

Research Article

A Computational Auditory Scene Analysis-Enhanced Beamforming Approach for Sound Source Separation

L. A. Drake,¹ J. C. Rutledge,² J. Zhang,³ and A. Katsaggelos (EURASIP Member)⁴

¹JunTech Inc., 2314 E. Stratford Ct, Shorewood, WI 53211, USA

²Computer Science and Electrical Engineering Department, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

³Electrical Engineering and Computer Science Department, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA

⁴Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA

Correspondence should be addressed to L. A. Drake, ladrake@ieee.org

Received 1 December 2008; Revised 18 May 2009; Accepted 12 August 2009

Recommended by Henning Puder

Hearing aid users have difficulty hearing target signals, such as speech, in the presence of competing signals or noise. Most solutions proposed to date enhance or extract target signals from background noise and interference based on either location attributes or source attributes. Location attributes typically involve arrival angles at a microphone array. Source attributes include characteristics that are specific to a signal, such as fundamental frequency, or statistical properties that differentiate signals. This paper describes a novel approach to sound source separation, called computational auditory scene analysis-enhanced beamforming (CASA-EB), that achieves increased separation performance by combining the complementary techniques of CASA (a source attribute technique) with beamforming (a location attribute technique), complementary in the sense that they use independent attributes for signal separation. CASA-EB performs sound source separation by temporally and spatially filtering a multichannel input signal, and then grouping the resulting signal components into separated signals, based on source and location attributes. Experimental results show increased signal-to-interference ratio with CASA-EB over beamforming or CASA alone.

Copyright © 2009 L. A. Drake et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

People often find themselves in cluttered acoustic environments, where what they want to listen to is mixed with noise, interference, and other acoustic signals of no interest. The problem of extracting an acoustic signal of interest from background clutter is called sound source separation and, in psychoacoustics, is also known as the “cocktail party problem.” Such “hearing out” of a desired signal can be particularly challenging for hearing aid users who often have reduced localization abilities. Sound source separation could allow them to distinguish better between multiple speakers, and thus, hear a chosen speaker more clearly. Separated signals from a sound source separation system can be further enhanced through techniques such as amplitude compression for listeners with sensorineural hearing loss and are also suitable for further processing in other applications, such as teleconferencing, automatic speech recognition, automatic transcription of ensemble music, and modeling the human auditory system.

There are three main approaches to the general sound source separation problem: blind source separation methods, those that use location attributes, and those that use source attributes. Blind source separation techniques separate sound signals based on the assumption that the signals are “independent,” that is, that their n th-order joint moments are equal to zero. When 2nd-order statistics are used, the method is called principal component analysis (PCA); when higher-order statistics are used, it is called independent component analysis (ICA). Blind source separation methods can achieve good performance. However, they require the observation data to satisfy some strict assumptions that may not be compatible with a natural listening environment. Besides the “independence” requirement, they can also require one or more of the following: a constant mixing process, a known and fixed number of sources, and an equal number of sources and observations [1]. Location and source attribute-based methods do not require any of these, and thus, are effective for a wider range of listening environments.

Location attributes describe the physical location of a sound source at the time it produced the sound. For example, a sound passes across a microphone array from some direction, and this direction, called “arrival angle,” is a location attribute. One location-attribute-based technique is binaural CASA [2–4]. Based on a model of the human auditory system, binaural sound source separation uses binaural data (sound “heard” at two “ears”) to estimate the arrival angle of “dominant” single-source sounds. It does this by comparing the binaural data’s interaural time delays and interaural intensity differences to a look-up table and selecting the closest match. While binaural CASA performance is impressive for a two microphone array (two ears), improved performance may be achieved by using larger arrays—as in beamforming. In addition to lifting the two microphone restriction of binaural CASA, microphone array processing is also more amenable to quantitative performance analysis since it is a mathematically derived approach.

Beamforming uses spatially sampled data from an array of two or more sensors to estimate arrival angles and waveforms of “dominant” signals in the wavefield. Generally, the idea is to combine the sensor measurements in some way so that desirable signals add constructively, while noise and interference are reduced. Various beamforming methods (taken and adapted from traditional array processing for applications such as radar and sonar) have been developed for and applied to speech and other acoustic signals. A review of these “microphone array processing” methods can be found in [5]. Regardless of which specific location method is chosen, however, and how well it works, it still cannot separate signals from the same (or from a close) location since location is its cue for separation [6].

In this paper, we present a novel technique combining beamforming with a source attribute technique, monaural CASA. This category of source attribute methods models how human hearing separates multispeaker input to “hear out” each speaker individually. Source attributes describe the state of the sound source at the time it produces the sound. For example, in the case of a voiced speech sound, fundamental frequency (F0) is a source attribute that indicates the rate at which the speaker’s glottis opens and closes. Monaural CASA [3, 7–12] is based on a model of how the human auditory system performs monaural sound source separation. It groups “time-frequency” signal components with similar source attributes, such as fundamental frequency (F0), amplitude modulation (AM), onset/offset times and timbre. Such signal component groups then give the separated sounds.

Location-attribute techniques can separate signals better in some situations than source-attribute techniques can. For example, since location attributes are independent of signal spectral characteristics, they can group harmonic and inharmonic signals equally well. Source-attribute techniques such as monaural CASA, on the other hand, have trouble with inharmonic signals. Similarly, when a signal changes its spectral characteristics abruptly, for example, from a fricative to a vowel in a speech signal, the performance of location-attribute techniques will not be affected. Source-attribute

techniques, on the other hand, may mistakenly separate the fricative and the vowel—assigning them to different sound sources.

Source-attribute techniques can also perform better than location-attribute methods in some situations. Specifically, they can separate sound mixtures in which the single-source signals have close or equal arrival angles. Their complementary strengths suggest that combining these two techniques may provide better sound source separation performance than using either method individually. Indeed, previously published work combining monaural and binaural CASA shows that this is a promising idea ([3, 13]).

In this paper, we exploit the idea of combining location and source attributes further by combining beamforming with monaural CASA into a novel approach called CASA-Enhanced Beamforming (CASA-EB). The main reason for using beamforming rather than binaural CASA as the location-attribute technique, here, is that beamforming may provide higher arrival angle resolution through the use of larger microphone arrays and adaptive processing. In addition, beamforming is more subject to quantitative analysis.

2. CASA-EB Overview

We begin by introducing some notation and giving a more precise definition of sound source separation. Suppose a multisource sound field is observed by an array of M acoustic sensors (microphones). This produces M observed mixture signals:

$$y[m, n] = \sum_{k=1}^K x_k[m, n] + w[m, n], \quad m = 1, 2, \dots, M, \quad (1)$$

where n is the time index, m is the microphone index, $x_k[m, n]$ is the k th source signal as observed at the m th microphone, and $w[m, n]$ is the noise in the observation (background and measurement noise). The goal of sound source separation, then, is to make an estimate of each of the K single-source signals in the observed mixture signals

$$\hat{x}_k[n], \quad k \in \{1, 2, \dots, K\}, \quad (2)$$

where $\hat{}$ is used to indicate estimation, and the estimate $\hat{x}_k[n]$ may differ from the source signal by a delay and/or scale factor.

In our CASA-EB approach, sound source separation is achieved in two steps. As shown in Figure 1, these are signal analysis and grouping. In the signal analysis step, the array observations, $y[m, n]$, are transformed into signal components in a 3D representation space with dimensions: time frame ρ , frequency band ω , and arrival angle band ϕ (see illustration in Figure 2). This is accomplished in two substeps—temporal filtering of $y[m, n]$ through a bandpass filterbank, followed by spatial filtering of the resulting band-pass signals. In the grouping step, selected signal components from this 3D CASA-EB representation space are grouped to form the separated single-source signals (see the illustration in Figure 3). Grouping consists of three substeps—selecting

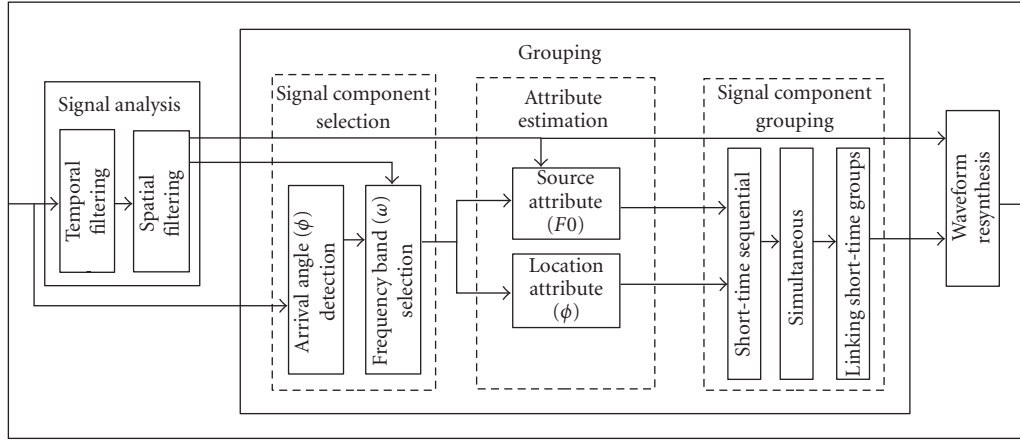
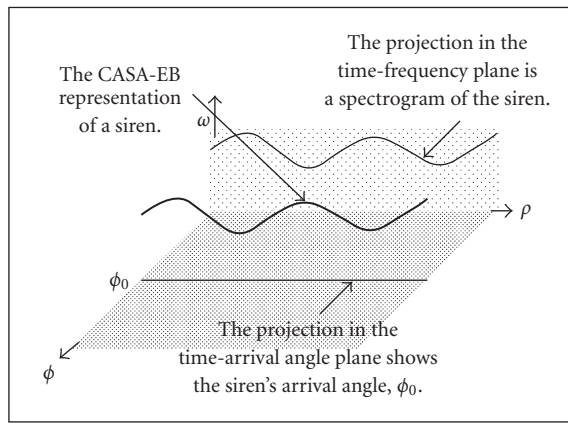
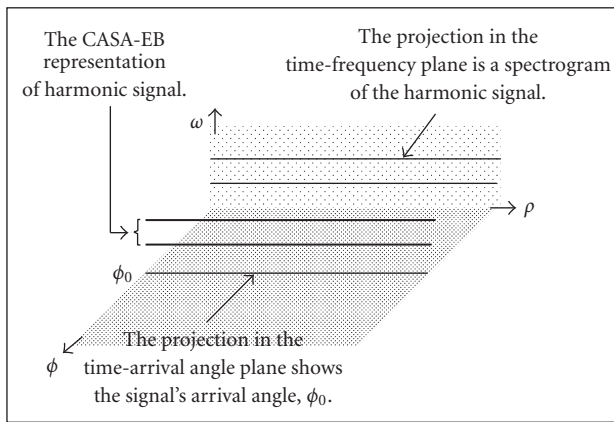


FIGURE 1: Block diagram of CASA-EB.



(a)



(b)

FIGURE 2: CASA-EB representations of a siren (a), and a simple harmonic signal (b). The projections on the time-frequency plane (signal's spectrogram) and time-arrival angle planes (signal's arrival angle path) are also shown.

signal components to group, estimating their attributes, and finally grouping selected signal components that share common attribute values.

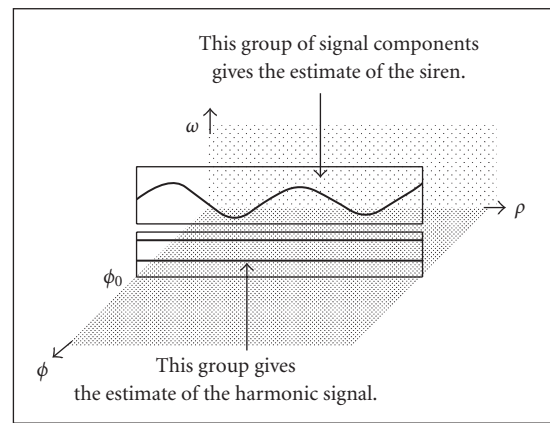


FIGURE 3: Separated signals from a two-signal mixture. This figure shows separated signal component groups from an example mixture signal—the sum of the two signals shown in Figure 2. The signal component groups are formed by collecting together signal components with similar location and source attributes (details in Section 4).

A summary of the CASA-EB processing steps and the methods used to implement are given in Table 1. The details of these are described below—signal analysis in Section 3 and grouping in Section 4. Then, Section 5 discusses how waveforms of the separated single-source signals can be synthesized from their signal component groups. Finally, after this presentation of the CASA-EB method, experimental results are presented in Section 6.

3. CASA-EB Representation Space

As just described, the first step in our approach is signal analysis. The array observations $y[m, n]$ are filtered along both the temporal and spatial dimensions to produce “frequency components”

$$Y[\phi, \omega, n] = \mathcal{T}_{\phi}\{y_{\omega}[m, n]\}, \text{ with} \tag{3}$$

$$y_{\omega}[m, n] = y[m, n] * h_{\omega}[n],$$

TABLE 1: Summary of CASA-EB methods.

	Processing block	Method
Signal analysis	Temporal filtering Spatial filtering	Gammatone filterbank Delay-and-sum beamformer
Grouping	Signal component selection (ϕ) Signal component selection (ω) Attribute estimation (F0) Attribute estimation (ϕ) Signal component grouping (short-time sequential) Signal component grouping (simultaneous) Signal component grouping (linking short-time groups)	STMV beamforming Signal detection using MDL criterion Autocorrelogram From $P[\phi, \omega, \rho]$ Kalman filtering with Munkres' optimal data assn algorithm Clustering via a hierarchical partitioning algorithm Munkres' optimal data assn algorithm
Waveform resynthesis	Over frequency Over time	Adding together grouped signal components Overlap-add

where $h_\omega[n]$ is a bandpass temporal filter associated with the frequency band indexed by ω , and \mathcal{T}_ϕ is a spatial transform associated with the arrival angle band indexed by ϕ (details of these signal analyses follow below).

The “frequency components” $Y[\phi, \omega, n]$ are used later in the processing (Section 4.2) for estimation of a grouping attribute, fundamental frequency, and also for waveform resynthesis. The signal components to be grouped in CASA-EB are those of its 3D representation shown in Figure 2; these are the power spectral components of the $Y[\phi, \omega, n]$, obtained in the usual way as the time-average of their magnitudes squared

$$P[\phi, \omega, \rho] = \frac{1}{N_\omega} \sum_{n'=\rho T-(N_\omega-1)/2}^{\rho T+(N_\omega-1)/2} |Y[\phi, \omega, n']|^2, \quad (4)$$

where the $P[\phi, \omega, \rho]$ are downsampled from the $Y[\phi, \omega, n]$ with downsampling rate T , that is, $\rho = n/T$, and N_ω is the number of samples of $Y[\phi, \omega, n]$ in frequency band ω that are used to compute one sample of $P[\phi, \omega, \rho]$.

3.1. Temporal Filtering. For the *temporal filterbank*, $h_\omega[n]$, $\omega \in \{1, 2, \dots, \Omega\}$, we have used a modified gammatone filterbank. It consists of constant-Q filters in high frequency bands (200 to 8000 Hz) and constant-bandwidth filters in lower frequency bands (below 200 Hz) (Constant-Q filters are a set of filters that all have the same quotient (Q), or ratio of center frequency to bandwidth.). Specifically, the constant-Q filters are the fourth-order gammatone functions,

$$h_\omega[n] = \alpha^\omega \cdot e^{-\beta(\alpha^\omega n T_s)} (\alpha^\omega n T_s)^3 e^{j2\pi f_s/2(\alpha^\omega n T_s)} u[n], \quad (5)$$

where the frequency band indices ($\omega = 1, 2, \dots, 75$) are in reverse order, that is, the lower indices denote higher frequencies, f_s and T_s are the sampling frequency and sampling period, $u[n]$ is the unit step function, and α and β are parameters that can be used to adjust filter characteristics such as bandwidths and spacing on the frequency axis. For

CASA-EB, $\alpha = 0.95$, and $\beta = 2000$ work well. The constant-bandwidth filters are derived by downshifting the lowest frequency constant-Q filter ($\omega = 75$) by integer multiples of its bandwidth

$$h_\omega[n] = h_{75}[n] e^{-j2\pi(\omega-75)B_{75}n}, \quad (6)$$

where $\omega = 76, 77, \dots, 90$, and B_{75} is the bandwidth of the lowest frequency constant-Q filter.

The modified gammatone filterbank is used for temporal filtering because it divides the frequency axis efficiently for CASA-EB. Specifically, for CASA, the frequency bands are just narrow enough that the important spectral features of a signal (such as harmonics in low frequencies and formants in high frequencies) can be easily distinguished from each other. For beamforming, the bands are narrow enough to limit spatial filtering errors to an acceptable level.

3.2. Spatial Filtering. The *spatial transform*, \mathcal{T}_ϕ , that we are using is the well-known delay-and-sum beamformer

$$\begin{aligned} \mathcal{T}_\phi\{y_\omega[m, n]\} &= \frac{1}{M} \sum_{m=1}^M y_\omega[m, n] \cdot e^{j2\pi(m-1)f_\phi}, \text{ with} \\ f_\phi &= f_\omega \frac{d}{C} \sin \phi, \\ \phi &\in \left[-\frac{\pi}{2}, +\frac{\pi}{2}\right], \end{aligned} \quad (7)$$

where f_ω is the center frequency of frequency band ω , d is the distance between adjacent microphones in a uniform linear array, and C is the speed of sound at standard temperature and pressure.

Delay-and-sum beamforming is used here for the signal analysis in our general solution to the sound source separation problem because it does not cancel correlated signals, for example, echos (as MV beamforming can), and does not require a priori information or explicit modeling of target signals, interferers, or noise (as other data adaptive

beamforming can). Its drawback is that, since it has relatively low arrival angle resolution, each signal component will contain more interference from neighboring arrival angle bands. In CASA-EB, this is ameliorated somewhat by the additional separation power provided by monaural CASA. For specific applications, CASA-EB performance may be improved by defining signal and/or noise models and using a data adaptive beamformer.

In summary, the 3D CASA-EB representation space consists of signal components $P[\phi, \omega, \rho]$ generated by filtering a temporally and spatially sampled input signal along both of these dimensions (to produce frequency components $Y[\phi, \omega, n]$), and then, taking the average magnitude squared of these.

4. CASA-EB Grouping to Separate Single-Source Signals

As described previously, the second step in CASA-EB is to group signal components from the time-frequency-arrival angle space into separated single-source signal estimates. Grouping consists of three steps: selecting the signal components for grouping, estimating their location and source attributes, and finally, grouping those with similarly valued attributes to form the separated single-source signal estimates. The details of these three steps are given in the following three subsections.

4.1. Signal Component Selection. In this step, the set of all signal components ($P[\phi, \omega, \rho]$) is pruned to produce a subset of “significant” signal components, which are more likely to have come from actual sound sources of interest and to constitute the main part of their signals. Grouping is then performed using only this subset of signals. Experience and experimental results indicate that this type of before-grouping pruning does not adversely affect performance and has the following two benefits. First, it reduces the computational complexity of grouping and second, it increases grouping robustness (since there are fewer spurious signal components to throw the grouping operation “off-track”). Now, we describe the signal component selection process in more detail.

4.1.1. Arrival Angle Detection. This process begins with pruning away signal components from arrival angles in which it is unlikely there is any audible target sound, that is, from angles within which the signal power is low. There are a variety of ways to detect such low-power arrival angles. For example, a simple way is, for a given time frame ρ , to add up the power spectral components $P[\phi, \omega, \rho]$ in each arrival angle band ϕ

$$P[\phi] = \sum_{\omega} P[\phi, \omega, \rho]. \quad (8)$$

In this work, we are using a wideband adaptive beamformer by Krolik—the steered minimum variance (STMV) beamformer [14]. This wideband method is an adaptation of

Capon’s [15] narrowband minimum variance (MV) beamformer. The MV beamformer is a constrained optimization method that produces a spatial spectral estimate in which power is minimized subject to the constraint of unity gain in the look direction, that is,

$$\min_{\mathbf{w}} [\mathbf{w}^+ \cdot \mathbf{R}_f \cdot \mathbf{w}] \quad \text{subject to } \mathbf{a}_f^+(\phi) \cdot \mathbf{w} = 1, \quad (9)$$

where \mathbf{w} is the beamformer weight vector, \mathbf{R}_f is the covariance matrix of a narrowband array observation vector with frequency f , $^+$ indicates conjugate transpose, and $\mathbf{a}_f(\phi) = [1 \ e^{-j2\pi f t_1(\phi)} \ \dots \ e^{-j2\pi f t_{M-1}(\phi)}]^T$ is the “steering vector.” The solution to (9) gives the MV beamformer spatial spectral estimate:

$$P_f[\phi] = [\mathbf{a}_f^+(\phi) \cdot \mathbf{R}_f^{-1} \cdot \mathbf{a}_f(\phi)]^{-1}. \quad (10)$$

To apply this narrowband method to a wideband signal, one could just filter the wideband array observations, apply the narrowband method individually in each band, and then sum up the results across frequency. This “incoherent” wideband method, however, does not take full advantage of the greater statistical stability of the wideband signal—a goal of wideband methods such as STMV beamforming. To achieve this goal, a wideband method must use a statistic computed across frequency bands.

In light of the above, STMV beamforming is an adaptation of MV beamforming in which a wideband composite covariance matrix ($\mathbf{R}^{\text{st}}[\phi]$ defined below) is used in place of the narrowband one, and the steering vector in the constraint is adjusted appropriately (more on this below):

$$\min_{\mathbf{w}} [\mathbf{w}^+ \cdot \mathbf{R}^{\text{st}}[\phi] \cdot \mathbf{w}] \quad \text{subject to } \mathbf{1}^T \cdot \mathbf{w} = 1, \quad (11)$$

where $\mathbf{1}$ is an $M \times 1$ vector of ones. The STMV beamformer solution is

$$P[\phi] = [\mathbf{1}^T \cdot \mathbf{R}^{\text{st}}[\phi]^{-1} \cdot \mathbf{1}]^{-1}. \quad (12)$$

To compute the wideband composite covariance matrix $\mathbf{R}^{\text{st}}[\phi]$ from the array observation vectors, some preprocessing is performed first. The $y[m, n]$ are bandpass filtered (as in (3)), and then the resulting narrowband signals are “presteered” as follows:

$$y_{\omega}^{\text{st}}[m, n] = \mathbf{T}^{\text{st}}[f_{\omega}, \phi] \cdot y_{\omega}[m, n], \quad (13)$$

where f_{ω} is the center frequency of frequency band ω , the steering matrix $\mathbf{T}^{\text{st}}[f_{\omega}, \phi]$ is a diagonal matrix with diagonal elements $[1 \ e^{j2\pi f_{\omega} t_1(\phi)} \ \dots \ e^{j2\pi f_{\omega} t_{(M-1)}(\phi)}]$, and $t_m(\phi)$ is the time delay between the m th sensor and a reference sensor (sensor 1) for a narrowband signal $e^{-j2\pi f_{\omega} t}$ from angle ϕ . Such presteering has the effect of zeroing out inter-sensor time delays $t_m(\phi)$ in narrowband signals from angle ϕ . For example, for the narrowband signal $\mathbf{s}(t) = [1 \ e^{-j2\pi f_{\omega} t_1(\phi)} \ \dots \ e^{-j2\pi f_{\omega} t_{M-1}(\phi)}]$,

$$\mathbf{T}^{\text{st}}[f_{\omega}, \phi] \cdot \mathbf{s}(t) = \mathbf{1}. \quad (14)$$

Thus, the effect of preprocessing the wideband array observations is to make the steering vectors equal for

all frequency bands ($\mathbf{a}_f(\phi) = \mathbf{1}$), and this provides a frequency-independent steering vector to use in the STMV beamformer's unity-gain constraint.

Now, given the presteered array observations, the wide-band composite covariance matrix is simply

$$\begin{aligned} \mathbf{R}^{\text{st}}[\phi] &= \sum_{\omega=l}^h \sum_{n=n_0}^{n_0+(N-1)} \mathbf{y}_{\omega}^{\text{st}}[m, n] \cdot \mathbf{y}_{\omega}^{\text{st}*}[m, n], \\ &= \sum_{\omega=l}^h \mathbf{T}^{\text{st}}[f_{\omega}, \phi] \cdot \mathbf{R}_{\omega} \cdot \mathbf{T}^{\text{st}*}[f_{\omega}, \phi], \end{aligned} \quad (15)$$

where \mathbf{R}_{ω} is the covariance matrix of $\mathbf{y}_{\omega}[m, n]$, and the summations run from frequency band l to h and from time index n_0 to $n_0 + (N - 1)$.

The advantage of Krolik's technique over that of (8) and other similar data-independent beamforming techniques is that it provides higher arrival angle resolution. Compared to other data adaptive methods, it does not require a priori information about the source signals and/or interference, does not cancel correlated signals (as MV beamforming is known to do), and is not vulnerable to source location bias (as other wideband adaptive methods, such as the coherent signal-subspace methods, are [16]).

4.1.2. Frequency Band Selection. Now, for each detected arrival angle band, ϕ_0 , the next step is to select the significant signal components from that arrival angle band. This is done in two steps. First, high-power signal components are detected, and low-power ones pruned. Then, the high-power components are further divided into peaks (i.e., local maxima) and their neighboring nonpeak components. Although all the high-power components will be included in the separated signals, only the peak components need to be explicitly grouped. Due to the nature of the gammatone filterbank we are using, the non-peak components can be added back into the separated signal estimates later at signal reconstruction time, based on their relationship with a peak. Consider the following. Since the filterbank's neighboring frequency bands overlap, a high-power frequency component sufficient to generate a peak in a given band is also likely to contribute significant related signal power in neighboring bands (producing non-peak components). Thus, these non-peak components are likely to be generated by the same signal feature as their neighboring peak, and it is reasonable to associate them.

Low-power signal components are detected and pruned using a technique by Wax and Kailath [17]. In their work, a covariance matrix is computed from multichannel input data, and its eigenvalues are sorted into a low-power set (from background noise) and a high-power set (from signals). The sorting is accomplished by minimizing an information theoretic criterion, such as Akaike's Information Criterion (AIC) [18, 19] or the Minimum Description Length (MDL) criterion [20, 21]). The MDL is discussed here

since it is the one used in CASA-EB. From [17], it is defined as

$$\begin{aligned} \text{MDL} &= -\log \left(\frac{\prod_{i=\lambda+1}^L l_i^{1/(L-\lambda)}}{(1/(L-\lambda)) \cdot \sum_{i=\lambda+1}^L l_i} \right)^{(L-\lambda)N_t} \\ &\quad + \frac{1}{2} \lambda (2L - \lambda) \log N_t, \end{aligned} \quad (16)$$

where $\lambda \in \{0, 1, \dots, L - 1\}$ is the number of possible signal eigenvalues and the parameter over which the MDL is minimized, L is the total number of eigenvalues, l_i is the i th largest eigenvalue, and N_t is the number of time samples of the observation vectors used to estimate the covariance matrix. The λ that minimizes the MDL (λ_{\min}) is the estimated number of signal eigenvalues, and the remaining $(L - \lambda_{\min})$ smallest eigenvalues are the detected noise eigenvalues. Notice, this MDL criterion is entirely a function of the $(L - \lambda)$ smallest eigenvalues, and not the larger ones. Thus, in practice, it distinguishes between signal and noise eigenvalues based on the characteristics of the background noise. Specifically, it detects a set of noise eigenvalues with relatively low and approximately equal power. Wax and Kailath use this method to estimate the number of signals in multichannel input data. We use it to detect and remove the $(L - \lambda_{\min})$ low-power, noise components $P[\phi, \omega, \rho]$ —by treating the $P[\phi, \omega, \rho]$ as the eigenvalues in their method. We chose this method for noise detection because it works based on characteristics of the noise, rather than relying on arbitrary threshold setting.

In summary, signal component selection/pruning is accomplished in two steps. For each fixed time frame ρ , high power arrival angle bands are detected, and signal components from low power arrival angle bands are removed. Then, in high power arrival angle bands, low-power signal components are removed and high-power signal components are divided into peaks (for grouping) and non-peaks (to be added back into the separated signal estimates after grouping, at signal reconstruction time).

4.2. Attribute Estimation. In the previous section, we described how signal components in the CASA-EB representation can be pruned and selected for grouping. In this section, we describe how to estimate the selected signal components' attributes that will be used to group them. In this work, we estimate two types of signal attributes, location attributes and source attributes. As described in the introduction, these are complementary. Used together, they may allow more types of sound mixtures to be separated and produce more completely separated source signals.

4.2.1. Location Attribute. For a selected signal component, $P[\phi, \omega, \rho]$, the location attribute used in CASA-EB is its arrival angle band, or simply its ϕ index. This is the delay-and-sum beamformer steering angle from the spatial filtering step in Section 3.

4.2.2. Source Attribute. Source attributes are features embedded in a signal that describe the state of the signal's source

at the time it produced the signal. In the previous work, several different source attributes have been used, including F0 [2, 3, 8–11, 22, 23], amplitude modulation [8], onset time [9, 23], offset time [9], and timbre [24]. In this work, we use an F0 attribute. Since F0 is the most commonly used, its use here will allow our results to be compared to those of others more easily. Next, we discuss F0 estimation in more detail.

There are two main approaches to F0 estimation: spectral peak-based and autocorrelation-based methods. The spectral peak-based approach is straightforward when there is only one harmonic group in the sound signal. In this case, it detects peaks in the signal's spectrum and estimates F0 by finding the greatest common divisor of their frequencies. However, complications arise when the signal contains more than one harmonic group. Specifically, there is the added “data association problem,” that is, the problem of determining the number of harmonic groups and which spectral peaks belong to which harmonic groups. The autocorrelation-based approach handles the data association problem more effectively and furthermore, as indicated in [25], also provides more robust F0 estimation performance. Hence, an autocorrelation-based method is used in this work.

The basic idea behind the autocorrelation method is that a periodic signal will produce peaks in its autocorrelation function at integer multiples of its fundamental period, and these can be used to estimate F0. To use F0 as an attribute for grouping signal components, however, it is also necessary to be able to associate the signal components $P[\phi, \omega, \rho]$ with the F0 estimates. This can be done using an extension of the autocorrelation method—the autocorrelogram method.

Detailed descriptions of the autocorrelogram method can be found in [9–11, 25–30]. To summarize here, the steps of this method are the following. First, an input signal $X[n]$ is filtered either by a set of equal-bandwidth bandpass filters covering the audible range of frequencies, or more often, by a filtering system based more closely on the human auditory system, such as a gammatone filterbank. This filtering produces the bandpass signals $X_\omega[n]$. Then, to form the autocorrelogram, an autocorrelation of the filtered signal is computed in each band and optionally normalized by the signal power in the band:

$$\text{acm}[\omega, \tau] = \frac{R_{X_\omega}[\tau]}{R_{X_\omega}[0]}. \quad (17)$$

For an illustration, see Figure 4. Next, a summary autocorrelogram is computed by combining the narrowband autocorrelations over frequency and optionally applying a weighting function to emphasize low-frequency peaks:

$$\text{sacm}[\tau] = \frac{1}{\Omega} \sum_{\omega=1}^{\Omega} \text{acm}[\omega, \tau] \cdot w[\tau], \quad (18)$$

where

$$w[\tau] = \exp\left[\frac{-\tau}{N_\tau}\right] \quad (19)$$

is a low frequency emphasis function, and N_τ is the number of time lags at which the autocorrelogram is computed.

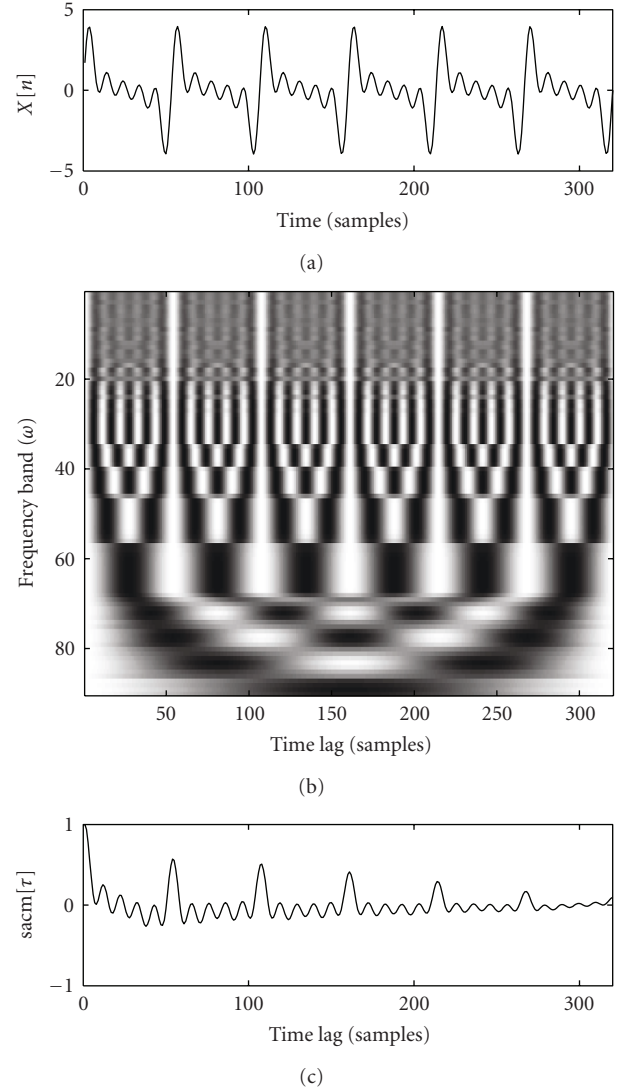


FIGURE 4: Autocorrelogram representation of a sum of sinusoids. The signal, $X[n] = \sum_{r=1}^5 \sin(2\pi 300r \cdot nT_s)$, with $T_s = 1/16000$ s/sample is shown in (a). (b) shows the power-normalized autocorrelogram, $\text{acm}[\omega, \tau] = R_{X_\omega}[\tau]/R_{X_\omega}[0]$, where $R_{X_\omega}[\tau]$ is the autocorrelation of the filtered signal, $X_\omega[n] = X[n] * h_\omega[n]$. Here, the maximum value is displayed in white, the minimum in black. Finally, the summary autocorrelogram, $\text{sacm}[\tau] = ((1/\Omega) \cdot \sum_{\omega=1}^{\Omega} \text{acm}[\omega, \tau]) \cdot w[\tau]$ is shown in (c).

For an example of the summary autocorrelogram, see Figure 4. Finally, F0 estimates are made based on peaks in the summary autocorrelogram, and overtones of these are identified by associating peaks in the autocorrelogram with the F0-estimate peaks in the summary autocorrelogram.

For CASA-EB, we are using the following implementation of the autocorrelogram method. In each time frame ρ , an autocorrelogram and summary autocorrelogram are computed for each detected arrival angle band ϕ_0 (from Section 4.1), and a single F0 analysis is made from each such autocorrelogram/summary autocorrelogram pair. That

is, for each ϕ_0 , an autocorrelogram and summary autocorrelogram are computed from the temporally and spatially filtered signal, $Y[\phi_0, \omega, n]$, $\omega \in \{1, 2, \dots, \Omega\}$ and $n \in \{\rho T - N_\tau/2 + 1, \dots, \rho T + N_\tau/2\}$, where we used $N_\tau = 320$ (equivalent to 20 milliseconds). Then, for this arrival angle band and time frame, the F0 estimation method of Wang and Brown [11] is applied, producing a single F0 estimate made from the highest peak in the summary autocorrelogram

$$\hat{F}0[\phi_0, \rho], \quad (20)$$

and a set of flags, indicating for each $P[\phi_0, \omega, \rho]$, whether it contains a harmonic of $\hat{F}0[\phi_0, \rho]$ or not

$$\text{FN}[\phi_0, \omega, \rho], \quad \omega \in \{1, 2, \dots, \Omega\}. \quad (21)$$

Here, $\text{FN}[\phi_0, \omega, \rho] = 1$ when band ω contains a harmonic, and 0 otherwise. Details of the implementation are the following.

Temporal filtering is done with a gammatone filterbank because its constant-Q filters can resolve important low-frequency features of harmonic signals (the fundamental and its lower frequency harmonics) better than equal-bandwidth filterbanks with the same number of bands (Low frequency harmonics are important since, in speech for example, they account for much of the signal power in vowels). These better-resolved, less-mixed low frequency harmonics can give better F0 estimation results (F0 estimates and related harmonic flags, FN's), since they produce sharper peaks in the autocorrelogram, and these sharper peaks are easier for the F0 estimation algorithm to interpret. Spatial filtering (new to autocorrelogram analysis) is used here because it provides the advantage of reducing interference in the autocorrelogram when multiple signals from different spatial locations are present in the input.

The autocorrelogram is computed as described previously, including the optional power normalization in each frequency band. For the summary autocorrelogram, however, we have found that F0 estimation is improved by using just the lower frequency bands that contain the strongest harmonic features. Thus,

$$\text{sacm}[\tau] = \frac{1}{74} \sum_{\omega=17}^{90} \text{acm}[\omega, \tau] \cdot w[\tau], \quad (22)$$

where the bands, 90 to 17, cover the frequency range, 0, to 3500 Hz, the frequency range of a vowel's fundamental and its lower harmonics.

Finally, an F0 analysis is performed using the autocorrelogram/summary autocorrelogram pair, according to the method of Wang and Brown [11]. Their method is used in CASA-EB to facilitate comparison testing of CASA-EB'S monaural CASA to their monaural CASA (described in Section 6). The details of the method are the following. First, a single F0 is estimated based on the highest peak in the summary autocorrelogram:

$$\hat{F}0[\phi_0, \rho] = \frac{f_s}{\tau_m}, \quad (23)$$

where f_s is the temporal sampling frequency of the input signal $y[m, n]$, and τ_m is the time lag of the highest peak in the summary autocorrelogram. Then, the associated overtones of this F0 are identified by finding frequency bands in the autocorrelogram with peaks at, or near, τ_m . Specifically, this is done as follows. A band ω is determined to contain an overtone, that is, $\text{FN}[\phi_0, \omega, \rho] = 1$, when

$$\frac{R_{X_\omega}[\tau_m]}{R_{X_\omega}[0]} > \Theta_d, \quad (24)$$

and $\Theta_d = 0.90$ is a detection threshold. Wang and Brown used $\Theta_d = 0.95$. For CASA-EB, experiments show that Θ_d 's in the range of 0.875 to 0.95 detect overtones well [31]. This F0 estimation method amounts to estimating F0 and detecting its overtones for a single "foreground signal," and treating the rest of the input mixture signal as background noise and interference. Although this limits the number of signals for which an F0 estimate is made (one per autocorrelogram), it also helps by eliminating the need to estimate the number of harmonic signals. Further, it provides more robust F0 estimation since, from each autocorrelogram, an F0 estimate is only made from the signal with the strongest harmonic evidence (the highest peak in the summary autocorrelogram).

Notice that in our application, the number of signals for which F0 estimates can be made is less limited since we have more than one autocorrelogram per time frame (one for each detected arrival angle). Additionally, our F0 estimates may be better since they are made from autocorrelograms with less interharmonic group interference. Such interference is reduced since the autocorrelograms are computed from the spatially filtered signals, $Y[\phi_0, \omega, n]$, $\omega \in \{1, 2, \dots, \Omega\}$, that are generally "less mixed" than the original input mixture signal $y[m, n]$ because they contain a smaller number of harmonic groups with significant power.

4.3. Signal Component Grouping. Recall that sound source separation consists of two steps: signal analysis (to break the signal into components such as $P[\phi, \omega, \rho]$), and signal component grouping (to collect the components into single source signal estimates). Grouping collects together signal components according to their attributes (estimated in Section 4.2), and ideally, each group only contains pieces from a single source signal.

Grouping is typically done in two stages: simultaneous grouping clusters together signal components in each time frame ρ that share common attribute values, and sequential grouping tracks these simultaneous groups across time. In the previous work, many researchers perform simultaneous grouping first and then track the resulting clusters [2, 3, 10, 22, 32]. For signals grouped by the F0 source attribute, for example, the simultaneous grouping step consists of identifying groups of harmonics, and the sequential grouping step consists of tracking their fundamental frequencies. A primary advantage of simultaneous-first grouping is that it can be real-time amenable when the target signals' models are known a priori. However, when they are not known, it can be computationally complex to determine the correct

signal models [10], or error-prone if wrong signal models are used.

Some researchers have experimented with sequential-first grouping [8, 9]. In this case, the sequential grouping step consists of tracking individual signal components, and the simultaneous grouping step consists of clustering together the tracks that have similar source attribute values in the time frames in which they overlap. Although this approach is not real-time amenable since tracking is performed on the full length of the input mixture signal before the resulting tracks are clustered, it has the advantage that it controls error propagation. It does this by putting off the more error-prone decisions (simultaneous grouping's signal modeling decisions) until later in the grouping process.

In this work, we strike a balance between the two with a *short-time sequential-first* grouping approach. This is a three-step approach (illustrated in Figure 5). First, to enjoy the benefits of sequential-first grouping (reduced error-propagation) without suffering long time delays, we start by tracking individual signal components over a few frames. Then, these short-time frequency component tracks are clustered together into short-time single-source signal estimates. Finally, since signals are typically longer than a few frames, it is necessary to connect the short-time signal estimates together (i.e., to track them). The details of these three steps are given next.

4.3.1. Short-Time Sequential Grouping. In this step, signal components are tracked for a few frames (six for the results presented in this paper). Recall from Section 4.1 that the signal components that are tracked are the perceptually significant ones (peak, high-power components from arrival angle bands in which signals have been detected). Limiting tracking to these select signal components reduces computational complexity and improves tracking performance.

Technically, tracking amounts to estimating the *state* of a target (e.g., its position and velocity) over time from related observation data. A target could be an object, a system, or a signal, and a sequence of states over time is called a *track*. In our application, a target is a signal component of a single sound source's signal (e.g., the n th harmonic of a harmonic signal), its state consists of parameters (e.g., its frequency) that characterize the signal component, and the observation data in each frame ρ consists of the (multi source) signal components $P[\phi, \omega, \rho]$.

Although we are tracking multiple targets (signal component sequences), for the sake of simplicity, we first consider the tracking of a single target. In this case, a widely used approach for tracking is the Kalman filter [33]. This approach uses a linear system model to describe the dynamics of the target's internal state and observable output, that is,

$$\begin{aligned} \mathbf{x}[\rho + 1] &= \mathbf{A}[\rho] \cdot \mathbf{x}[\rho] + \mathbf{v}[\rho], \\ \mathbf{z}[\rho + 1] &= \mathbf{C}[\rho + 1] \cdot \mathbf{x}[\rho + 1] + \mathbf{w}[\rho + 1]. \end{aligned} \quad (25)$$

Here, $\mathbf{x}[\rho + 1]$ is the target's state and $\mathbf{z}[\rho + 1]$ is its observable output in time frame $(\rho + 1)$, $\mathbf{A}[\rho]$ is the state transition matrix, $\mathbf{C}[\rho + 1]$ is the matrix that transforms the current

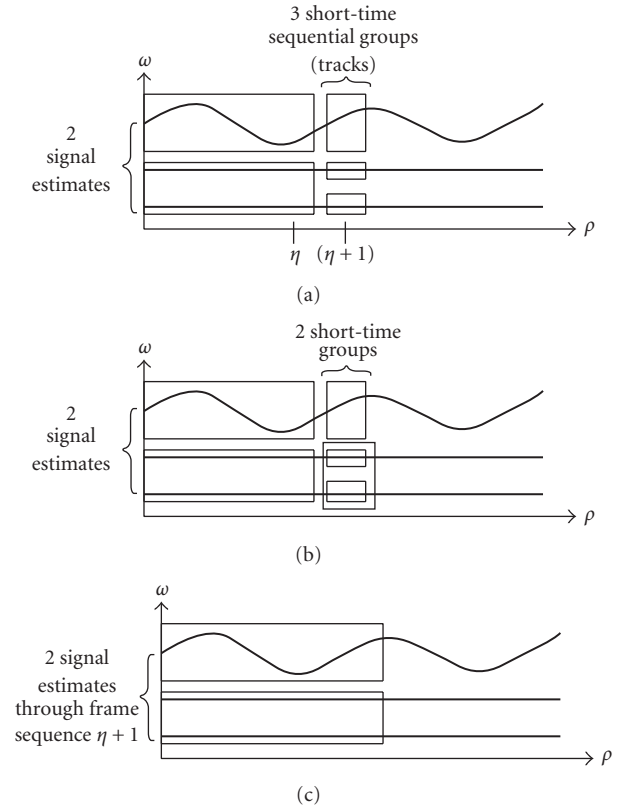


FIGURE 5: Illustration of *short-time sequential-first* grouping. Here the input signal is a mixture of the two single-source signals shown in Figure 2. (a) The graph shows short-time tracks in time segment $(\eta + 1)$ with completed signal estimate groups through time segment η . Here, time segment η consists of time frames $\rho \in \{\eta T', \dots, (\eta + 1)T' - 1\}$, and $T' = 6$. (b) The graph shows simultaneous groups of the short-time tracks shown in (a). (c) The graph shows completed signal estimate groups through time segment $(\eta + 1)$.

state of the track to the output, and $\mathbf{v}[\rho]$ and $\mathbf{w}[\rho]$ are zero-mean white Gaussian noise with covariance matrices $\mathbf{Q}[\rho]$ and $\mathbf{R}[\rho]$, respectively. Based on this model, the Kalman filter is a set of time-recursive equations that provides optimal state estimates. At each time $(\rho + 1)$, it does this in two steps. First, it computes an optimal prediction of the state $\mathbf{x}[\rho + 1]$ from an estimate of the state $\mathbf{x}[\rho]$. Then, this prediction is updated/corrected using the current output $\mathbf{z}[\rho + 1]$, generating the final estimate of $\mathbf{x}[\rho + 1]$.

Since the formulas for Kalman prediction and update are well known [33], the main task for a specific application is reduced to that of constructing the linear model, that is, defining the dynamic equations (see (25)). For CASA-EB, a target's output vector, $\mathbf{z}[\rho]$, is composed of its frequency and arrival angle bands, and its internal state, $\mathbf{x}[\rho]$, consists of its frequency and arrival angle bands, along with their rates of change:

$$\begin{aligned} \mathbf{z}[\rho] &= [\phi \quad \omega]^T, \\ \mathbf{x}[\rho] &= \left[\phi \quad \frac{d}{dt}\phi \quad \omega \quad \frac{d}{dt}\omega \right]^T. \end{aligned} \quad (26)$$

The transition matrices of the state and output equations are defined as follows:

$$A[\rho] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (27)$$

$$C[\rho] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

where this choice of $A[\rho]$ reflects our expectation that the state changes slowly, and this $C[\rho]$ simply picks the output vector ($[\phi \ \omega]^T$) from the state vector.

When there is more than one target, the tracking problem becomes more complicated. Specifically, at each time instant, multiple targets can produce multiple observations, and generally, it is not known which target produced which observation. To solve this problem, a data association process is usually used to assign each observation to a target. Then, Kalman filtering can be applied to each target as in the single target case.

While a number of data association algorithms have been proposed in the literature, most of them are based on the same intuition—that an observation should be associated with the target most likely to have produced it (e.g., the “closest” one). In this work, we use an extension of Munkres’ optimal data association algorithm (by Burgeois and Lassalle [34]). A description of this algorithm can be found in [35]. To summarize briefly here, the extended Munkres algorithm finds the best (lowest cost) associations of observations to established tracks. It does this using a cost matrix with H columns (one per observation) and $J+H$ rows (one per track plus one per observation), where the $(j, h)^{\text{th}}$ element is the cost of associating observation h to track j , the $(J+h, h)^{\text{th}}$ element is the cost of initiating a new track with observation h , and the remaining off-diagonal elements in the final H rows are set to a large number such that they will not affect the result.

The cost of associating an observation with a track is a function of the distance between the track’s predicted next output and the observation. Specifically, we are using the following distance measure:

$$\text{cost}_{j,h} = \begin{cases} |\hat{\omega}_j - \omega_h|, & \text{when } |\hat{\omega}_j - \omega_h| \leq 1 \text{ and } \phi_h = \phi_j, \\ 2\gamma, & \text{otherwise,} \end{cases} \quad (28)$$

where $\hat{\omega}_j$ is the prediction of track j ’s next frequency (as computed by the Kalman filter), ω_h and ϕ_h are the frequency and arrival angle of observation h , respectively, and track j ’s arrival angle band ϕ_j is constant. Finally, γ is an arbitrary large number used here so that if observation h is outside track j ’s validation region, ($|\hat{\omega}_j - \omega_h| > 1$ or $\phi_h \neq \phi_j$), then observation h will not be associated with track j . Note that this cost function means that frequency tracks change their frequency slowly (≤ 1 frequency band per time frame),

and sound sources do not move (since ϕ_j is held constant). In subsequent work, the assumption of unmoving sources could be lifted by revising the cost matrix and making adjustments to the simultaneous grouping step (described next in Section 4.3.2).

Finally, the cost of initiating a new track is simply set to be larger than the size of the *validation region*

$$\text{cost}_{J+h,h} = \gamma, \quad (29)$$

and the remaining costs in the last H rows are set equal to 2γ so that they will never be the low cost choice.

4.3.2. Simultaneous Grouping. In this step, the short-time tracks from the previous step are clustered into short-time signal estimates based on the similarity of their source and location attribute values. There are a variety of clustering methods in the literature (refer to pattern recognition texts, such as [36–40]). In CASA-EB, we use the hierarchical partitioning algorithm that is summarized next.

Partitioning is an iterative approach that divides a measurement space into k disjoint regions, where k is a predefined input to the partitioning algorithm. In general, however, it is difficult to know k a priori. Hierarchical partitioning addresses this issue by generating a hierarchy of partitions—over a range of different k values—from which to choose the “best” partition. The specific steps are the following. (1) Initialize k to be the minimum number of clusters to be considered. (2) Partition the signal component tracks into k clusters. (3) Compute a performance measure to quantify the quality of the partition. (4) Increment k by 1 and repeat steps 2–4, until a stopping criterion is met, or k reaches a maximum value. (5) Select the best partition based on the performance measure computed in step 3.

To implement the hierarchical partitioning algorithm, some details remain to be determined: the minimum and maximum number of clusters to be considered, the partitioning algorithm, the performance measure, and a selection criterion to select the best partition based on the performance measure. For CASA-EB, we have made the following choices. For the minimum and maximum numbers of clusters, we use the number of arrival angle bands in which signals have been detected, and the total number of arrival angle bands, respectively.

For partitioning algorithms, we experimented with a deterministic one, partitioning around medoids (PAMs), and a probabilistic one, fuzzy analysis (FANNY)—both from a statistics shareware package called R [41, 42]. (R is a reimplement of S [43, 44] using Scheme semantics. S is a very high level language and an environment for data analysis and graphics. S was written by Richard Becker, John M. Chambers, and Allan R. Wilks of AT&T Bell Laboratories Statistics Research Department.) The difference between the two is in how measurements are assigned to clusters. PAM makes hard clustering assignments; that is, each measurement is assigned to a single cluster. FANNY, on the other hand, allows measurements to be spread across multiple clusters during partitioning. Then, if needed, these fuzzy assignments can be hardened at the end (after the last

iteration). For more information on PAM and FANNY, refer to [37]. For CASA-EB, we use FANNY since it produces better clusters in our experiments.

Finally, it remains to discuss performance measures and selection criteria. Recall that the performance measure's purpose in hierarchical partitioning is to quantify the quality of each partition in the hierarchy. Common methods for doing this are based on "intracluster dissimilarities" between the members of each cluster in a given partition (small is good), and/or on "intercluster dissimilarities" between the members of different clusters in the partition (large is good). As it turns out, our data produces clusters that are close together. Thus, it is not practical to seek clusters with large inter-cluster dissimilarities. Rather, we have selected a performance measure based on intra-cluster dissimilarities. Two intra-cluster performance measures were considered: the maximum intra-cluster dissimilarity in any single cluster in the partition, and the mean intra-cluster dissimilarity (averaged over all clusters in the partition). The maximum intra-cluster dissimilarity produced the best partitions for our data and is the one we used. The details of the dissimilarity measure are discussed next.

Dissimilarity is a measure of how same/different two measurements are from each other. It can be computed in a variety of ways depending on the measurements being clustered. The measurements we are clustering are the source and location attribute vectors of signal component tracks. Specifically, for each short-time track j in time segment η , this vector is composed of the track's arrival angle band ϕ_j , and its F0 attribute in each time frame ρ of time segment η in which the track is active. Recall (from Section 4.2), this F0 attribute is the flag $FN[\phi_j, \omega_j[\rho], \rho]$ that indicates whether the track is part of the foreground harmonic signal or not, in time frame ρ . Here, $\rho \in \{\eta T', \dots, (\eta + 1)T' - 1\}$, T' is the number of time frames in short-time segment η , and $\omega_j[\rho]$ is track j 's frequency band in time frame ρ .

Given this measurement vector, dissimilarity is computed as follows. First, since we do not want to cluster tracks from different arrival angles, if two tracks (j_1 and j_2) have different arrival angles, their dissimilarity is set to a very large number. Otherwise, their dissimilarity is dependent on the difference in their F0 attributes in the time frames in which they are both active

$$d_{j_1, j_2} = \frac{\sum_{\rho=\eta T'}^{(\eta+1)T'-1} \mathcal{D} \cdot w_{j_1, j_2}[\rho]}{\sum_{\rho=\eta T'}^{(\eta+1)T'-1} w_{j_1, j_2}[\rho]}, \quad (30)$$

where \mathcal{D} denotes $|FN_{j_1}[\phi_{j_1}, \omega_{j_1}[\rho], \rho] - FN_{j_2}[\phi_{j_2}, \omega_{j_2}[\rho], \rho]|$ and $w_{j_1, j_2}[\rho]$ is a flag indicating whether tracks j_1 and j_2 are both active in time frame ρ , or not:

$$w_{j_1, j_2}[\rho] = \begin{cases} 1, & \text{if tracks, } j_1 \text{ and } j_2, \\ & \text{are both active in time frame } \rho, \\ 0, & \text{otherwise.} \end{cases} \quad (31)$$

If there are no time frames in which the pair of tracks are both active, it is not possible to compute their dissimilarity.

In this case, d_{j_1, j_2} is set to a neutral value such that their (dis)similarity will not be a factor in the clustering. Since the maximum dissimilarity between tracks is 1 and the minimum is 0, the neutral value is 1/2. For such a pair of tracks to be clustered together, they must each be close to the same set of other tracks. Otherwise, they will be assigned to different clusters.

Now that we have a performance measure (maximum intra-cluster dissimilarity), how should we use it to select a partition? It may seem reasonable to select the one that optimizes (minimizes) the performance measure. This selection criterion is no good though; it selects a partition in which each measurement is isolated in a separate cluster. A popular strategy used in hierarchical clustering is to pick a partition based on *changes* in the performance measure, rather than on the performance measure itself [37, 38, 40]. For CASA-EB, we are using such a selection criterion. Specifically, in keeping with the nature of our data (which contains a few, loosely connected clusters), we have chosen the following selection criterion. Starting with the minimum number of clusters, we select the first partition (the one with the smallest number of clusters, k) for which there is a significant change in performance from the previous partition (with $(k - 1)$ clusters).

4.3.3. Linking Short-Time Signal Estimate Groups. This is the final grouping step. In the previous steps, we have generated short-time estimates of the separated source signals (clusters of short-time signal component tracks). In this step, these short-time signal estimates will be linked together to form full-duration signal estimates. This is a data association problem. The short-time signal estimates in each time segment η must be associated with the previously established signal estimates through time segment $(\eta - 1)$. For an illustration, see Figure 5. To make this association, we rely on the fact that signals usually contain some long signal component tracks that continue across multiple time segments. Thus, these long tracks can be used to associate short-time signal estimates across segments. The idea is that a signal estimate's signal component tracks in time segment $(\eta - 1)$ will continue to be in the same signal in time segment η , and similarly, signal component tracks in a short-time signal estimate in time segment η will have their origins in the same signal in preceding time segments. The details of our processing are described next.

For this data association problem, we use the extended Munkres algorithm (as described in Section 4.3.1) with a cost function that is based on the idea described previously. Specifically, the cost function is the following:

$$\text{cost}_{g_k[\rho], c_\ell[\eta]} = \frac{\mathcal{A}_{k, \ell} - \mathcal{B}_{k, \ell}}{\mathcal{A}_{k, \ell}}, \quad (32)$$

where $g_k[\rho]$ is the k th signal estimate through the $(\eta - 1)^{\text{st}}$ time segment (i.e., $\rho < \eta T'$), $c_\ell[\eta]$ is the ℓ^{th} short-time signal estimate in time segment η , $\mathcal{A}_{k, \ell}$ is the power in the union of all their frequency component tracks,

$$\mathcal{A}_{k, \ell} = \sum_{j \in \{g_k[\rho] \cup c_\ell[\eta]\}} \mathcal{P}_j, \quad (33)$$

\mathcal{P}_j is the power in track j (defined below), $\mathcal{B}_{k,\ell}$ is the power in all the frequency component tracks that are in both $g_k[\rho]$ and $c_\ell[\eta]$,

$$\mathcal{B}_{k,\ell} = \sum_{j \in \{g_k[\rho] \cap c_\ell[\eta]\}} \mathcal{P}_j, \quad (34)$$

and \mathcal{P}_j is computed by summing all the power spectral density components along the length of track j ,

$$\mathcal{P}_j = \sum_{\rho=j^{\text{start}}}^{\min((\eta+1)T'-1, j^{\text{stop}})} P[\phi_j, \omega_j[\rho], \rho]. \quad (35)$$

This cost function takes on values in the range of 0 to 1. The cost is 0 when all the tracks in cluster $c_\ell[\eta]$ that have their beginning in an earlier time sequence are also in cluster track $g_k[\rho]$, and vice versa. The cost is 1 when $c_\ell[\eta]$ and $g_k[\rho]$ do not share any of the same signal component tracks.

Finally, notice that this cost function does not treat all tracks equally; it gives more weight to longer and more powerful tracks. To see this, consider two clusters: $c_{\ell_1}[\eta]$ and $c_{\ell_2}[\eta]$ that each contains one shared track with $g_k[\rho]$. Let the shared track in $c_{\ell_1}[\eta]$ be long and have high power, and let the shared track in $c_{\ell_2}[\eta]$ be short and have low power. Then, $\mathcal{B}_{k,1}$ will be larger than $\mathcal{B}_{k,2}$, and thus $\text{cost}_{k,1}[\eta] < \text{cost}_{k,2}[\eta]$. Although both $c_{\ell_1}[\eta]$ and $c_{\ell_2}[\eta]$ have one continuing track segment from $g_k[\rho]$, the one with the longer, stronger shared track is grouped with it. In this way, the cost function favors signal estimates that keep important spectral structures intact.

5. CASA-EB Waveform Synthesis

The preceding processing steps complete the separation of the mixture signal into the single-source signal estimates $g_k[\rho]$. However, the signal estimates are still simply groups of signal components. In some applications, it may be desirable to have waveforms (e.g., to listen to the signal estimates, or to process them further in another signal processing application such as an automatic speech recognizer).

Waveform reconstruction is done in two steps. First, in time frame ρ , a short-time waveform is generated for each group, $g_k[\rho]$, that is active (i.e., nonempty) in the time frame. Then, full-length waveforms are generated from these by connecting them together across time frames. The implementation details are described next.

In the first step, for each currently active group, its short-time waveform is generated by summing its short-time narrowband waveforms $Y[\phi, \omega, n]$ over frequency:

$$\hat{x}_k^\rho[n] = \sum_{\substack{\phi, \omega \text{ s.t.} \\ P[\phi, \omega, \rho] \in g_k[\rho]}} Y[\phi, \omega, n], \quad (36)$$

where $n \in \{\rho - (T-1)/2 \dots \rho + (T-1)/2\}$. In the second step, these short-time waveforms are connected together across time into full-length waveforms by the standard overlap-add algorithm,

$$\hat{x}_k[n] = \sum_{\rho} \sum_{r=-(T-1)/2}^{(T-1)/2} v[r] \cdot \hat{x}_k^\rho[r], \quad (37)$$

where we have chosen to use a Hanning window, $v[\cdot]$, because of its low sidelobes and reasonably narrow main lobe width.

6. Experimental Results

For a sound source separation method, such as CASA-EB, it is important that it both separate mixture signals completely and that the separated signals have good quality. The experiments described in Section 6.2 assess CASA-EB's ability to do these. Specifically, they test our hypothesis that combining monaural CASA and beamforming, as in CASA-EB, provides more complete signal separation than either CASA or beamforming alone, and that the separated signals have low spectral distortion.

Before conducting these experiments, a preliminary experiment is performed. In particular, to make the comparison of CASA-EB to monaural CASA meaningful, first we need to verify that the performance of the monaural CASA in CASA-EB is inline with other previously published CASA methods. Since it is not practical to compare our CASA technique to every previously proposed technique (there are too many and there is no generally accepted standard), we selected a representative technique for comparison—that of van der Kouwe, Wang and Brown [1]. We chose their method for three reasons. First, a clear comparison can be made since their testing method is easily reproducible with readily-available test data. Second, comparison to their technique can provide a good check for ours since the two methods are similar; they both use the same grouping cue and a similar temporal analysis filter, $h_\omega[n]$. The main differences are that our technique contains spatial filtering (which theirs does not), and it uses tracking/clustering for grouping (while their technique uses neural networks for grouping). Finally, they (Roman, Wang and Brown) have also done work separating signals based on location cues (binaural CASA) [4], and some preliminary work combining source attributes (F0 attribute) and location attributes (binaural CASA cues)—see [13] by Wrigley and Brown.

6.1. Preliminary Signal Separation Experiments: Monaural CASA. To compare our monaural CASA technique to that of [1], we tested our technique using the same test data and performance measure as they used to test theirs. In this way, our results can be compared directly to their published results. The test data consists of 10 mixture signals from the data set of [8]. Each mixture consists of a speech signal (v8) and one of ten interference signals (see Table 2).

The performance measure is the SIR gain (signal to interference ratio) (this SIR gain is the same as the SNR gain in [1]; we prefer the name SIR gain since it is a more accurate description of what is computed), that is, the difference between the SIRs before and after signal separation:

$$\Delta \text{SIR} = \text{SIR}_{\text{after}} - \text{SIR}_{\text{before}}, \quad (38)$$

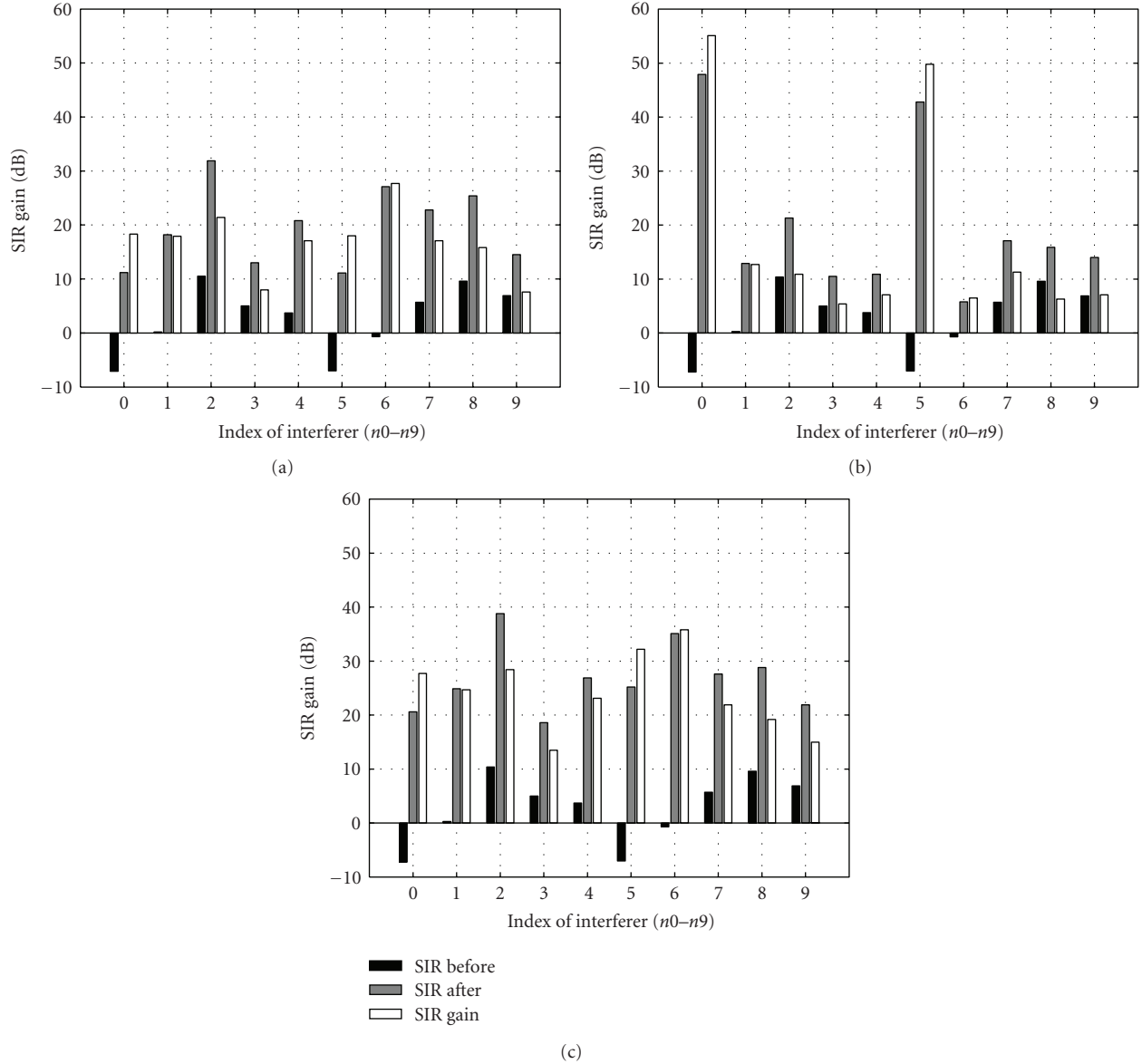


FIGURE 6: SIR gains of $\widehat{v8}$ estimates from beamforming (a), CASA (b) and CASA-EB (c). The horizontal axes in the graphs specify the test mixture by the index of the interferer. The three bars shown for each indicate the SIR of $v8$ in the mixture (black), the SIR of the separated $\widehat{v8}$ (gray), and the SIR gain (white). To summarize these results, the mean SIR gains are 16.9 dB (for beamforming on mixtures with $\pi/2$ radians of source separation), 17.2 dB (for monaural CASA) or 8.4 dB (for monaural CASA without the n_0 and n_5 results), and 24.2 dB (for CASA-EB on mixtures with $\pi/2$ radians of source separation).

where

$$\begin{aligned} \text{SIR}_{\text{after}} &= 10 \log \left(\frac{P_{v8 \in \widehat{v8}}}{P_{nx \in \widehat{v8}}} \right), \\ \text{SIR}_{\text{before}} &= 10 \log \left(\frac{P_{v8 \in v8+nx}}{P_{nx \in v8+nx}} \right). \end{aligned} \quad (39)$$

Here, $P_{v8 \in \widehat{v8}}$ is the power (or amount) of the speech signal ($v8$) in its estimate (i.e., the separated signal $\widehat{v8}$), $P_{nx \in \widehat{v8}}$ is the power (or amount) of interference (nx) in $\widehat{v8}$, $P_{v8 \in v8+nx}$ is the power of $v8$ in the test mixture ($v8 + nx$), and $P_{nx \in v8+nx}$ is the power of nx in ($v8 + nx$), where nx is one of $\{n_0, n_1, \dots, n_9\}$.

SIR is a useful measure in the sense that it tells us how well interference has been removed by signal separation—the higher the SIR, the more interference-free the separated signal.

In a typical experiment, we ran our monaural CASA algorithm on each of the ten mixture signals, and the resultant SIRs (before and after) along with the SIR gains are shown in the upper panel of Figure 6. Specifically, this figure contains 10 groups of lines (black, gray, and white), indexed from 1 to 10 on a horizontal axis, one for each mixture signal in the test data. For example, the results at index 5 are for mixture ($v8 + n_5$). In each group (i.e., for each mixture

TABLE 2: Voiced speech signal $v8$ and the interference signals ($n0$ – $n9$) from Cooke’s 100 mixtures [8].

ID	Description	Characterization
$v8$	Why were you all weary?	
$n0$	1 kHz tone	Narrowband, continuous, structured
$n1$	White noise	Wideband, continuous, unstructured
$n2$	Series of brief noise bursts	Wideband, interrupted, unstructured
$n3$	Teaching laboratory noise	Wideband, continuous, partly structured
$n4$	New wave music	Wideband, continuous, structured
$n5$	FM signal (siren)	Locally narrowband, continuous, structured
$n6$	Telephone ring	Wideband, interrupted, structured
$n7$	Female TIMIT utterance	Wideband, continuous, structured
$n8$	Male TIMIT utterance	Wideband, continuous, structured
$n9$	Female utterance	Wideband, continuous, structured

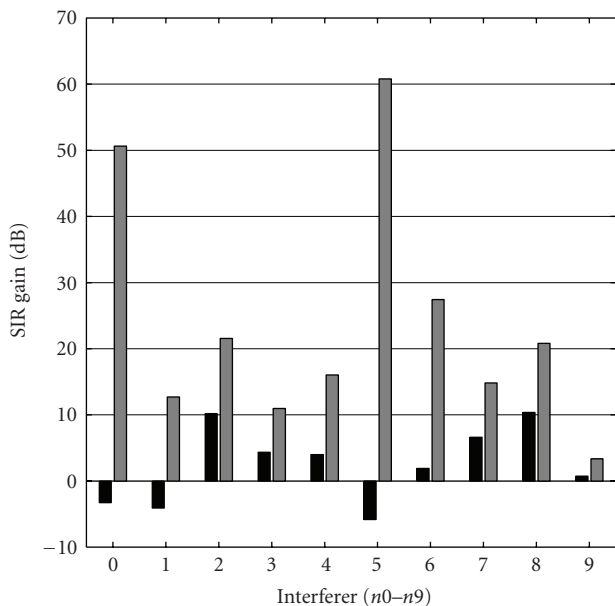


FIGURE 7: SIR gains of $\hat{v}8$ estimates, from Wang, Brown, and van der Kouwe et al.’s monaural CASA. The horizontal axis specifies the test mixture by its interferer. The two lines shown for each indicate the SIR of $v8$ in the mixture (black), and the SIR of the separated $v8$ (gray).

signal), the height of the black line is the SIR of the original mixture signal, the height of the gray line is the SIR of the signal estimate after CASA separation ($\hat{v}8$), and the height of the white line is their difference, that is, the SIR gain achieved by CASA separation.

For comparison’s sake, Wang, Brown, and van der Kouwe’s results on the mixture signals of Table 2 are shown in Figure 7, organized in the same way as in Figure 6. From these figures, we can see that the performance of our CASA technique is similar to theirs. The main differences are from the $n6$ and $n9$ mixture signals; their method performed better for $n6$, CASA-EB for $n9$. Thus, our CASA technique can be considered comparable to this published CASA technique.

6.2. Main Signal Separation Experiments: CASA-EB. To test our hypothesis that the combined approach, CASA-EB, separates mixture signals more completely than the individual techniques (CASA and beamforming) used alone, we ran all three on mixture signals of the same speech ($v8$) and interference ($n0$ – $n9$) signals and compared the resulting SIR gains. To assess the quality of the separated signals, we also computed their LPC cepstral distortions.

For monaural CASA, the test data was exactly the same as that used in Section 6.1. For beamforming and CASA-EB, however, array data was simulated from the speech and interference signals, and the mixture signals were made from these. We chose to simulate the array data rather than to record the speech-interference mixture signals through a microphone array because simulation provides data that is specific to the room it is recorded in. The disadvantage of this approach is that the simulated array data may not be entirely realistic (e.g., it does not include room reverberations). For the array data simulation, we used a method described in [31] on a uniform linear array of 30 microphones. Each of the ten mixture signals, as measured at the array, is composed of the speech ($v8$) and one interference signal ($n0$ – $n9$), where $v8$ ’s arrival angle is $+\pi/4$ and the interference signal’s is $-\pi/4$ radians from broadside.

6.2.1. Signal Separation Completeness. The SIR gains of the separated signals from beamforming, monaural CASA and CASA-EB are shown in Figures 6(a), 6(b), and 6(c), respectively. The results show a definite advantage for CASA-EB over either beamforming or monaural CASA alone for all but two exceptions (the narrowband interferers, $n0$ and $n5$) addressed below. Specifically, the mean SIR gains for beamforming, monaural CASA and CASA-EB are 16.9, 17.2, and 24.2 dB, respectively. Note that the mean SIR gain for monaural CASA would be 8.4 if you leave out the results from the mixtures made with the narrowband interferers, $n0$ and $n5$.

Now, we consider the two exceptions, that is, the mixtures ($v8 + n0$) and ($v8 + n5$) for which CASA-alone achieves near-perfect performance, and CASA-EB does not. Why does CASA remove $n0$ and $n5$ so well? To find an answer, we first notice that unlike other interferers, $n0$ and $n5$

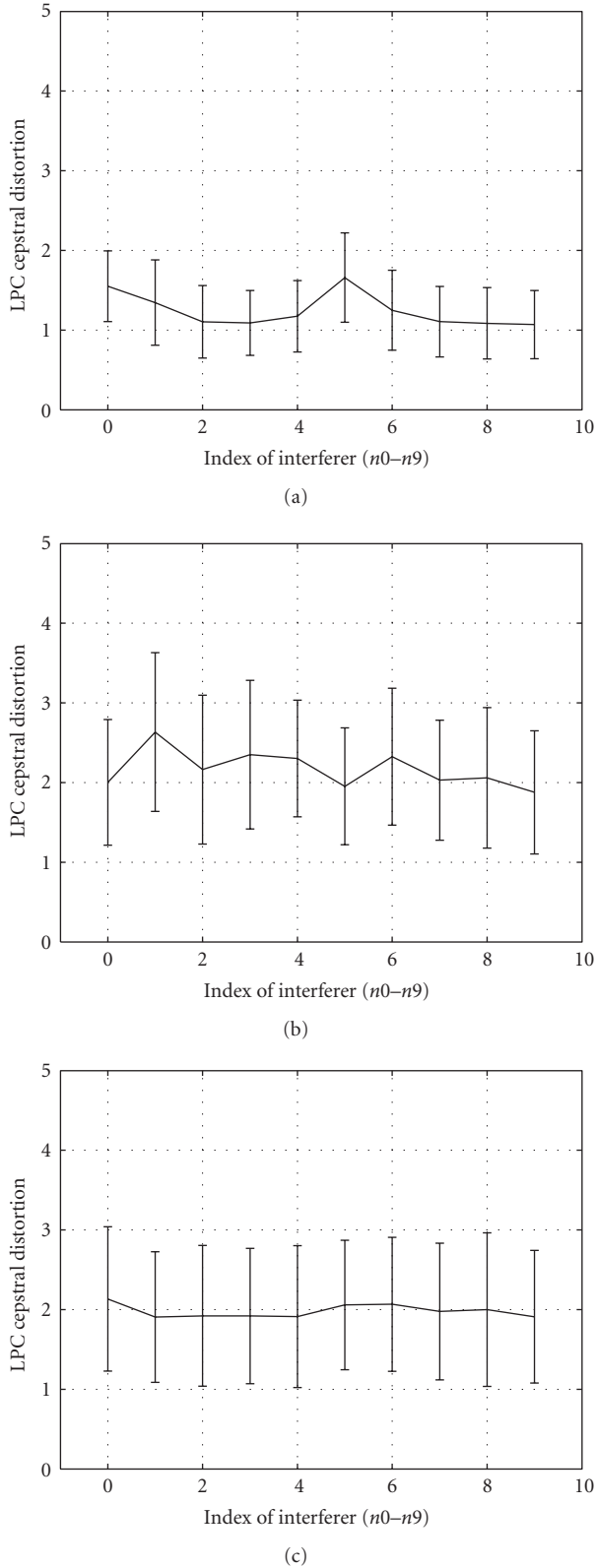


FIGURE 8: LPC cepstral distortions of $\widehat{v8}$ estimates from beamforming (a), CASA (b), and CASA-EB (c). As in Figures 6 and 7, the horizontal axes in the graphs specify the test mixture by the index of the interferer. The value plotted is the mean LPC cepstral distortion over the duration of the input mixture, $v8 + nx$, $nx \in \{n0, n1, \dots, n9\}$; the error bars show the standard deviations.

are narrowband and, in any short period of time, each has its power concentrated in a single frequency or a very narrow frequency band. Now, recall that our CASA approach separates a signal from interference by grouping harmonic signal components of a common fundamental, and rejecting other signal components. It does this by first passing the signal-interference mixture through a filter bank (the $h_\omega[n]$ defined in Section 3), that is, decomposing it into a set of subband signals. Then, the autocorrelation for each subband is computed, forming an autocorrelogram (see Figure 4(b)), and a harmonic group (a fundamental frequency and its overtones) is identified (as described in Section 4.2). After such harmonics are identified, the remaining signal components (interferers) are rejected.

When an interferer is narrowband (such as $n0$ and $n5$), it is almost certain that it will be contained entirely in a single subband. Furthermore, if the interferer has a lot of power (as in $v8 + n0$ and $v8 + n5$), it is going to affect the location of the autocorrelogram peak for that subband. Either the peak in the subband will correspond to the period of the interferer, if it is strong relative to the other signal content in the subband, or the peak will at least be pulled towards the interferer. When we use CASA, this will cause the subband to be rejected from the signal estimate, and as a result the interferer will be completely rejected. This is why CASA works so well in rejecting narrowband interferers.

When CASA-EB is used, the CASA operation is preceded by spatial filtering (beamforming). When the interferer and the signal come from different directions (as is the case in $v8 + n0$ and $v8 + n5$), this has the affect of reducing the power of the interferer in the subband that it is in. As a result, the autocorrelogram peak in that subband will be much less affected by the interferer compared to the CASA alone case, and as a result, the subband may not be rejected in the signal reconstruction, leading to a smaller SIR improvement than when CASA is used alone. However, we would like to point out that CASA-EB's performance in this case (on mixtures with narrowband interferers), although not as good as CASA-alone's dramatic performance, is still quite decent thanks to the spatial filtering that reduced the interferers' power.

6.2.2. Perceptual Quality of Separated Signals. The mean LPC cepstral distortions of the separated signals ($\widehat{v8}$) from beamforming, monaural CASA, and CASA-EB are shown in Figures 8(a), 8(b), and 8(c), respectively. Here, LPC cepstral distortion is computed as:

$$d[r] = \sqrt{\frac{1}{F+1} \cdot \sum_{f=0}^F (\ln(P_{v8}[f]) - \ln(P_{\widehat{v8}}[f]))^2}, \quad (40)$$

where $r = n/T_d$ is the time index, $T_d = 160$ is the length of signal used to compute $d[r]$, $P_{v8}[f]$ is the LPC power spectral component of $v8$ at frequency f (computed by the Yule-Walker method), and $F = 60$ corresponds to frequency $fs/2$.

The results show that beamforming produces low distortion (1.24 dB averaged over the duration of the separated

signal \widehat{v}_8 and over all 10 test mixtures), CASA introduces somewhat higher distortion (2.17 dB), and CASA-EB is similar to monaural CASA (1.98 dB). The fact that beamforming produces lower distortion than CASA may be because distortion in beamforming comes primarily from incomplete removal of interferers and noise, while in CASA, additional distortion comes from the removal of target signal components when the target signal has frequency content in bands that are dominated by interferer(s). Thus, beamforming generally passes the entire target signal with some residual interference (generating low distortion), while CASA produces signal estimates that can also be missing pieces of the target signal (producing more distortion).

6.2.3. Summary. In summary, CASA-EB separates mixture signals more completely than either individual method alone and produces separated signals with rather low spectral distortion (~ 2 dB LPC cepstral distortion). Lower spectral distortion can be had by using beamforming alone, however, beamforming generally provides less signal separation than CASA-EB and cannot separate signals from close arrival angles.

7. Conclusion

In this paper, we proposed a novel approach to acoustic signal separation. Compared to most previously proposed approaches which use either location or source attributes alone, this approach, called CASA-EB, exploits both location and source attributes by combining beamforming and auditory scene analysis. Another novel aspect of our work is in the signal component grouping step, which uses clustering and Kalman filtering to group signal components over time and frequency.

Experimental results have demonstrated the efficacy of our proposed approach; overall, CASA-EB provides better signal separation performance than beamforming or CASA alone, and while the quality of the separated signals suffers some degradation, their spectral distortions are rather low (~ 2 dB LPC cepstral distortion). Although beyond the scope of this current work, to demonstrate the advantage of *combining* location and source attributes for acoustic signal separation, further performance improvements may be achieved by tuning CASA-EB's parts. For example, using a higher resolution beamformer may allow CASA-EB to produce separated signals with lower residual interference from neighboring arrival angles, and using a larger set of source attributes could improve performance for harmonic target signals and accommodate target signals with nonharmonic structures.

References

[1] A. J. W. van der Kouwe, D. Wang, and G. J. Brown, "A comparison of auditory and blind separation techniques for speech segregation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 189–195, 2001.

[2] P. N. Denbigh and J. Zhao, "Pitch extraction and separation of overlapping speech," *Speech Communication*, vol. 11, no. 2-3, pp. 119–125, 1992.

[3] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3, pp. 209–222, 1999.

[4] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[5] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, New York, NY, USA, 2001.

[6] L. Drake, A. K. Katsaggelos, J. C. Rutledge, and J. Zhang, "Sound source separation via computational auditory scene analysis-enhanced beamforming," in *Proceedings of the 2nd IEEE Sensor Array and Multichannel Signal Processing Workshop*, Rosslyn, Va, USA, August 2002.

[7] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, no. 3-4, pp. 141–177, 2001.

[8] M. Cooke, *Modelling auditory processing and organisation*, Ph.D. dissertation, The University of Sheffield, Sheffield, UK, 1991.

[9] G. Brown, *Computational auditory scene analysis: a representational approach*, Ph.D. dissertation, The University of Sheffield, Sheffield, UK, 1992.

[10] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. dissertation, MIT, Cambridge, Mass, USA, April 1996.

[11] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[12] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[13] S. N. Wrigley and G. J. Brown, "Recurrent timing neural networks for joint F0-localisation based speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 1, pp. 157–160, Honolulu, Hawaii, USA, April 2007.

[14] J. Krolik, "Focused wide-band array processing for spatial spectral estimation," in *Advances in Spectrum Analysis and Array Processing*, S. Haykin, Ed., vol. 2 of *Prentice Hall Signal Processing Series and Prentice Hall Advanced Reference Series*, chapter 6, pp. 221–261, Prentice-Hall, Englewood-Cliffs, NJ, USA, 1991.

[15] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[16] D. N. Swingler and J. Krolik, "Source location bias in the coherently focused high-resolution broad-band beamformer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 1, pp. 143–145, 1989.

[17] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[18] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281, 1973.

- [19] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [20] G. Schwartz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.
- [21] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.
- [22] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *The Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, 1976.
- [23] U. Baumann, "Pitch and onset as cues for segregation of musical voices," in *Proceedings of the 2nd International Conference on Music Perception and Cognition*, February 1992.
- [24] G. Brown and M. Cooke, "Perceptual grouping of musical sounds: a computational model," *The Journal of New Music Research*, vol. 23, no. 2, pp. 107–132, 1994.
- [25] Y. Gu, "A robust pseudo perceptual pitch estimator," in *Proceedings of the 2nd European Conference on Speech Communication and Technology (EUROSPEECH '91)*, pp. 453–456, 1991.
- [26] M. Weintraub, *A theory and computational model of auditory sound separation*, Ph.D. dissertation, Stanford University, Stanford, UK, 1985.
- [27] R. Gardner, "An algorithm for separating simultaneous vowels," *British Journal of Audiology*, vol. 23, pp. 170–171, 1989.
- [28] M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90)*, vol. 1, pp. 357–360, Albuquerque, NM, USA, April 1990.
- [29] P. F. Assmann and Q. Summerfield, "Modeling the perception of concurrent vowels: vowels with different fundamental frequencies," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 680–697, 1990.
- [30] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery—I: pitch identification," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866–2882, 1991.
- [31] L. Drake, *Sound source separation via computational auditory scene analysis (casa)-enhanced beamforming*, Ph.D. dissertation, Northwestern University, December 2001.
- [32] U. Baumann, "A procedure for identification and segregation of multiple auditory objects," in *Proceedings of the NATO Advanced Study Institute on Computational Hearing*, S. Greenberg and M. Slaney, Eds., pp. 211–215, International Computer Science Institute, Berkeley, Calif, USA, 1998.
- [33] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, vol. 179 of *Mathematics in Science and Engineering*, Academic Press, New York, NY, USA, 1988.
- [34] F. Burgeois and J. C. Lassalle, "Extension of the Munkres algorithm for the assignment problems to rectangular matrices," *Communications of the ACM*, vol. 14, no. 12, pp. 802–804, 1971.
- [35] S. S. Blackman, *Multiple-Target Tracking with Radar Applications*, Artech House, Boston, Mass, USA, 1986.
- [36] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, Reading, Mass, USA, 1974.
- [37] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, NY, USA, 1990.
- [38] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, NY, USA, 1999.
- [39] B. D. Ripley and N. Hjort, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, UK, 1995.
- [40] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.
- [41] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.
- [42] R. Gentleman, R. Ihaka, and R Core Team, "R version 0.63.1," December 1998, a statistical computation and graphics system. Re-implementation of the S language, using Scheme semantics, <http://www.stat.auckland.ac.nz/r/r.html>.
- [43] R. A. Becker, J. M. Chambers, and A. R. Wilks, *The New S Language*, Chapman & Hall, London, UK, 1988.
- [44] J. M. Chambers and T. J. Hastie, *Statistical Models in S*, Chapman & Hall, London, UK, 1992.