Comparison of a target-equalization-cancellation approach and a localization approach to source separation

Jing Mi, Matti Groll, and H. Steven Colburn^{a)}

Hearing Research Center, Department of Biomedical Engineering, Boston University, 44 Cummington Mall, Boston, Massachusetts 02215, USA

(Received 11 May 2017; revised 19 October 2017; accepted 19 October 2017; published online 14 November 2017)

Interaural differences are important for listeners to be able to maintain focus on a sound source of interest in the presence of multiple sources. Because interaural differences are sound localization cues, most binaural-cue-based source separation algorithms attempt separation by localizing each time-frequency (T-F) unit to one of the possible source directions using interaural differences. By assembling T-F units that are assigned to one direction, the sound stream from that direction is enhanced. In this paper, a different type of binaural cue for source-separation purposes is proposed. For each T-F unit, the target-direction signal is cancelled by applying the equalization-cancellation (EC) operation to cancel the signal from the target direction; then, the dominance of the target in each T-F unit is determined by the effectiveness of the cancellation. Specifically, the energy change from cancellation is used as the criterion for target dominance for each T-F unit. Source-separation performance using the target-EC cue is compared with performance using localization cues. With simulated multi-talker and diffuse-babble interferers, the algorithm based on target-EC cues yields better source-separation performance than the algorithm based on localization cues, both in direct comparison with the ideal binary mask and in measured speech intelligibility for the separated target streams. © 2017 Acoustical Society of America. <https://doi.org/10.1121/1.5009763>

[LRB]

Pages: 2933–2941

I. INTRODUCTION

Normal-hearing listeners are adept at focusing on the sound source of interest when in the presence of other sound sources; this phenomenon is called the cocktail party effect (Cherry, 1953). In contrast, hearing-impaired listeners often struggle in challenging listening environments (e.g., those that are noisy and/or that have multiple simultaneous sources). Automatic speech recognition systems often yield compromised performance in such environments as well. Therefore, ongoing efforts in the computational acoustic scene analysis (CASA) domain have been devoted to building hearing-system-inspired cocktail-party processors that can separate out the source of interest from the background noise (cf. Wang and Brown, 2006).

The overall ability of human listeners to solve the cocktail party problem relies on many acoustic cues, including pitch, common onsets/offsets and binaural cues (cf. Bronkhorst, 2015). Binaural cues are among the most useful and well-modeled cues. In fact, binaural cues and their processing have been studied in the psychoacoustic domain for decades and various models have been suggested to explain binaural interaction mechanisms, including the coincidence-detection model (Jeffress, 1948), the equalization-cancellation (EC) model (Durlach, 1963) and the auditory-nerve-based model (Colburn, 1977, 1973). Among these models, the EC model is capable of quantitatively predicting binaural hearing benefits in multiple listening conditions, including tone-in-noise detection (Durlach, 1963), speech-in-noise intelligibility (Zurek, 1992)

and speech-on-speech interference (Wan *et al.*, 2010). However, when it comes to source separation algorithms in the CASA domain, the EC model has not been widely applied. One possible reason for this is that the EC model was formulated to cancel a single noise source at a time, while in many situations of interest, there are often multiple interfering sources.

Most binaural source separation algorithms in CASA use interaural time difference (ITD), interaural level difference (ILD), and interaural cross-correlation coefficient (IACC) as binaural features (Liu and Zhang, 2014; Mandel *et al.*, 2010; Yilmaz and Rickard, 2004; Roman *et al.*, 2003). In those algorithms, each time-frequency (T-F) unit is localized to one of the possible source directions based on ITD and ILD. By assembling all the T-F units that are allocated to one source's direction, the sound stream made by that source is separated from the mixture. The IACC is defined as the peak value of the cross-correlation function (Yin *et al.*, 1987) and is sometimes used to eliminate those T-F units that have low cross-correlation between the waveforms at the two ears (Alinaghi *et al.*, 2013). The performance of such localization-based algorithms mainly depends on two factors. The first factor is the accurate extraction of ITD, ILD, and IACC from short time slices. ITD can be calculated using the time-domain cross-correlation function between the signals at the two ears; however, for high-frequency channels, the cross-correlation functions are multi-modal within the physiologically reasonable ITD range, thus there will be confusion in determining the fine-structure ITD. To resolve the confusion caused by multi-modality, top-down integration of across-frequency information is necessary.

^{a)}Electronic mail: colburn@bu.edu

This integration is acceptable for sound localization algorithms, but cannot be applied to source separation algorithms that make separate decisions for individual T-F units. Previous source-separation algorithms either ignore high-frequency ITD (Yilmaz and Rickard, 2004), use top-down information to alleviate the confusion (Mandel *et al.*, 2010) or use envelope ITD for high-frequency channels (Roman *et al.*, 2003). The second factor is how the ITD, ILD, and IACC cues are combined to improve localization of individual T-F units. Joint ITD-ILD distributions have been constructed by either supervised learning for each spatial configuration (Roman *et al.*, 2003) or by assuming conditional-independency of ITD and ILD distributions given their respective direction-specific values (Mandel *et al.*, 2010). For the former method, it is laborious to acquire enough training samples for the many possible spatial configurations in realistic listening environments, especially considering the uncertainty about the number of sources and the uncertainty of each source's location. For the latter method, it is not always reasonable to assume that ITD and ILD distributions are conditionally independent. In reality, the interfering source is often binaurally correlated, for example, interfering speech from a person nearby. Therefore, after conditioning on intrinsic direction-specific ITD and ILD, the ITD and ILD distributions may still show correlation, as they are affected by the same interfering source.

In this study, we developed a previously proposed binaural processing scheme (Mi and Colburn, 2016) that can be used for source separation purposes and that does not involve the localization of individual T-F unit. The scheme is inspired by the EC model, together with the target-cancellation idea in the work of Roman *et al.* (2006). In the Roman study, a 10-s white noise in the target-direction was used to train an adaptive filter that could be applied to the left or right target signal to make it identical to the signal at the other ear so that the difference between these outputs would result in cancellation of target-direction signals. The adaptive filter could later be applied to the sound mixture to cancel the signal from the target direction in the presence of interferers. In the current study, instead of learning a cancellation filter with a sample of target-alone stimuli, the signal components are equalized using HRTFs from the known target direction and then cancelled. To be more specific, this target-EC process can be considered as a steerable-null filter but with pre-stored filter coefficients, and the method applied here can be described as using a simple target-EC process as an alternative to the adaptive filter in Roman's study. In both methods, after cancelling the target, the output-to-input ratio (OIR) of cancellation process is calculated for individual T-F units. Based on the OIR, the target source dominance is estimated for each T-F unit. The basic rationale is based on the observation that, if a T-F unit has a large energy decrease after the target EC, which corresponds to a small OIR, then the T-F unit is likely to be dominated by the target; if the energy is not significantly reduced after the target-EC processing, which corresponds to a large OIR, then the T-F unit is likely to be dominated by noise sources. In other words, based on the OIR value, a binary mask of target dominance for each T-F

unit is estimated; then the mask is used to manipulate the original sound mixture to pass the target-dominant T-F units and to block the interferer-dominant T-F units. This proposed binaural processing scheme was described and applied to model human performance in multi-talker conditions [using data from Marrone *et al.* (2008)] in Mi and Colburn (2016). A similar scheme of applying this target-EC model to source separation has been reported in Deshpande and Braasch (2016). In the current paper, instead of focusing on explaining available human data with the proposed target-EC binaural processing scheme, we focus on the engineering problem, that is, which group of binaural cues—localization cues (ITD, ILD, and IACC) or target-EC cues (OIR)—is better for source-separation purposes.

The rest of this paper is organized into four sections. Section II introduces the basic structure under which the two groups of cues are compared and then presents the details of the cue extractions. Section III describes the three simulated listening conditions under which the comparisons are done and the two ways of assessing the algorithms' performance. Section IV compares the performance of the classifiers that use the two different groups of binaural cues (localization cues and target-EC cues). Finally, Sec. V summarizes the findings and discusses the relevance of the target-EC model in both engineering solutions and psychoacoustic modeling.

II. MODEL FRAMEWORK

Most binaural source separation algorithms in the CASA domain follow the framework shown in Fig. 1, so here the comparison between the two groups of binaural cues is presented under the same framework. There are four steps in this framework: (1) peripheral processing, (2) extraction of binaural cues, (3) estimation of a binary mask based on the pattern of the binaural-cue distribution, and (4) resynthesis of the target stream using the estimated binary mask. Because the main aim of this study is to evaluate which group of binaural cues is better for source separation purposes, we construct source separation algorithms based on different combinations of binaural cues. For all algorithms considered, the direction of the target source and the direction-related interaural parameters are assumed to be known. The interaural parameters are calculated from the head-related transfer functions. In the following paragraphs, each of the four steps is specified in detail.

The peripheral processing step (step 1) is simulated with 32 gammatone filters. The center frequencies of the filters range from 80 Hz to 6 kHz; they are spaced uniformly on a logarithmic scale. The filtered signals are divided into 20-ms time slices using a Hamming window, with 50% overlap between successive windows.

In the binaural-cue extraction step (step 2), we extract two groups of binaural cues. The first group is based on interaural difference cues, specifically, ITD distance, ILD distance and IACC, which are specified in detail in the next paragraph. Because ITD and ILD are also the cues for sound localization, this group is referred to as "localization cues." The second group of binaural cues only includes the OIR of the target-EC operation. This is referred to as the target-EC cue and is explained in detail below.

FIG. 1. General source-separation framework. The binaural mixtures are either speech mixtures from multiple talkers (as shown here) or speech masked by diffuse background babble. The binaural mixtures first pass through filter banks and then each filtered signal is divided into short time slices. For each T-F unit, two groups of binaural cues (localization cues and EC cue) are extracted. The distributions of binaural cues are then analyzed to estimate a binary mask. Last, the estimated binary mask is used to manipulate the original binaural mixture to obtain the separated target stream.

For the localization cues, in frequency channels below 1200 Hz, the fine-structure ITD is calculated; in frequency channels above 1200 Hz, the envelope ITD is calculated. Both ITDs are calculated by identifying the peak location of the cross-correlation function of the left and right signals or signal envelopes of the corresponding frequency channels. The ILD is the energy ratio between two ears, expressed in decibel units. The distance (absolute difference) between the calculated ITD and target-direction ITD, and the distance between the calculated ILD and target-direction ILD are used, instead of directly using the calculated short-time ITD and ILD. The reason for this “distance” choice is that our goal is to formulate the problem as a binary classification problem and to use a simple classifier to solve the problem. Therefore, we only want to differentiate each T-F unit into either a target-dominant or a noise-dominant unit, and we do not care which noise source is dominant. For example, suppose there are two spatially separated noise sources located at different sides relative to the target direction. The ITD, ILD distributions of T-F units dominated by the different noise sources will be on different sides of the ITD, ILD distributions of the target-dominant T-F units. Thus, a linear classifier would have difficulty classifying each T-F unit into one of only two possible classes (target-dominant or noise-dominant), as there are actually three populations. In contrast, with ITD-distance and ILD-distance, the target-dominant T-F units have ITD and ILD close to the intrinsic target-direction ITD and ILD, and so the target-dominant T-F units are close to the origin in the ITD-distance and ILD-distance plane. Similarly, the noise-dominant T-F units are far away from the origin. Thus, a linear binary classifier would be enough to separate the target-dominant units from the noise-dominant units.

For the target-EC cue, we first calculate the output of a T-F unit using the following equation:

$$Y_i(t) = L_i(t) - \alpha_i R_i(t - \tau_i). \quad (1)$$

In this equation, $L_i(t)$ and $R_i(t)$ represent the filtered left-ear and right-ear waveforms for the i th frequency channel before the target-EC processing; the variables τ_i and α_i stand for the intrinsic interaural time difference and amplitude

ratio between two ears for the target direction; and $Y_i(t)$ represents the i th channel output after EC processing. The output-to-input (OIR) ratio is defined as

$$OIR(i, j) = 10 \log_{10} \frac{\int |Y_{i,j}(t)|^2 dt}{0.5 \left(\int |L_{i,j}(t)|^2 dt + \int |R_{i,j}(t)|^2 dt \right)}. \quad (2)$$

In this equation, $Y_{i,j}(t)$ represents the target-cancelled output in the i th frequency channel and the j th time slice; and $L_{i,j}(t)$ and $R_{i,j}(t)$ represent the left and right filtered signals of that T-F unit (before the target-cancellation). As described above, the OIR is small whenever the target dominates the inputs. When the SNR is positive, the OIR is negatively correlated with short-time SNR, because the larger the SNR, the more completely is the signal cancelled. In contrast, when the SNR is negative, the OIR stays almost uniform for a large range of SNRs, because there is not much signal to be cancelled at negative SNR region [cf. Fig. 2 in [Mi and Colburn \(2016\)](#)].

In the pattern analysis step (step 3), several binary masks are estimated, with each mask based on either the localization cues or the target-EC cues as extracted in step 2. In the current study, the estimation is mainly done with the logistic regression (LR) classifier. For the training of the classifier, either the localization cues or the target-EC cue are provided as the features and the ideal binary mask (IBM) is provided as the labels. The IBM is a binary mask in that weights each T-F unit by either zero or unity, depending on the signal-to-noise ratio within that unit. The IBM is ideal in that it calculates the signal-to-noise ratio for each unit by assuming complete knowledge of all waveforms (target and interferers) within the unit. The IBM is commonly regarded as a reasonable goal for source separation algorithms ([Wang, 2005](#)). Separate logistic regression classifiers are trained for each frequency channel and for each listening condition. In real-life applications, it is unrealistic to have each classifier specifically trained for each listening condition; however, the goal here is to study which group of binaural cues is better for source separation purpose, so we primarily focus on the cue itself. Strategies for making the source separation algorithms more realistic are considered in Sec. V.

Also in step 3, in addition to the logistic regression classifiers, a support vector machine (SVM) classifier with a Gaussian kernel (Cortes and Vapnik, 1995) is used to estimate the mask with localization cues. The SVM classifier would improve the performance in cases for which the non-linear interaction between the localization cues has a large effect on the classification results. Basically, the SVM classifier with Gaussian kernel aims to find a non-linear hyperplane that separates two classes with maximized margins between the data points and the hyperplane. In the Gaussian-kernel SVM classifier, there is a kernel parameter that controls the smoothness of the hyperplane and a penalty parameter that controls the mis-classification tolerance of the classifier. The choice of these two parameters is critical to the classifier's performance; thus, a five-fold cross-validation using the training data is performed to choose functional parameters for the classifier. In the five-fold cross-validation, the training samples are partitioned into five equal-sized subsample groups. In each repetition, four subsample groups are used as training data and the remaining subsample group is used as validation data. The procedure is repeated five times and the best parameters from each repetition are averaged to yield the final kernel and penalty parameters. This tuning of parameters is not needed for the logistic regression classifiers, only for the SVM classifier.

The last step (step 4) reconstructs the target stream by applying the estimated binary mask to the original dichotic stimuli. This reconstruction is done using the COCHLEAGRAM analysis/synthesis toolbox (Wang and Brown, 2007) and a detailed description of the reconstruction process can be found in Wang and Brown (2006). This step provides binaural output waveforms for each of the binary masks as they are applied to the input.

III. EVALUATION

The source separation algorithms using different groups of binaural cues were evaluated in three simulated listening conditions: the anechoic multi-talker condition, the reverberant multi-talker condition, and the diffuse-babble condition. The anechoic head-related transfer functions used were taken from the CIPIC database (Algazi *et al.*, 2001) and the reverberant room response was simulated with an image-model-based (Allen and Berkley, 1979) room simulator.¹ The simulated room was of size $6\text{ m} \times 10\text{ m} \times 3\text{ m}$ and the sources were at a distance of 1.5 m from the listener. The uniform absorption coefficient of the wall was set as 0.147 and the resulting reverberation time was 0.7 s. In the anechoic and reverberant multi-talker conditions, a female target-talker was placed in the front, and two female masker-talkers were placed symmetrically at $\pm 60^\circ$ to the target-talker (as shown in Fig. 1). The coordinate response measure corpus (Bolía *et al.*, 2000) was used as the speech source for all talkers, mainly because many experiments studying spatial release from masking in multi-talker conditions used this corpus. In the diffuse-babble condition, the diffuse babble was simulated by combining 23 talkers that were located in the horizontal plane at 15° intervals in all directions except the target direction. The speech signals used for these 23

talkers were randomly selected from Harvard IEEE sentences (Rothauser *et al.*, 1969). The final mixture of the 23 talkers' spoken sentences was interaurally uncorrelated and was not intelligible, thus this condition is here referred to as the "diffuse-babble" condition. For each listening condition, the training set and testing set each contained 20 stimuli, and each stimulus contained 32 frequency channels and about 150 time frames.

The performance of the constructed source separation algorithms was evaluated in two ways: (1) the estimated mask was compared with the IBM described above; (2) the speech intelligibility of the separated target stream was measured with human subjects. The details of these assessments are explained below.

The first evaluation was to compare the estimated mask with the IBM. As described explicitly above, the IBM is an energy-based mask that preserves the T-F regions with SNRs above a certain threshold and that silences the T-F regions with SNRs below a certain threshold. In our calculation of the IBM, the threshold was set to 0 dB. For reverberant stimuli, we only considered the direct-path target signal as the target in the IBM calculation.²

In the second evaluation, seven normal-hearing subjects were recruited to evaluate the intelligibility of the separated speech waveforms. The CRM sentences used in the stimuli follow the structure "Ready <call sign> go to <color> <number> now" (Bolía *et al.*, 2000). Subjects were asked to choose the color and number that they heard in the target stream using a MATLAB GUI to indicate their choice. The level of the combined maskers was fixed at 70 dB sound pressure level and the target level varied adaptively. The speech reception threshold (the T/M ratio at which subjects achieved 50% correct response rate) was recorded. There were 15 types of trials (3 listening condition $\times$ 5 types of stimuli) and each type of trial was done with three repetitions. The three listening conditions, (1) anechoic multi-talker, (2) reverberant multi-talker, and (3) diffuse babble, are specified in detail above, at the beginning of this section. For each listening condition, there were five types of stimuli. Three of the stimuli were generated by applying different binary masks to the original dichotic mixture. The three binary masks were (i) the IBM (described above), (ii) the binary mask estimated with the target-EC-cue-based logistic regression classifier (LR + target-EC cue), and (iii) the binary mask estimated with the localization-cues-based logistic regression classifier (LR + localization cues). (The logistic regression classifiers are trained using training samples that are at -3 dB overall SNR. These trained classifiers were used for all the tested SNR levels.) Note that the SVM classifier with localization cues was not tested in this listening experiment. The other two types of stimuli were the "unprocessed collocated" and "unprocessed spatially separated" mixtures. For the multi-talker conditions, "collocated" means that all the talkers are placed in front, and "spatially separated" means that the two maskers are symmetrically separated from the target. For the diffuse-babble condition, "collocated" means that the left side babble is mixed with the target and the mixture is presented diotically, and "separated" means that the binaurally generated diffuse

babble is mixed with the same target in each ear and the mixture is presented dichotically.

IV. RESULTS

A. Comparison with the IBM

To quantify the pairwise similarity between each of the three estimated masks (i.e., the two localization-estimated masks and the target-EC-estimated mask) and the IBM, the hit-minus-false-alarm (HIT-FA) rate, comparing against the IBM, is calculated for each of the three classifiers in the three listening conditions for all frequency channels and averaged over channels. The HIT rate is defined as the percentage of target-dominant T-F units (T-F units with label 1 in the IBM) that are correctly labeled as 1 by the estimated classifier; the FA rate is defined as the percentage of interferer-dominant T-F units (T-F units with label 0 in the IBM) that are labeled as 1 by the estimated classifier. The HIT-FA rate has been shown to be closely correlated with speech intelligibility (Kim and Loizou, 2010) in stable noise conditions and it is a widely used metric in the CASA domain. The results were calculated with the overall SNR level setting at -3 dB. In Fig. 2, the HIT-FA rates averaged over frequency channels are plotted for each type of classifier and for each listening condition. The three classifiers involved are: the logistic regression classifier using localization cues (LR + localization cues), the logistic regression classifier using the target-EC cue (LR + target-EC cue) and the support vector machine (SVM) classifier using localization cues (SVM + localization cues).

From the bar graph in Fig. 2, it can be seen that the logistic regression classifier using the target-EC cue yields the highest HIT-FA rates in all three listening conditions, showing the best agreement with the IBM. To our surprise, using localization cues, the SVM classifier only performs

FIG. 2. HIT-FA rates of the three classifiers (indicated by different shades) in three conditions, using the IBM as the standard. The HIT rate is the percentage of correctly classified target-dominant T-F units and the FA rate is the percentage of incorrectly classified interferer-dominant T-F units. In this figure, “LR” represents the logistic regression classifier and “SVM” represents the support vector machine classifier. The signal-to-noise ratio of the evaluated mixture is -3 dB in all the cases plotted.

better than the logistic regression classifier in the diffuse-babble condition. For both anechoic and reverberant multi-talker conditions, the SVM classifier performs slightly worse than the logistic regression classifier when they both use localization cues. This implies that in some cases, ITD distance and ILD distance can be considered as independent while in other cases (e.g., diffuse-babble condition), they cannot. In the multi-talker condition, because all the speech sources are sparse in the T-F domain, each T-F unit is likely to be dominated by a single source (Yilmaz and Rickard, 2004). In this case, the ITD distance and ILD distance of the target-dominant T-F units can be considered as independent because the contamination by other sound sources in each T-F unit can be ignored. Thus, the logistic regression classifier using localization cues might do better than the SVM classifier because there is no significant interaction between ITD and ILD. For the diffuse-babble condition, even though there are 23 simultaneous interferers, the noise waveforms at the two ears might still be correlated in many T-F units. In this case, the ITD distance and ILD distance are not independent, so the SVM classifier performs better than the logistic regression classifier using localization cues as it takes the interaction between different cues into consideration.

The performance increase obtained by using the target-EC cue instead of the localization cues stems from two benefits. The first benefit is that, when extracting the target-EC cue from the binaural waveform, the ITD does not need to be calculated because the target direction is known; thus, the multi-modality confusion in determining ITD from the cross-correlation function is avoided. Although, for the localization cues, the envelope ITD is used for frequency channels above 1200 Hz, the cross-correlation functions of 500 to 1200 Hz binaural signals can still have multiple peaks in the human physiological range (i.e., within ± 1 ms). The second benefit is that the target-EC cue combines the information from ITD, ILD, and IACC together, in the sense that it depends on all three parameters. To be more specific, for a T-F unit to yield a good cancellation, both ITD and ILD should be close to the target direction parameters and the waveforms at both ears should correlate well with each other. Thus, in the calculation of the target-EC cue, the interactions among ITD, ILD, and IACC are naturally taken into consideration without building statistical models to capture their relationships.

Among the three stimulus conditions, the anechoic multi-talker condition yielded the best match to the IBM (i.e., the highest overall HIT-FA rates) and the reverberant multi-talker condition yielded the worst match to the IBM. This is within our expectation, as it is well known that reverberation can be very harmful for binaural cues. Nevertheless, we note that binaural cues are not completely destroyed even in strong reverberant conditions. With proper extraction of binaural information, notably using the target-EC process, the binaural cue could still be a very useful source separation cue.

B. Speech intelligibility evaluation

Figure 3 shows the means and standard deviations of the measured speech reception thresholds (SRTs) for the seven

FIG. 3. (Color online) Speech threshold measured for five types of stimuli (indicated by colors as labeled) in three listening environments (grouped along the axis). For each stimulus-environment combination, the mean and standard deviation are calculated from the measurements of seven normal hearing listeners.

subjects in the various environments with the several processing schemes. In both of the multi-talker conditions (i.e., anechoic and reverberant), the separated SRT is lower than that of the collocated, as one would expect. In the anechoic multi-talker condition, all three binary-mask-processed stimuli lower the SRTs by at least 10 dB, relative to the unprocessed separated mixture condition. In the reverberant multi-talker condition, the change in SRT achieved by applying binary masks is much smaller, which might be due to different reasons for different types of binary masks. For the IBM, because only the energy of the direct-path target is included as the target and the remainder is considered as noise, for the same SNR, the number of T-F units that are preserved in the reverberant condition is much less than the number preserved in the anechoic condition. Therefore, at a low SNR level, the IBM in the reverberant case might preserve too few T-F units. For the binary masks estimated with the binaural cues, because the strong reverberation severely distorted the binaural representation, the estimations for the binary masks become much less accurate. However, it is important to note that subjects can still achieve lower SRTs with the output of the binaural-cue-based source separation algorithms than with the unprocessed spatially-separated mixture (listening binaurally).

In the diffuse-babble condition, the thresholds with the diffuse babble (indicated as “separated”) and the collocated babble (indicated as “collocated”) are very close, suggesting that binaural information is not useful for the enhancement of the target in this scenario. Also, because the babble produces little informational masking (Durlach *et al.*, 2003), the SRT of the unprocessed stimuli (separated and collocated) in the diffuse-babble condition is much lower than that in multi-talker conditions. Applying the IBM to the mixture in this condition only reduces the SRT by a small amount. The SRT with the target-EC-based binary mask is similar to the SRT with unprocessed stimuli; in contrast, the SRT with localization-cues-based binary mask is even higher than the

SRT of unprocessed stimuli. If we are going to build a source-separation processor based on binaural cues, the bottom line is that the processor should not harm the intelligibility in any circumstance. Therefore, the target-EC cue would be a better candidate for such a processor than the localization cues.

It is important to mention that, at very low SNRs, all of the binary-mask processed stimuli sound very fragmented. But as this is a closed set speech task, subjects can often successfully predict the correct key words with stimuli synthesized from a limited number of T-F units, when those T-F units contain critical information to decode some of the syllables. Thus, it would be expected that for an open-set speech task, if only a limited number of T-F units are preserved, subjects would not be able to identify the word accurately with the few syllables heard. Therefore, in the future, the effectiveness of the binary mask processing strategy needs to be evaluated further with an open-set speech corpus. Pilot experiments using the IEEE corpus (instead of the CRM corpus) conducted by the authors show that the normal-hearing subjects can achieve comparable performance with target-EC processed stimuli and with the unprocessed stimuli in several listening conditions. This is promising in that the target-EC stimuli do not lose information and hinder the understanding of the target speech (consistent with results presented in this paper). Replacing the binary mask with a continuous-valued mask might also be a way to preserve more target information; this is discussed further in Sec. V.

V. DISCUSSION

This paper reports the development of a binaural processing scheme for individual T-F units, specifically one based on the effectiveness of cancellation of signals from the target direction. This target-EC based processing is compared with the processing based on localization cues (ITD, ILD, and IACC) in terms of source-separation performance. Binaural hearing contributes to auditory scene analysis by enabling us to localize and separate sound sources; therefore, many previous computational studies solve the source separation problem by location-based T-F units grouping with ITD, ILD cues. For these grouping algorithms to work, the direction of each sound source is usually assumed to be known; however, localization of multiple simultaneous sources is not an easy computational task, and there is no evidence that the human listeners need to know the directions of all sound sources to stay focused on the target stream. Experiments have only shown that human listeners can accurately localize a target source in multi-source conditions and that the accurate localization of the target is closely related to the successful understanding of the target (Yost *et al.*, 1996). It would be a burden for the brain to localize each source in a very noisy environment such as a restaurant or a bar, even without considering the complexity brought by moving sources. In this context, we note that the target-EC-cue-based source separation scheme only requires the target direction to be known, which is a minimal assumption and which is consistent with our subjective impressions.

In addition to better performance and fewer assumptions, the target-EC-cue-based source separation scheme also demands fewer computational resources. To calculate the localization cues in the time domain, the cross-correlation of each T-F unit needs to be calculated and the peak of the cross-correlation has to be identified. In contrast, to calculate the target-EC cue, we could process a long period of signal in bulk, with simple operations including shifting in time, adjusting the amplitudes, and subtraction. Though the computation power is a lesser concern nowadays, it is still better to have smaller computational loads for small electronic devices, such as the hearing aids and audio-enhancement headsets. Moreover, computation-light algorithms are more efficient in real-time applications.

The aim of this study was to compare which group of cues—localization cues or the target-EC cue—is better for source separation performance. Each classifier (the logistic regression classifier or the SVM classifier) is just a medium to conduct the comparison. In this study, the classifiers used are trained individually for each specific listening environment, which would be difficult to realize in real environments. However, we should also realize that it is impossible to train a classifier that works universally for every environment. Ideally, the decision rule of the classifier should be dynamically adjusted according to the intelligibility and quality of the separated signal. For example, in the target-EC-cue-based classifier, the OIR threshold controls how many T-F units are preserved. There is a tradeoff between the direction-selectiveness and the intelligibility of the final separated signal. If the intelligibility of the separated output could be measured, then the algorithm would be able to adjust the OIR threshold in the direction that will increase the intelligibility score. Such an intelligibility measure should be able to work without the target-alone signal; otherwise it will not be useful in real-world decisions. Unfortunately, the calculation of most current speech intelligibility/quality metrics requires the target-alone signal, such as the Speech Intelligibility Index and the short-time objective intelligibility (Taal *et al.*, 2010; ANSI, 1997). Therefore, a closed-loop design for the source-separation algorithm is not currently feasible. As the “puzzle principle” states, “we can program a computer to solve any problem by trial and error, without knowing how to solve it in advance, provided only that we have a way to recognize when the problem is solved” (Minsky, 1986). We cannot build an intelligent cocktail party processor without a standard to evaluate its performance. The human brain, in contrast, possesses a set of agents to evaluate the intelligibility and quality of the separated stream, and the feedback signal from those evaluation agents in turn changes the behavior of the “source-separation” agents.

In the classification setting, each T-F unit has only two states: target-dominated or noise-dominated. The binary mask in the classification setting has the advantage of simplicity, at the price of introducing artificial noise. There are recent studies suggesting that replacing the binary mask with a soft mask could lead to speech quality improvement (Narayanan and Wang, 2013). The values in the soft mask are continuous real numbers and are correlated with

probabilities of a T-F unit being dominated by the target. The binary mask in the target-EC-based source-separation scheme could be easily replaced with a soft mask, as described by Eq. (3). In general, the smaller the OIR, the larger the probability that the T-F unit is dominated by the target and the T-F unit should be amplified more; the larger the OIR, the larger the probability that the T-F unit is dominated by the interferers and the T-F unit should be suppressed more. So the soft mask needs to be formulated so that the OIR is inversely correlated with the value in the soft mask. In Eq. (3), the OIR is projected to the (0,1) range through a logistic function; and then the difference between one and the output of the logistic function is defined as the value of the soft-mask,

$$\text{output_mixture}(i,j) = \text{input_mixture}(i,j) \times \left(1 - \frac{1}{1 + e^{-\beta * \text{OIR}(i,j)}}\right). \quad (3)$$

In this equation, $\text{input_mixture}(i,j)$ represents the unprocessed left/right ear mixture in the i th frequency channel and the j th time slice; $\text{OIR}(i,j)$ represents the logarithm of the output-to-input ratio of the same T-F unit, so that good cancellation leads to negative values of $\text{OIR}(i,j)$ [and the $\text{output_mixture}(i,j)$ approaches the $\text{input_mixture}(i,j)$ in Eq. (3)]; and β is a positive constant, which controls the steepness of the logistic function: the larger the constant β , the steeper the logistic function. When β approximates plus infinity, the soft mask in this scheme will become a binary mask again. When β approximates zero, the soft mask value will be almost not affected by the OIR and the resulted output will not be spatially focused. We found that setting β to a value between 0.5 and 1 leads to reasonable separation performance. In an informal listening test, we found the output of the specified target-EC-soft-mask scheme sounds very similar to the output of the previous target-EC-binary-mask scheme in the three listening conditions tested. This is surprising because in the binary mask case the threshold is acquired through pre-training and the thresholds are different for different conditions, while in the soft-mask case the parameter of the OIR projection function was chosen almost arbitrarily and stays the same for all conditions (the constant β was set equal to unity in all three conditions). However, we should interpret the result of the informal listening tests with caution; formal tests are required to study the feasibility of the soft-mask scheme.

In terms of engineering solutions for the cocktail party problem, there are three general categories: blind source separation (BSS) (Bell and Sejnowski, 1995), beamforming (Levin, 1964), and computational auditory scene analysis (CASA) (Wang and Brown, 2006). The former two techniques utilize signals from multiple receptors to derive the target signal, based on assumptions of the target/noise statistics. Certain formulations of the BSS and beamforming have been shown to be equivalent to each other (Araki *et al.*, 2003). The CASA technique inspired by the human hearing, separates out the target signal by applying a mask to the original mixture; the mask is estimated based on cues that are relevant to human perception of multiple sound sources.

The CASA approach differs from the other two techniques in two aspects: first, it uses signals from no more than two receptors and it can still separate the target source when the number of sound sources exceeds the number of receptors; second, the transformations applied to the original signal are non-linear, while in the other two techniques the target signal is generally derived via linear combinations of the original multi-channel signals. The target-EC-based source-separation algorithm discussed in this paper is a CASA algorithm. The target-cancellation idea is not novel in the beamforming domain (Griffiths and Jim, 1982) and it is often used to block the target signal in order to better model the noise reference. In this paper, the result of target-cancellation is directly used as a criterion to select T-F units; and the noise estimation is not part of the source-separation process. As shown in Roman *et al.* (2006), for two-channel stimuli, with intelligibility of the separated output as the evaluation standard, the adaptive beamforming method only performs better than binaural-cues-based CASA methods when there is only one interferer. When the number of interferers exceeds the number of receptors, the performance of beamforming method degrades faster than the CASA method. Another benefit of CASA algorithms over beamforming is that binaural cues are preserved in the separated stimuli, so the listeners can still have spatial awareness. In terms of sound quality, CASA methods tend to introduce noise to the separated signal due to the non-linear transformations, and the separated signal can become very fragmented in low SNR setting due to the sparsity of the preserved T-F units. There are also algorithms that perform T-F unit manipulations on the original binaural stimuli based on the beamforming output [cf. review by Baumgärtel *et al.* (2015)]. Those algorithms have been shown to outperform traditional beamforming algorithms in terms of speech intelligibility/quality improvement while still maintaining the binaural cues in the original stimuli, which makes them promising.

In terms of psychoacoustic models for the multi-talker condition, the EC model and its variations can explain human data in many listening scenarios (Wan *et al.*, 2010; Lavandier and Culling, 2010; Beutelmann and Brand, 2006; Hawley *et al.*, 2004). Even though the EC process is formulated to cancel one noise source at a time, the EC process can still possibly attenuate multiple noise sources when multiple noise sources are at the same side of the listeners. Therefore, the predictions of the EC model can often match human data when interferers are located close to each other on one side of the listener (Wan *et al.*, 2010). In the special case of speech-envelope-modulated interferers, a short-term EC processor is able to use the sparseness of interferers to partially cancel the dominant source, even when they are symmetrically distributed with respect to the target; however, for the case that the speech interferers are symmetrically distributed to the target, the short-time EC model still fails to predict the large amount of spatial release from masking observed in psychoacoustic experiments (Wan *et al.*, 2014). Although the case of symmetrically distributed interferers is a very special spatial configuration that seldom exists in daily scenes, the large spatial benefit observed suggests that the brain has a good strategy for this condition.

The T-F glimpsing strategy has been shown to be a viable strategy (Schoenmaker and van de Par, 2016). Different strategies could be chosen based on the specific listening scenario. For example, if there are multiple widely separated interferers, the T-F glimpsing strategy might be better than the noise-cancellation strategy (e.g., EC process); if the interferers are close to each other, the noise-cancellation strategy might be more convenient to operate than the T-F glimpsing strategy. The target-EC processing scheme proposed in this paper can be viewed as a T-F selection strategy inspired by the classic EC model; however, it only uses the EC calculation as a way to cancel the target, and thereby follows a completely different strategy from the classic EC model.

The target-EC-based binaural processing scheme should be evaluated in more realistic listening environments in the future. The multiple talker conditions in this paper used the CRM corpus as the speech source for the target and interfering talkers, so that the results could be directly compared with many psychoacoustic experiments that were done using the CRM corpus (Marrone *et al.*, 2008; Brungart *et al.*, 2001). With the CRM corpus, the voice activity of the target and maskers are almost simultaneous; however, in real life, multiple people often talk sequentially and the conversations are corrupted by dynamic background noise. Future studies could focus on developing methods to simulate such listening environments and use more realistic simulations to test the proposed target-EC-based source separation approach.

ACKNOWLEDGMENTS

We would like to thank the editor and the reviewers for their time spent reviewing the manuscript and for their valuable comments helping us to improve the quality of the manuscript. The work presented in this document was supported by the National Institute of Deafness and Communication Disorders, Grant No. R01 DC 00100.

¹The software is written by Joseph G. Desloge. The software is available upon request.

²The definition of the IBM in the reverberant condition is not clear regarding whether the reverberant part of the target should be considered as target energy as well. It has been shown in psychoacoustic experiments that the IBM definition that only considers the direct-path target signal as the target leads to the largest intelligibility improvement. Including early reflections does not bring additional benefit and including the late reverberant tail will even harm the intelligibility (Roman and Woodruff, 2013).

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, Catalog No. 01TH8575, pp. 99–102.

Alinaghi, A., Wang, W., and Jackson, P. J. (2013). "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 684–688.

Allen, J. B., and Berkley, D. A. (1979). "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.* **65**, 943–950.

ANSI (1997). *American National Standard Methods for the Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

Araki, S., Makino, S., Hinamoto, Y., Mukai, R., Nishikawa, T., and Saruwatari, H. (2003). "Equivalence between frequency-domain blind

- source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP J. Adv. Sign. Process.* **2003**, 1157–1166.
- Baumgärtel, R. M., Krawczyk-Becker, M., Marquardt, D., Völker, C., Hu, H., Herzke, T., Coleman, G., Adiloğlu, K., Ernst, S. M. A., Gerkmann, T., Doclo, S., Kollmeier, B., Hohmann, V., and Dietz, M. (2015). "Comparing binaural pre-processing strategies I," *Trends Hear.* **19**, 1–16.
- Bell, A. J., and Sejnowski, T. J. (1995). "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.* **7**, 1129–1159.
- Beutelmann, R., and Brand, T. (2006). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **120**, 331–342.
- Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.* **107**, 1065–1066.
- Bronkhorst, A. W. (2015). "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Atten. Percept. Psychophys.* **77**, 1465–1487.
- Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.* **110**, 2527–2538.
- Cherry, E. C. (1953). "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.* **25**, 975–979.
- Colburn, H. S. (1973). "Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination," *J. Acoust. Soc. Am.* **54**, 1458–1470.
- Colburn, H. S. (1977). "Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise," *J. Acoust. Soc. Am.* **61**, 525–533.
- Cortes, C., and Vapnik, V. (1995). "Support-vector networks," *Mach. Learn.* **20**, 273–297.
- Deshpande, N., and Braasch, J. (2016). "A binaural model using equalization/cancellation and simulated head movements to localize and extract one speaker from a mixture," *J. Acoust. Soc. Am.* **139**, 1993.
- Durlach, N. I. (1963). "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.* **35**, 1206–1218.
- Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003). "Note on informational masking (L)," *J. Acoust. Soc. Am.* **113**, 2984–2987.
- Griffiths, L., and Jim, C. (1982). "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Ant. Propag.* **30**, 27–34.
- Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.* **115**, 833–843.
- Jeffress, L. A. (1948). "A place theory of sound localization," *J. Comp. Physiol. Psych.* **41**, 35–39.
- Kim, G., and Loizou, P. C. (2010). "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Trans. Audio Speech Lang. Process.* **18**, 2080–2090.
- Lavandier, M., and Culling, J. F. (2010). "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.* **127**, 387–399.
- Levin, M. J. (1964). "Maximum-likelihood array processing," technical report (M.I.T. Lincoln Laboratory, Lexington, MA).
- Liu, H., and Zhang, J. (2014). "A binaural sound source localization model based on time-delay compensation and interaural coherence," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1424–1428.
- Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (2010). "Model-based expectation-maximization source separation and localization," *Trans. Audio Speech Lang. Process.* **18**, 382–394.
- Marrone, N., Mason, C. R., and Kidd, G. (2008). "Tuning in the spatial dimension: Evidence from a masked speech identification task," *J. Acoust. Soc. Am.* **124**, 1146–1158.
- Mi, J., and Colburn, H. S. (2016). "A binaural grouping model for predicting speech intelligibility in multitalker environments," *Trends Hear.* **20**, 1–12.
- Minsky, M. (1986). *The Society of Mind* (Simon and Schuster, New York), p. 73.
- Narayanan, A., and Wang, D. (2013). "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7092–7096.
- Roman, N., Srinivasan, S., and Wang, D. (2006). "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Am.* **120**, 4040–4051.
- Roman, N., Wang, D., and Brown, G. J. (2003). "Speech segregation based on sound localization," *J. Acoust. Soc. Am.* **114**, 2236–2252.
- Roman, N., and Woodruff, J. (2013). "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *J. Acoust. Soc. Am.* **133**, 1707–1717.
- Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969). *IEEE Recommended Practice for Speech Quality Measurements* (IEEE Transactions on Audio Electroacoustics, New York).
- Schoenmaker, E., and van de Par, S. (2016). "Intelligibility for binaural speech with discarded low-SNR speech components," in *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing*, edited by P. van Dijk, D. Başkent, E. Gaudrain, E. de Kleine, A. Wagner, and C. Lanting (Springer International Publishing, Cham), pp. 73–81.
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217.
- Wan, R., Durlach, N. I., and Colburn, H. S. (2010). "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.* **128**, 3678–3690.
- Wan, R., Durlach, N. I., and Colburn, H. S. (2014). "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," *J. Acoust. Soc. Am.* **136**, 768–776.
- Wang, D. (2005). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Springer, New York), pp. 181–197.
- Wang, D., and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE, New York), pp. 1–395.
- Wang, D., and Brown, G. J. (2007). cochleagram [computer program], <http://web.cse.ohio-state.edu/~wang.77/pnl/shareware/cochleagram/> (Last viewed October 31, 2017).
- Yilmaz, O., and Rickard, S. (2004). "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sign. Process.* **52**, 1830–1847.
- Yin, T. C., Chan, J. C., and Carney, L. H. (1987). "Effects of interaural time delays of noise stimuli on low-frequency cells in the cat's inferior colliculus. III. Evidence for cross-correlation," *J. Neurophys.* **58**, 562–583.
- Yost, W. A., Dye, R. H., and Sheft, S. (1996). "A simulated 'cocktail party' with up to three sound sources," *Percept. Psychophys.* **58**, 1026–1036.
- Zurek, P. M. (1992). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston).