, Kluwer Academic, Norwell MA, 2005

3 WHAT IS THE GOAL OF CASA?

Different evaluation criteria tend to reflect different goals of computational models, whether or not they are explicitly laid out. This raises the question of what should be the goal of CASA? This is a very important question since its answers bear directly on the research agenda and determine whether computational efforts lead to real progress towards ultimately solving the CASA problem.

To address this question, it might be helpful to put CASA in a broad context of perception since CASA purports to model auditory scene analysis, which is a major process of auditory perception. So a larger question is, what is the goal of perception? This question, raised in the most general form, would fall under the realm of philosophy, and indeed philosophers have debated this issue for centuries. What we are concerned here is the information processing perspective, which is shared by human and machine perception. From this perspective, Gibson (1966) considers perceptual systems as ways of seeking and extracting information about the environment from the sensory input. In the visual domain, Marr (1982) states that the purpose of vision is to produce a visual description of the environment for the viewer. By extrapolating Marr's statement to the auditory domain, the purpose of audition would be to produce an auditory description of the environment for the listener. It is worth noting, according to the above views, that perception is a process private to the perceiver despite the fact that the physical environment is common to different perceivers.

According to Bregman (1990), the goal of ASA is to produce separate auditory streams from sound mixtures, each stream corresponding to an acoustic event. This would imply that the goal of CASA is to computationally extract individual streams from sound mixtures. To make this description more meaningful, however, further constraints need to be observed:

- To qualify as a stream a sound must be audible on its own. In other words, the intensity of the sound at the eardrum must exceed a certain sound level, referred to as the absolute threshold (Moore, 2003).
- The number of streams that can be segregated at a time must be limited. This limit is directly related to the capacity of auditory attention. In a comprehensive account, Cowan recently concluded that the capacity of attention is about four (Cowan, 2001). This implies that the auditory system cannot segregate more than 4 streams simultaneously. While a listener may be able to segregate up to 4 tones or steady vowels, in a very noisy environment such as a cocktail party, the attentional capacity may reduce to figure-ground

separation, i.e. attending to only a foreground stream with a general awareness of the background.

- A fundamental fact in auditory perception is auditory masking (Moore, 2003). Roughly speaking, auditory masking refers to the phenomenon that within a critical band a stronger signal tends to mask a weaker one. When a sound is masked, it is eliminated from perception as if the sound never reached the ear.
- ASA results depend on sound types. Say we listen to mixtures of two equally-loud sounds. If the sounds are two tones well separated in frequency or two speech utterances, we can readily segregate them. On the other hand, if the sounds are white noise and pink noise we are completely incapable of any segregation.

With the above analysis in mind, we now discuss some alternative CASA objectives. The first objective, which might be called the gold standard, is simply to segregate all sound sources from a sound mixture. If this standard could be reached, it would be the ideal goal of CASA, at least from an engineering standpoint. On the other hand, the goal is clearly beyond what the human listener can do; just observe for yourself how many conversations you can follow in a cocktail party. It is probably also an unrealistic computational goal if the system has just one or two microphones.

Another alternative objective is to enhance ASR. This objective has the advantage that it directly relates to one of the primary motivations for CASA research. The objective is also straightforward to evaluate as discussed in the last section. This objective has several drawbacks. One drawback is that it is narrowly focused on speech. Although speech is a vital type of acoustic signal for humans, it is by no means the only important signal to us. What about music, or other environmental sounds? For music in particular, it is hard to characterize music perception as a recognition process. A deeper issue with recognition as the goal is that perceiving is more than recognizing (Treisman, 1999). Perceiving has, in addition of recognition, all the current details of events, such as how they sound like, where they are, whether they are approaching or receding, and many other details about them. Such details are crucial for the perceiver to decide how to act. Also it is not clear how the ASR objective can account for the fact that new things unheard before can be perceived as well.

The third alternative is to enhance human listening. A main advantage of this objective is the close coupling with auditory perception. Also a primary motivation of studying CASA is to improve hearing prosthesis for listeners with hearing impairment as well as hearing of normal listeners in very noisy environments. However, this objective is specifically tailored to human listening and there are other applications that do not directly involve humans,

such as audio information retrieval. There are also practical difficulties for computational researchers in terms of required expertise for conducting human experiments.

These alternative objectives have their advantages and disadvantages. A desirable objective should be generally consistent with the above analysis on human auditory scene analysis, and be comprehensive enough to apply to different types of acoustic signal and different application domains. The objective should not consider just recognition performance or human listening, but at the same time it should be consistent with such criteria. The simplicity of the objective and easiness to apply are also desirable so that a researcher need not wait for a long time to find out how well a provisional model works. In the next section, I present the ideal time-frequency mask as a putative goal of CASA.

4 IDEAL BINARY MASK AS THE GOAL OF CASA

As discussed in Section 3, the gold-standard objective is probably unrealistic. A more realistic objective is to segregate a target signal from the mixture. Then the objective becomes that of figure-ground separation. This begs the question of what should be regarded as the target? Generally speaking, what the target is depends on external input as well as intention; it is closely related to the study of attention, in particular what attracts attention (Pashler, 1998). From a practical standpoint, what constitutes the target is task-dependent and often unambiguous. For the purpose of our discussion, we assume that the target is known. We also assume, for the sake of evaluation, the availability of premixing target signal and interference.

A widely accepted representation in CASA is the two-dimensional time-frequency representation where the time dimension consists of a sequence of time frames and the frequency dimension consists of a bank of auditory filters (e.g. gammatone filters). This representation is consistent with accounts of human ASA and auditory physiology. Within this representation, the key consideration behind the notion of the ideal binary mask is to retain the time-frequency regions of a target sound that are stronger than the interference, and discard the regions that are weaker than the interference. More specifically, an ideal mask is a binary matrix, where 1 indicates that the target energy is stronger than the interference energy within the corresponding T-F unit and 0 indicates otherwise. This definition implies a 0-dB SNR criterion for mask generation, and other SNR criteria are possible too (see below). Figure 12.1 illustrates the ideal mask for a mixture of a male utterance and a female utterance, where the male utterance is regarded as target. The overall SNR of the mixture is 0 dB. The top left panel

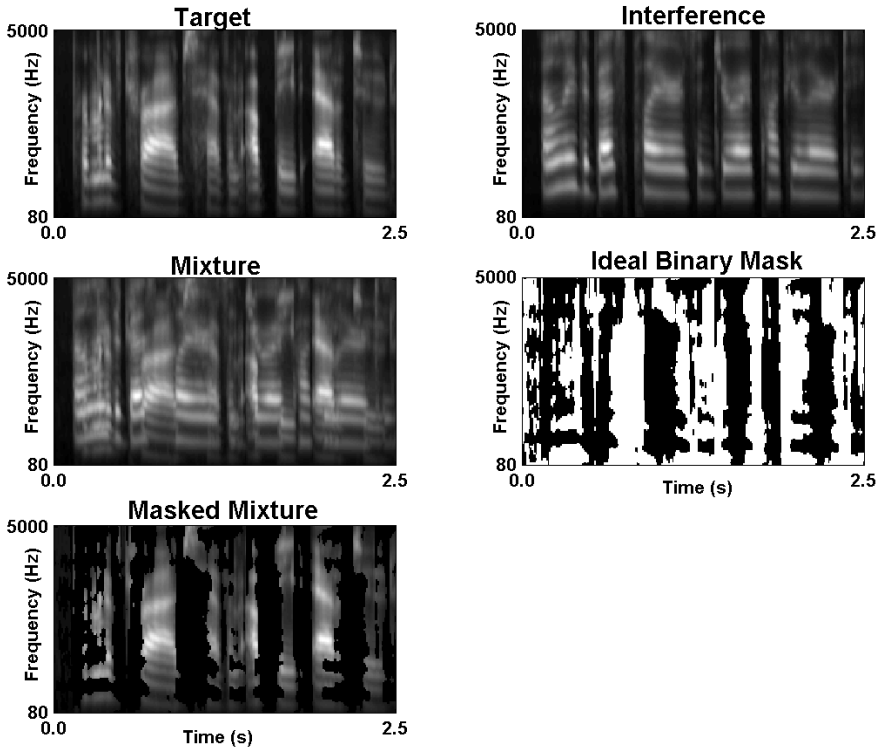

Figure 12.1. Illustration of the ideal binary mask. **Top left**: Two-dimensional T-F representation of the target utterance ("Primitive tribes have an upbeat attitude"). The figure displays the rectified responses of the gammatone filterbank with 128 channels. **Top right**: Corresponding representation of the interfering utterance ("Only the best players enjoy popularity"). **Middle left**: Corresponding representation of the mixture. **Middle right**: Ideal T-F binary mask, where white pixels indicate 1 and black pixels 0. **Bottom left**: Masked mixture using the ideal binary mask.

of Figure 12.1 shows the T-F representation of the target utterance, the top right panel the representation of the interfering utterance, and the middle left panel the representation of the mixture. For this mixture, the ideal mask is shown in the middle right panel. The bottom left panel of the figure shows the result of ideal masking on the mixture. Compared with the original mixture, the masked mixture is much closer to the clean target. Listening to the masked mixture one can clearly hear the target utterance while no trace of interference is audible.

Binary masks have been used as an output representation in the CASA literature (Brown and Cooke, 1994; Wang and Brown, 1999). Related to binary masks is the observation that different speech utterances tend to be orthogonal in a high-resolution time-frequency representation because the

energy of a single utterance tends to be sparsely distributed (Jourjine et al., 2000; Roweis, 2001). This observation obviously does not hold when an acoustic background is babble noise or contains broadband intrusions. To my knowledge, the papers of Hu and Wang (2001) and Roman et al. (2001) are the earliest studies that suggest the use of the ideal binary mask (see also Roman et al., 2003; Hu and Wang, 2004). Note that the definition of the ideal binary mask does not assume orthogonality among sound sources.

The ideal binary mask has a number of desirable properties:

- Flexibility. With the same mixture, the definition leads to different masks depending on what the target is. It is consistent with the perceptual observation that the same environment can be perceived in different ways by different perceivers.
- Well-definedness. The ideal mask is well defined no matter how many intrusions are in the scene. One may also identify multiple targets from the same mixture, with multiple processors that have different target definitions.
- The ideal binary mask sets the ceiling performance for all binary masks.
- The ideal mask is broadly consistent with ASA constraints in terms of audibility and segregation capacity. In particular, it has direct correspondence with the auditory masking phenomenon.

When a gammatone filterbank is used for generating the time-frequency representation, a technique introduced by Weintraub (1985) can be used to resynthesize a waveform signal from a binary mask (see also Brown and Cooke, 1994; Wang and Brown, 1999). One can then conduct listening tests on resynthesized signal. The ideal binary mask produces high quality resynthesized target unless the mixture SNR is very low.

Recent research on missing-data speech recognition provides an effective bridge between a segregated mask and ASR (Cooke et al., 2001). The main idea of missing-data recognition is to adapt the standard HMM recognizer so that recognition decisions are based only on reliable T-F units while marginalizing unreliable or missing T-F units. Cooke et al. (2001) found that the *a priori* mask - defined according to whether the mixture energy is within 3 dB of the target energy - used in conjunction missing-data recognition yields excellent recognition performance. Similar performance is obtained by Roman et al. (2003) using the ideal binary mask. Moreover, the study of Roman et al. (2003) found that deviations from the ideal binary mask lead to gradual degradation in speech recognition performance.

The ideal binary mask has been recently tested in human speech intelligibility experiments. As noted earlier, the definition of the ideal mask uses the 0-dB SNR criterion within individual T-F units. However, one can produce different ideal masks using different local SNR criteria. Brungart et al. (in preparation) tested a range of local SNR criteria around 0 dB using ideal masking on speech mixtures involving one target talker and 1 to 3 competing talkers. All talkers have equal overall loudness, or the SNR between the target and a single competing talker is zero. Their experiments showed that, within the local SNR range from -5 dB to 5 dB, ideal masking produces intelligibility scores near 100% in all mixtures involving 2, 3, and 4 talkers. In addition, the intelligibility score decreases systematically towards higher or lower SNR criteria. Note that for a fixed mixture a very high SNR criterion leads to a mask with very few 1's, hence very little target energy; a very low SNR criterion leads to a mask close to an all-1 mask, hence very little segregation. Their results also show that, for mixtures with very low SNR, ideal masking improves speech intelligibility dramatically (see also Roman et al., 2003~~1~~).

Finally an analogy may be drawn between auditory binary masking and visual occlusion. Figure 12.2 illustrates occlusion with a natural image of water lilies, where a lily in the front occludes the objects in the back. Visual occlusion may be considered as an instance of binary masking, in which the pixels of a front surface are assigned 1 in the mask and those of the occluded surfaces are assigned 0. Moreover, when an observer attends to a particular object in an image (say the lily near the center of Figure 12.2), this process of attending is analogous to ideal binary masking where the pixels of the attended object correspond to 1's in the mask and the remaining pixels correspond to 0's.

Figure 12.2. A natural image of water lilies.

In P. Divenyi (Ed.), "*Speech Separation by Humans and Machines*," pp. 181-197, Kluwer Academic, Norwell MA, 2005

5 ESTIMATION OF THE IDEAL BINARY MASK

The ideal binary mask clearly quantifies the computational goal of CASA. Guided by this goal, we have made conscious effort to compute the ideal mask. This section describes two models that explicitly estimate the ideal binary mask.

5.1 Monaural segregation of voiced speech

Voiced speech segregation has been a primary topic in CASA. For voiced speech, harmonicity is the essential cue for segregation. Earlier CASA models can segregate much of the low-frequency energy, but have trouble segregating high-frequency components. It is well-known that the auditory system can resolve the first few harmonics, while higher harmonics are unresolved. Psychoacoustic research suggests that the auditory system may use different mechanisms to deal with resolved and unresolved harmonics (Carlyon and Shackleton, 1994; Bird and Darwin, 1997). Subsequently, Hu and Wang (2003; 2004) developed a CASA model that employs different mechanisms in the low- and the high-frequency range. The model follows the general two-stage processing (see Section 1): Segmentation and grouping. Building on the output from the Wang-Brown model (1999) that works well in the low-frequency range, Hu and Wang proposed a psychoacoustically motivated method for tracking target pitch contours.

With the results of target pitch tracking, the model then labels individual T-F units. In the low-frequency range, a T-F unit is labeled by comparing its response periodicity and the extracted target pitch period. In the high-frequency range, wide bandwidths of auditory filters cause the filters to respond to multiple unresolved harmonics of voiced speech. These responses are amplitude modulated due to beats and combinational tones (Helmholtz, 1863). Furthermore, response envelopes fluctuate at the frequency that corresponds to the fundamental frequency of speech. Hence, the model labels a high-frequency unit by comparing its amplitude modulation (AM) rate with the extracted pitch frequency. To derive AM rates Hu and Wang have employed a sinusoidal modeling technique; specifically, a single sinusoid is used to model AM within a certain range of target pitch and the derivation of AM rates can then be formulated as an optimization problem. With appropriately chosen initial values, the optimization problem can be solved efficiently using an iterative gradient descent technique. With labeled T-F units, the model generates segments in the low-frequency range based on temporal continuity and cross-channel correlation between responses of adjacent frequency channels, and in the high-frequency range based on temporal continuity and common AM among adjacent filter responses. Segments thus formed then expand iteratively, and the resulting collection of

the segments with the target label gives the segregated target which is represented by a binary T-F mask.

Figure 12.3 illustrates the result of ideal mask estimation for voiced speech segregation. The top left panel of the figure shows the T-F representation of a voiced utterance which is the target. The top right panel shows the mixture of the utterance with a 'cocktail party' noise from Cooke (1993). The middle left panel shows the ideal binary mask for the mixture, and the middle right panel the estimated mask. The estimated mask is reasonably close to the ideal one. The bottom left panel gives the result of ideal masking on the mixture, and the bottom right panel the result of masking using the estimated mask.

The model of Hu and Wang (2004) produces substantially better performance than previous models, especially in the high-frequency range. In terms of systematic SNR evaluation, one may treat the resynthesized signal from the ideal binary mask as signal because the ideal mask represents

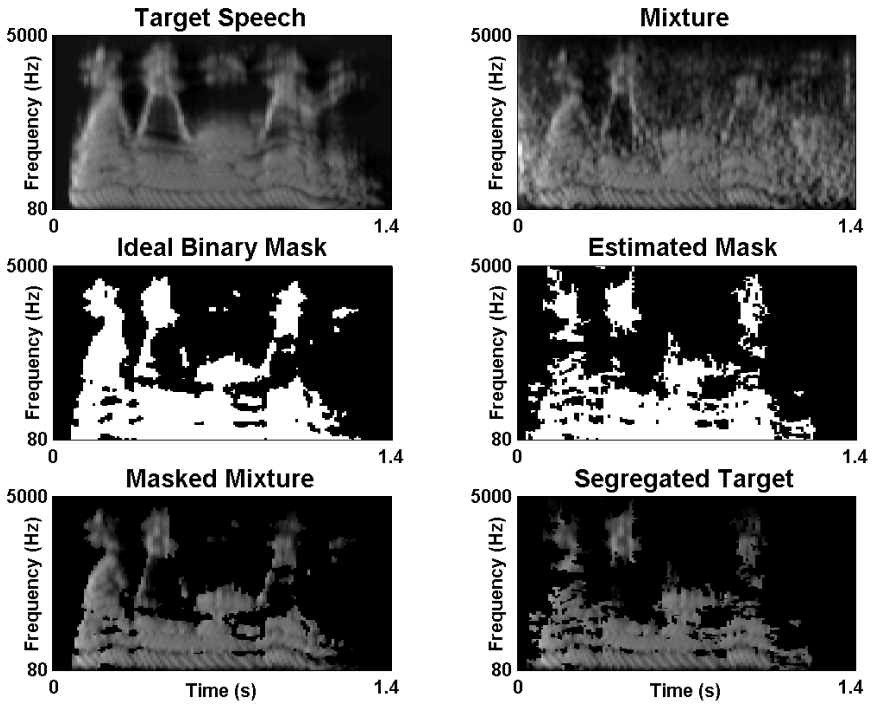

Figure 12.3. Ideal mask estimation for monaural speech segregation. **Top left**: T-F representation of the target utterance ("Why were you all weary"). **Top right**: T-F representation of the mixture of the target and the cocktail party noise. **Middle left**: Ideal binary mask for the mixture. **Middle right**: Estimated binary mask for the mixture. **Bottom left**: Masked mixture using the ideal mask. **Bottom right**: Masked mixture using the estimated mask.

the computational goal. The model then yields 5.2 dB improvement over the Wang and Brown model (1999), which had the representative performance of earlier CASA systems. It also has 6.4 dB gain over the standard spectral subtraction method in speech enhancement. Similar improvements are obtained with conventional SNR metric using premixing speech as signal.

5.2 Binaural speech segregation

It is well known that people can selectively attend to a single voice at a noisy cocktail party. Spatial location is believed to play an important role in cocktail party processing. How to simulate this perceptual ability, known as the cocktail-party problem (Cherry, 1953), is a great computational challenge.

Guided by the notion of the ideal binary mask, Roman et al. (2003) developed a new location-based approach to speech segregation. Their model uses the binaural cues of interaural time difference (ITD) and interaural intensity difference (IID) extracted from a KEMAR dummy head that realistically simulates the filtering process of the head, torso and external ear. They observe that, within a narrow frequency band, modifications to the relative energy of the target source to the interfering energy trigger systematic changes in the values of the binaural cues. For a given spatial configuration, this interaction produces characteristic clustering in the binaural feature space. Consequently, the model performs independent supervised learning for different spatial configurations and different frequency bands in the joint ITD-IID feature space. More specifically, they formulate the estimation of the ideal binary mask as a binary Bayesian classification problem, where the hypothesis is whether the target is stronger than the overall interference within a single T-F unit. Then a nonparametric method (kernel density estimation) is used to estimate likelihood functions in the ITD-IID space, which are then used in maximum *a posteriori* (MAP) decision making.

Figure 12.4 illustrates the result of estimating the ideal binary mask for natural speech segregation, using the same mixture shown in Figure 12.1. The top right panel shows the ideal binary mask, and the bottom right panel the estimated mask. The match between the two masks is excellent. Finally, the bottom left panel displays the result of masking the mixture using the estimated mask (cf. bottom left panel of Figure 12.1).

The resulting model was systematically evaluated in two-source and three-source configurations, and estimated binary masks approximate the ideal ones extremely well. In terms of conventional SNR evaluation, the model produces large and consistent SNR improvements over original mixtures. The SNR gains are as large as 13.8 dB in the two-source case and

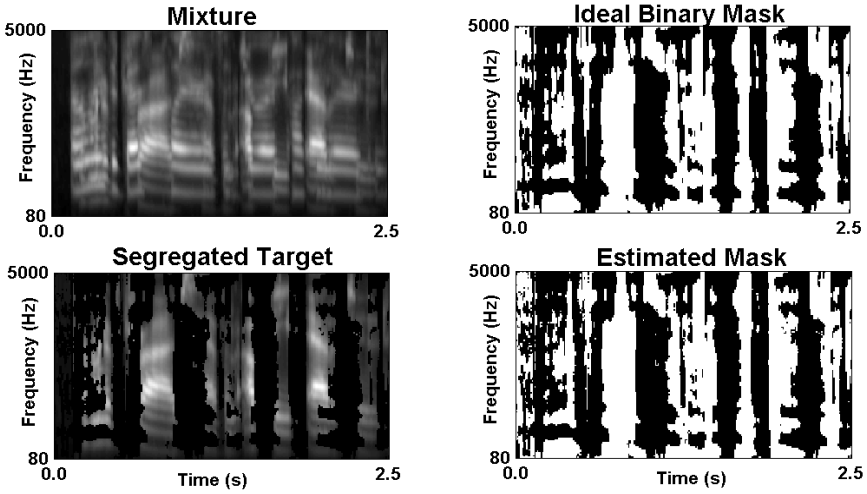

Figure 12.4. Ideal mask estimation for binaural speech segregation. **Top left**: the same mixture shown in Figure 1. **Top right**: Ideal binary mask for the mixture (also shown in Figure 1). **Bottom right**: Estimated binary mask. **Bottom left**: Masked mixture using the estimated mask.

11.3 dB in the three-source case. A comparison with the Bodden model (1993), which estimates a Wiener filter, shows that the Roman et al. model produces 3.5 dB improvement in the most favorable conditions for the Bodden model, and in other conditions the improvement is significantly greater. In addition to SNR evaluation, they performed an ASR evaluation by feeding estimated binary masks to a missing-data recognizer (Cooke et al., 2001), and the model yields large ASR improvements compared to direct recognition of mixtures. Also, the model was evaluated on speech intelligibility with human listeners. Because people excel at ASA and achieve near perfect intelligibility unless interference is severe, the tests used three low SNR levels: 0 dB, -5 dB and -10 dB (measured at the better ear). The general finding is that the algorithm improves human intelligibility for the tested conditions, and the improvement becomes larger as the SNR decreases - as large as an increase from an intelligibility score of 20% to 80% at -10 dB.

6 CONCLUSION

In his famous treatise of computational vision, Marr (1982) makes a compelling argument for separating different levels of analysis in order to understand complex information processing. In particular, the computational theory level, concerned with the goal of computation and general processing strategy, must be separated from the algorithm level, or the separation of

what from *how*. This chapter is an attempt at a computational-theory analysis of auditory scene analysis, where the main task is to understand the character of the CASA problem.

My analysis results in the proposal of the ideal binary mask as a main goal of CASA. This goal is consistent with characteristics of human auditory scene analysis. The goal is also consistent with more specific objectives such as enhancing ASR and speech intelligibility. The resulting evaluation metric has the properties of simplicity and generality, and is easy to apply when the premixing target is available. The goal of the ideal binary mask has led to effective ~~for~~ speech separation algorithms that attempt to explicitly estimate such masks.

7 ACKNOWLEDGMENTS

The author thanks G. Hu and N. Roman for their assistance in figure preparation. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (FA9550-04-1-0117).

References

Bird, J., and Darwin, C. J., 1997, Effects of a difference in fundamental frequency in separating two sentences, in: *Psychophysical and Physiological Advances in Hearing*, A. R. Palmer, *et al*, ed., Whurr, London,

Bodden, M., 1993, Modeling human sound-source localization and the cocktail-party-effect, *Acta Acust.* 1: 43-55.

Bregman, A. S., 1990, *Auditory Scene Analysis*, MIT Press, Cambridge MA.

Brown, G. J., and Cooke, M., 1994, Computational auditory scene analysis, *Computer Speech and Language* 8: 297-336.

Brungart, D., Chang, P., Simpson, B., and Wang, D. L., in preparation.

Carlyon, R. P., and Shackleton, T. M., 1994, Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *J. Acoust. Soc. Am.* 95: 3541-3554.

Cherry, E. C., 1953, Some experiments on the recognition of speech, with one and with two ears, *J. Acoust. Soc. Am.* 25: 975-979.

Cooke, M., 1993, *Modelling Auditory Processing and Organization*, Cambridge University Press, Cambridge U.K.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A., 2001, Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Comm.* 34: 267-285.

Cowan, N., 2001, The magic number 4 in short-term memory: a reconsideration of mental storage capacity, *Behav. Brain Sci.* 24: 87-185.

Ellis, D. P. W., 1996, *Prediction-driven computational auditory scene analysis*, Ph.D. Dissertation, MIT Department of Electrical Engineersing and Computer Science.

Gibson, J. J., 1966, *The Senses Considered as Perceptual Systems*, Greenwood Press, Westport CT.

Glotin, H., 2001, *Elaboration et étude comparative de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole: incorporation d'indices de voisement et de localisation*, Ph.D. Dissertation, Institut National Polytechnique de Grenoble.

Helmholtz, H., 1863, *On the Sensation of Tone* (A. J. Ellis, Trans.), Dover Publishers, Second English ed., New York.

Hu, G., and Wang, D. L., 2001, Speech segregation based on pitch tracking and amplitude modulation, in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79-82.

Hu, G., and Wang, D. L., 2003, Monaural speech separation, in: *Advances in Neural Information Processing Systems* (NIPS'02), MIT Press, Cambridge MA, pp. 1221-1228.

Hu, G., and Wang, D. L., 2004, Monaural speech segregation based on pitch tracking and amplitude modulation, *IEEE Trans. Neural Net.*, in press.

Hyvärinen, A., Karhunen, J., and Oja, E., 2001, *Independent Component Analysis*, Wiley, New York.

Jourjine, A., Rickard, S., and Yilmaz, O., 2000, Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures, in *Proceedings of IEEE ICASSP*, pp. 2985-2988.

Krim, H., and Viberg, M., 1996, Two decades of array signal processing research: The parametric approach, *IEEE Sig. Proc. Mag.* 13: 67-94.

Lee, T.-W., 1998, Independent Component Analysis: Theory and Applications, *Kluwer Academic*, Boston.

Lim, J., ed., 1983, *Speech Enhancement*, Prentice Hall, Englewood Cliffs NJ.

Marr, D., 1982, *Vision*, Freeman, New York.

McCabe, S. L., and Denham, M. J., 1997, A model of auditory streaming, *J. Acoust. Soc. Am.* 101: 1611-1621.

Moore, B. C. J., 1998, *Cochlear Hearing Loss*, Whurr Publishers, London.

Moore, B. C. J., 2003, *An Introduction to the Psychology of Hearing*, Academic Press, 5th ed., San Diego, CA.

Nakatani, T., and Okuno, H. G., 1999, Harmonic sound stream segregation using localization and its application to speech stream segregation, *Speech Comm.* 27: 209-222.

Norris, M., 2003, *Assessment and extension of Wang's oscillatory model of auditory stream segregation*, Ph.D. Dissertation, University of Queensland School of Information Technology and Electrical Engineering.

O'Shaughnessy, D., 2000, *Speech Communications: Human and Machine*, IEEE Press, 2nd ed., Piscataway NJ.

Pashler, H. E., 1998, *The Psychology of Attention*, MIT Press, Cambridge MA.

Roman, N., Wang, D. L., and Brown, G. J., 2001, Speech segregation based on sound localization, in *Proceedings of IJCNN*, pp. 2861-2866.

In P. Divenyi (Ed.), "*Speech Separation by Humans and Machines*," pp. 181-197, Kluwer Academic, Norwell MA, 2005

Roman, N., Wang, D. L., and Brown, G. J., 2003, Speech segregation based on sound localization, *J. Acoust. Soc. Am.* 114: 2236-2252.

Rosenthal, D. F., and Okuno, H. G., ed., 1998, *Computational Auditory Scene Analysis*, Lawrence Erlbaum, Mahwah NJ.

Roweis, S. T., 2001, One microphone source separation, in: *Advances in Neural Information Processing Systems* (NIPS'00), MIT Press,

Stubbs, R. J., and Summerfield, Q., 1988, Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners, *J. Acoust. Soc. Am.* 84: 1236-1249.

Stubbs, R. J., and Summerfield, Q., 1990, Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners, *J. Acoust. Soc. Am.* 87: 359-372.

Treisman, A., 1999, Solutions to the binding problem: progress through controversy and convergence, *Neuron* 24: 105-110.

van der Kouwe, A. J. W., Wang, D. L., and Brown, G. J., 2001, A comparison of auditory and blind separation techniques for speech segregation, *IEEE Trans. Speech Audio Process.* 9: 189-195.

van Veen, B. D., and Buckley, K. M., April 1988, Beamforming: A versatile approach to spatial filtering, *IEEE ASSP Magazine*, pp. 4-24.

Wang, D. L., 1996, Primitive auditory segregation based on oscillatory correlation, *Cognit. Sci.* 20: 409-456.

Wang, D. L., and Brown, G. J., 1999, Separation of speech from interfering sounds based on oscillatory correlation, *IEEE Trans. Neural Net.* 10: 684-697.

Weintraub, M., 1985, *A theory and computational model of auditory monaural sound separation*, Ph.D. Dissertation, Stanford University Department of Electrical Engineering.

Wrigley, S. N., and Brown, G. J., 2004, A computational model of auditory selective attention, *IEEE Trans. Neural Net.*, in press.

---

* Typo corrections made after publication are marked in the chapter.