Binaural Speech Separation Using Recurrent Timing Neural Networks for Joint F0-Localisation Estimation

Stuart N. Wrigley and Guy J. Brown

Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield S1 4DP, United Kingdom
s.wrigley@dcs.shef.ac.uk, g.brown@dcs.shef.ac.uk

Abstract. A speech separation system is described in which sources are represented in a joint interaural time difference-fundamental frequency (ITD-F0) cue space. Traditionally, recurrent timing neural networks (RTNNs) have been used only to extract periodicity information; in this study, this type of network is extended in two ways. Firstly, a coincidence detector layer is introduced, each node of which is tuned to a particular ITD; secondly, the RTNN is extended to become two-dimensional to allow periodicity analysis to be performed at each best-ITD. Thus, one axis of the RTNN represents F0 and the other ITD allowing sources to be segregated on the basis of their separation in ITD-F0 space. Source segregation is performed within individual frequency channels without recourse to across-channel estimates of F0 or ITD that are commonly used in auditory scene analysis approaches. The system is evaluated on spatialised speech signals using energy-based metrics and automatic speech recognition.

1 Introduction

When listening in natural environments, the ear is bombarded with energy from multiple sound sources. Despite these sounds being mixed together, the human auditory system has the ability to analyse and extract cognitive representations for the individual sounds that are present—possibly performing this task simultaneously for multiple sources. It has been proposed that the acoustic signal is subjected to an *auditory scene analysis* in which a number of cues are extracted and used to segregate sounds on the basis of them 'belonging' to same physical source [1]. Such cues include common periodicity, common onset/offset, proximity in frequency, etc.

Human speech perception is robust even in very challenging acoustic environments; conversely, automatic speech recognition (ASR) systems can be susceptible to relatively small changes in the background acoustics. For many years, there has been interest in developing computational models of auditory scene analysis (CASA; see [2] for a review) and one use of such models is as an aide to ASR systems. In this paper, we present a novel technique of computing a joint

A. Popescu-Belis, S. Renals, and H. Bourlard (Eds.): MLMI 2007, LNCS 4892, pp. 271–282, 2007.

harmonicity-location cue in a neurobiologically plausible manner which can be used to segregate concurrent talkers and produce a mask for use with 'missing data' automatic speech recognisers [3].

The remainder of this article is organized as follows. The next section describes the two grouping cues that will be used by our system. Section 3 provides a brief overview of the competing approaches to neural representations of sounds. Following this, recurrent timing neural networks (RTNNs) are described in detail. This is followed by the implementation details of the auditory front end and the way in which this is coupled to an array of RTNNs. We present a number of evaluation techniques which have been used to assess the system and describe their results. We conclude with a discussion of the presented work and directions for future work.

2 Grouping Cues

2.1 Harmonicity

Fundamental frequency (F0) is a potent grouping cue. When listening, harmonically related components tend to form perceptual wholes (streams), whereas differences in F0 promote segregation. For example, Brokx and Nooteboom found that listeners were better able to identify two simultaneous speech utterances if they had different F0s [4].

Further support for the role of F0 in grouping comes from a number of studies which have investigated the perception of 'double vowels'. In this paradigm, a pair of steady-state, synthetic vowels are presented simultaneously, with identical onset and offset, and subjects are required to identify both vowels. Scheffers [5] found that listeners were able to identify both simultaneous vowels more accurately when they were on different F0s than when they were on the same F0. From these studies, it was proposed that a F0-guided segregation strategy is used to separate, and subsequently identify, simultaneous sounds.

Despite the fact that listeners' recognition does improve with increasing F0, doubt has been cast upon the proposed F0-guided segregation strategy. Bird and Darwin [6] investigated the mechanisms by which the auditory system exploits F0 differences in separating two sentence-length utterances. They used a stimulus in which the low-pass part of the target sentence had the same F0 as the high-pass part of the interfering sentence. The remaining parts shared the same variable F0. If the auditory system relies on global mechanisms, performance would be impaired due to inappropriate grouping of low- and high-pass parts. It was found that listener performance on the band swapped stimuli was the same as on the unmanipulated stimuli up to 2 semitones. Thus, across-frequency grouping of components across the low- and high-frequency regions only occurred for F0 differences of 5 semitones and above, but not 2 semitones and below.

2.2 Location

Listeners can also take advantage of the differing signals reaching the two ears to determine the direction of a sound source [7]. Provided a sound is not in

the median plane[1], sound energy will reach the closer ear slightly before the further ear and also with a slightly higher intensity. These two cues are referred to as *interaural time difference* (ITD) and *interaural intensity difference* (IID), the latter caused by 'shadowing' due to the head. ITD will be the focus in this study. ITDs range from 0 s for a sound directly in front of the listener's head (i.e., at an azimuth of $\pm 0°$) to about 690 μs for a sound directly opposite one of the ears (i.e., at an azimuth of $\pm 90°$).

In general, the constituent energies of a sound originating from the same location will share approximately the same ITD (we note, however, that ITD coherence is eroded in reverberant conditions; see [2, Chap. 7]). Thus, across-frequency grouping by ITD ought to provide a powerful mechanism for segregating multiple voices. Indeed, across-frequency grouping by ITD has been employed by computational models of voice separation (e.g., [8,9]).

However, analogous to F0-based segregation, there is also evidence that across-frequency grouping does not occur for interaural time difference (ITD). A number of studies have drawn across-frequency grouping by ITD into question; Edmonds and Culling [10] studied this using target and interferer pairings each of which had been low- and high-pass filtered. Even when the low-pass portion of the target and the high-pass portion of the interferer were placed at the same ITD and the remaining portions placed at a different ITD, listeners performed as well as when both target portions were presented at a consistent ITD. When both target and interferer are placed at the same ITD, performance was significantly reduced. This suggests that the auditory system exploits differences in ITD independently within each frequency channel.

3 Neural Mechanisms

The precise mechanism by which the auditory system can exploit different grouping cues (the 'neural code') remains unclear. Taking the example of harmonicity, theories of pitch perception can be considered to lie on a continuum with 'place code' models and 'temporal code' models at the extremities. The place code states that the pitch of a periodic sound corresponds to the position of maximum excitation in some tonotopically organised site in the brain. In contrast, temporal models of pitch perception use the temporal fine structure of the auditory nerve firings to determine the pitch.

A class of neural networks called *timing nets* have been suggested as a means of explaining how the auditory system uses temporally-coded input to produce meaningful outputs [11,12]. Such networks consist of coincidence detectors and delay lines and can be considered to encapsulate a range of architectures which employ analyses of interspike intervals (e.g., auto-correlation and cross-correlation). A specific form of timing nets called recurrent timing neural networks (RTNNs) requires only one spike pattern as input and has been successfully used for periodicity analysis [12]. It should be emphasised, however, that the stimuli used in

[1] A vertical plane passing through the head such that all points on the plane are equidistant from both ears.

[12] consisted of synthetic, stationary F0s. In the study presented here, we extend this work to operate on natural speech and extend the network architecture such that interaural time delay is also represented within the same network. This novel architecture allows concurrent speech to be separated on the basis of a joint F0-location cue without need for across-channel grouping: all processing is strictly within-channel.

4 Recurrent Timing Neural Networks

An RTNN consists of a bank of coincidence detectors, all operating on the same stimulus; Fig. 1(b). Each node of the network has a recurrent input exhibiting a slightly different delay; Fig. 1(a). The pattern circulating in the recurrent delay loop re-emerges after τ milliseconds; this is then compared with the stimulus arriving at the node; if a coincidence is detected, the amplitude of the delay loop input is increased by a certain factor. Regardless of the detection of a coincidence, an attenuated version of the incoming signal is fed into the delay line: without this, there would be no circulating signal to produce coincidences. Thus, stimulus periodicities equal to a node's recurrent delay will be emphasised by that node. Over time, repeating temporal patterns are enhanced relative to the rest of the stimulus. Furthermore, multiple repeating patterns with different periodicities can be detected and encoded by such networks. Cariani showed that such a relatively simple network was able to successfully separate up to three concurrent synthetic vowels [12].

In this study, we wish to represent both pitch information and ITD information in the same feature space. To achieve this, we make two important alterations to

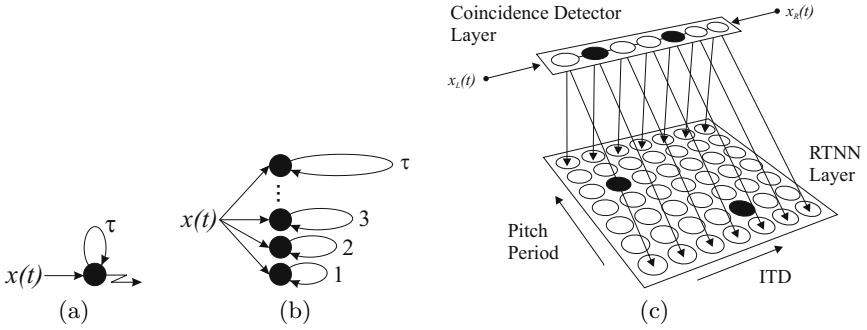

Fig. 1. (a) Coincidence detector with recurrent delay loop. (b) A group of coincidence detectors with recurrent delay loops of increasing length form a recurrent timing neural network (RTNN). Note that all nodes in the RTNN receive the same input. (c) 2D RTNN (bottom layer) with extra coincidence detector layer (top) allowing joint estimation of F0 and ITD. Downward connections are only shown for the front and back rows. Recurrent delay loops for the RTNN layer are omitted for clarity. $x_L(t)$ and $x_R(t)$ represent signals from the left and right ears respectively. Solid nodes represent activated coincidence detectors.

the architecture shown in Fig. 1(b). In order to compute the ITD feature, a one-dimensional row of coincidence detectors coupled by a delay line is introduced. In our system, 41 nodes are used, each coupled by a delay of equal duration (1 sample at 20 kHz or 50 μs). The end nodes of this row each receive one of the individual ear signals as shown in Fig. 1(c). Each signal propagates down the row; in essence, this performs a cross-correlation analysis equivalent to Jeffress' neural coincidence model of sound localisation [13]. Hence, the ITD of sounds to the right of the head are represented by coincidences in the left-hand half of the delay line (since they reach the right ear first and gain a headstart down the delay line travelling toward the left-hand edge) and vice versa.

Functionally, the delay line has the effect of performing an initial stage of source separation: activities due to spatially distinct sound sources will be emitted by their corresponding delay line node. At any one time, all nodes in the delay line will be emitting activity of some form (although only a small number will be responding strongly to sources at their best-ITD). In order to perform periodicity analysis on the output of each of these nodes, the one-dimensional network architecture shown in Fig. 1(b) is replicated to create a two-dimensional network in which each column performs periodicity analysis on the output of a single ITD delay line node; see Fig. 1(c). The activity of the two-dimensional layer, therefore, is a map with ITD on one axis and pitch on the other. Continuing the example shown in Fig. 1(c), the two-dimensional network is showing that the source nearest the right side of the head has a large pitch period, while the source towards the left side of the head has a small pitch period.

The ability to represent both F0 and ITD on the same feature space, allows the model to avoid a common problem in CASA: when multiple concurrent sources are present, how is the correct ITD associated with the correct F0? The two features are commonly computed in distinct feature spaces. In this model, they are automatically associated. Furthermore, it is easier to separate multiple sources in this feature space since it is unlikely that two sources will exhibit the same pitch and location simultaneously, thus being represented in different areas. Indeed, given a static separation of the sources, there is no need for explicit tracking of F0 or location: we simply connect the closest activity regions over time. A further advantage is that source separation can proceed within-channel without reference to a dominant F0 or dominant ITD estimate as required in an across-frequency grouping technique. Provided there is some separation in one or both of the cues, two activity regions (in the case of two simultaneous talkers) can be extracted and assigned to different sources.

5 The Model

5.1 Auditory Periphery

A bank of 20 gammatone filters [14] with overlapping passbands simulate the frequency analysis performed by the basilar membrane. Their centre frequencies range from 100 Hz to 8 kHz and are equally spaced on the equivalent rectangular

bandwidth (ERB) scale [15]. A gammatone filter of order n and centre frequency f Hz is defined as:

$$gt(t) = t^{n-1} \exp(-2\pi bt) \cos(2\pi ft + \phi) H(t) \tag{1}$$

where f is the centre frequency, ϕ is phase, b is related to bandwidth, and $H(t)$ is the unit step (Heaviside) function defined as $H(t) = 1$ if $t \geq 0$, $H(t) = 0$ otherwise. Here, we use fourth-order gammatone filters. Since later stages of the model only require periodicity information, the auditory nerve response is approximated by half-wave rectifying and cube root compressing the output of each filter.

5.2 RTNN

The RTNN layer consists of a grid of independent (i.e., unconnected) coincidence detectors with an input from the ITD estimation layer (described above) and a recurrent delay loop. For a node with a recurrent delay loop duration of τ whose input $x_\theta(t)$ is received from the ITD node tuned to an interaural delay of θ, the update rule is:

$$C(t) = \alpha x_\theta(t) + \beta x_\theta(t) C(t - \tau) \ . \tag{2}$$

The output of the node (and, hence, also about to enter the recurrent delay loop) is denoted by $C(t)$; $C(t-\tau)$ is the response which is just emerging from the delay loop. To ensure some form of signal is always circulating in the delay loop to allow coincidences to occur, α acts as an attenuator for the incoming signal. Note that α is sufficiently small so as not to dominate the node's response ($\alpha = 0.2$). Should a coincidence occur, the weight β determines the rate of adjustment and is dependent on τ such that coincidences at low pitches are de-emphasized [12]. Here, β increases linearly from 3 at the smallest recurrent delay loop length to 10 at the largest.

In order to perform the joint pitch-ITD analysis on each auditory nerve centre frequency, the model employs 20 independent networks (of the form shown in Fig. 1(c)), one for each frequency channel. Network activity is captured using a sliding temporal window in which activity over the duration of the window is averaged. In our model, we use a window size of 25 ms and a temporal shift of 5 ms. Therefore, for every frequency channel, a sequence of two-dimensional activity plots is built up over time.

Ultimately, the system should allow concurrent speakers to be separated and be transcribed by an automatic speech recognition system. Thus, it is necessary to know, at any time-frequency point during the signal, whether that point is dominated by the target speech or by some form of interference. The network activity plots (one generated every 5 ms per frequency channel) can be used to make an estimate of talker activity at a particular time-frequency point: a highly active node relative to the rest of the network indicates that the source at that F0-ITD combination is active. Specifically, a time-frequency binary mask for the target talker is created from the RTNN output. A time-frequency mask unit is set to 1 if the target talker's activity was greater than the target's mean

activity for that frequency channel, otherwise it is set to 0. Talker activity can be grouped across time frames by associating the closest active nodes in F0-ITD space (assuming the two talkers don't momentarily have the same ITD *and* F0).

6 Evaluation

The separation technique described above was evaluated on a number of speech mixtures drawn from the TIdigits Studio Quality Speaker-Independent Connected-Digit Corpus [16]. The sampling rate for the corpus is 20 kHz.

In order to investigate the influence of spatial separation on the ability of the system to successfully segregate concurrent speakers, three different target+interferer spatialisations were used: -40°+40°, -20°+20° and -10°+10°. For each scenario, 100 randomly selected utterance pairs were created, all of which were from male talkers. To avoid duration mismatches due to differing speaking rates across subjects, the target utterance consisted of five digits and the interferer utterance consisted of seven digits. Furthermore, the target was always on the left of the azimuth midline. The signals were spatialised by convolving them with head related transfer functions (HRTFs) measured from a KEMAR artificial head in an anechoic environment [17]. The two speech signals where then combined with a signal-to-noise ratio (SNR) of 0 dB. The SNR was calculated using the original, monaural, signals prior to spatialisation.

Three forms of evaluation were employed: assessment of the amount of target energy lost (P_{EL}) and interferer energy remaining (P_{NR}) in the mask [18, p. 1146]; target speaker SNR improvement; ASR performance improvement.

All three techniques require the use of an a priori binary mask (an 'optimal' mask). The a priori binary mask is formed by placing a 1 in any time-frequency units where the energy in the mixed signal is within 1 dB of the energy in the clean target speech (the regions which are dominated by target speech), otherwise they are set to 0. In other words, an a priori binary mask uses information about regions of uncorrupted speech within the mixture.

6.1 Signal Energy

In order to assess the quality of segregation based upon an energy metric, it is necessary to obtain a number of time-domain signals. These signals are derived from a resynthesis process which uses a binary mask to determine which time-frequency portions are required and which are to be discarded. The gammatone filter outputs for each frequency channel are divided into frames of size equal to the binary mask resolution. Each signal frame is then weighted by the value of the binary mask at that time-frequency point. Individual channels are recovered using the overlap-and-add method and these are summed across frequencies to yield a resynthesized signal. The percentage of target speech excluded from the segregated speech (P_{EL}), and the percentage of interferer included (P_{NR}) are defined to be [18, p. 1146]:

$$P_{\mathrm{EL}} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \ , \tag{3}$$

$$P_{\mathrm{NR}} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)} \ . \tag{4}$$

The clean target signal which has been resynthesized using the a priori binary mask is denoted by $I(n)$. $O(n)$ is the clean target signal which has been resynthesized using the RTNN produced mask (the actual separated signal produced by our system). $e_1(n)$ is the clean target signal which has been resynthesized using a mask in which 1s are present at all time-frequency points which are 1 in the a priori binary mask and 0 in the RTNN produced mask (the portions of the signal which ought to be present but are missing in the system's separated signal). $e_2(n)$ is the opposite of this; in other words, $e_2(n)$ is the clean target signal which has been resynthesized using a mask in which 1s are present at all time-frequency points which are 1 in the RTNN produced mask and 0 in the a priori binary mask (the portions of the signal which are present but should not be: remaining interferer).

In addition to resynthesizing the target $O(n)$, the interferer signal is also resynthesized using the RTNN generated target mask. This allows the calculation of SNR (an easily understood metric) before and after processing.

6.2 Automatic Speech Recognition

The third evaluation technique involves using an ASR system which can exploit the 'missing data' technique [3]. The task of ASR can be defined as the assignment of an acoustic observation x to a class of speech sound C. However, some components of x may be unreliable or missing due to an interfering sound source. In this situation, the likelihood of the acoustic model $f(x|C)$ cannot be established in the usual manner. To overcome this problem, the missing data approach partitions x into reliable and unreliable components, x_r and x_u. The reliable components x_r (whose values are known) are directly available to the classifier whereas the unreliable components x_u are uncertain.

More specifically, the marginal distribution $f(x_r|C)$ is used directly and the likelihood $f(x_u|C)$ is estimated by integrating over bounds (i.e., between zero and the observed energy) [3].

The missing data approach requires a process which identifies the reliable and unreliable components, x_r and x_u. Here, we use the RTNN time-frequency binary mask to indicate whether the acoustic evidence in each time-frequency region is reliable; units assigned 1 in the binary mask define the reliable areas of target speech whereas units assigned 0 represent unreliable regions.

The features used by the recogniser in this study are known as *auditory rate maps*. They are computed by calculating the instantaneous Hilbert envelope of each gammatone filter response [19]. This is smoothed by a low-pass filter with an 8 ms time constant, sampled at 5 ms intervals (to match the RTNN binary mask resolution), and cube root compressed to give a pseudo-spectrogram representation of auditory firing rate (Fig. 2).

Auditory rate maps were obtained for the training section of the corpus, and were used to train 12 word-level HMMs (a silence model, 'oh', 'zero' and '1' to '9') each consisting of 18 no-skip, straight-through states with observations modelled

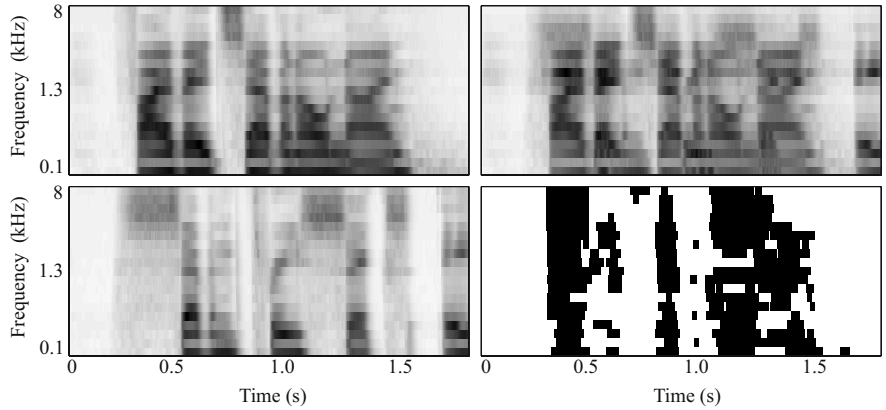

Fig. 2. The upper left panel shows the ratemap for the target utterance ('587o9') and the interfering utterance ('736883o') ratemap is shown below it. Darker areas indicate higher energy; the interfering ratemap has been truncated to match the duration of the target. The upper right panel shows the ratemap of the two utterances which have been spatialised to -40° (target) and 40° (interferer) and mixed at a monaural SNR of 0 dB. The bottom right panel shows the RTNN missing data mask for the target utterance—black pixels denote time-frequency regions dominated by the target utterance.

by a 12 component diagonal Gaussian mixture. Training was performed using HTK [20] and testing used Barker's Computational Auditory Scene Analysis Toolkit (CTK)[2].

6.3 Results

The results from our model are shown in Table 1; each value represents the average performance of 100 target and interferer utterance pairs for the three different spatial separations. In addition, the average performance across all spatial separations is included. The data in the table is split into three main classes: P_{EL} and P_{NR}, SNR improvement and ASR improvement. For comparison, a priori performance is also shown for SNR and ASR approaches. These values are calculated for the left ear (the ear closest to the target). Although the speech signals were mixed at 0 dB relative to the monaural signals, the actual SNR at the left ear for the spatialised signals will depend on the spatial separation of the two talkers (hence its inclusion in the table).

On average, the RTNN system removes over 91% of the interfering utterance with this figure rising to 94% at the most favourable separation. In addition to this, the amount of energy incorrectly removed from the target P_{EL} is approximately 11%. The SNR metric shows a significant improvement at all interferer positions; on average, our model exhibits an improvement factor of 3.7 when compared to the SNR before separation. Furthermore, it can be observed that

[2] Available from http://www.dcs.shef.ac.uk/∼jon/ctk.html

Table 1. Separation performance for concurrent speech at different interferer azimuth positions in degrees; 'pre' denotes performance before processing; 'RTNN' denotes performance after processing and 'a priori' denotes 'optimal' performance. ASR accuracy is (100% - word error rate).

	±10°	±20°	±40°	AVERAGE
SNR (dB) pre	1.64	3.13	5.19	3.32
SNR (dB) RTNN	10.03	11.55	14.49	12.02
SNR (dB) a priori	12.35	13.27	15.01	13.54
Mean P_{EL} (%)	10.62	12.74	10.22	11.19
Mean P_{NR} (%)	9.99	8.42	6.02	8.14
ASR Acc. (%) pre	15.00	22.20	28.20	21.80
ASR Acc. (%) RTNN	71.60	74.60	83.40	76.53
ASR Acc. (%) a priori	93.40	94.00	94.60	94.00

SNR performance approaches the a priori values at wider separations. Such SNR performance is supported by the low values for P_{NR} which indicate good levels of interferer rejection and relatively little target loss.

Importantly, the missing data ASR paradigm is tolerant to this relatively low level of target energy loss as indicated by the ASR accuracy performance. Indeed, the missing data ASR performance remains relatively robust when compared to the baseline system which used a unity mask (all time-frequency units assigned 1). We note that ASR performance also approaches the a priori values at wider angular separations. Furthermore, we predict that an increased sampling rate would produce improvements in performance for both SNR and ASR at smaller separations due to the higher resolution of the ITD sensitive layer (see below).

7 Conclusions

A number of grouping cues play important roles in the auditory system's ability to segregate competing sounds. Two of these cues are harmonicity and location. The neural coding strategy by which such cues are represented is the subject of continued debate. A type of neural network called recurrent timing neural networks has been proposed as a means of explaining how the auditory system uses temporally-coded input to produce meaningful outputs [11,12]. Such networks have been used successfully in previous studies to separate multiple concurrent synthetic vowels using periodicity information.

In this study we extended such one-dimensional networks to allow the architecture to represent sounds in a joint F0-ITD cue space. The system was evaluated using a much more challenging paradigm than the synthetic static vowels used in previous RTNN studies [11,12]. Here, the scenarios consist of concurrent real speech mixed at an SNR of 0 dB. Unlike stationary vowels, each

constituent signal contains fluctuating F0s and sections of unvoiced speech are common.

The results shown in Table 1 indicate high levels of interferer rejection with low levels of incorrectly removed target speech. The system retained an average of 88.81% of target speech energy and removed over 91% of the interferer. SNR values for the separated target speech also indicate good separation, and informal listening tests found that target speech extracted by the system was of good quality. SNR performance reported here (10.03 dB at the smallest separation) also compares well with those of [9], although direct comparison is difficult due to differing stimuli and spatial separations. The energy-based mechanism allowing unvoiced segments to be represented in the RTNN binary mask successfully included the utterances' fricatives. We note that the target signals commonly used in such evaluations tend to be voiced throughout (e.g., [18]), thus avoiding the problem of unvoiced energy.

The relatively wide spatial separations employed here were by necessity of the sampling rate of the speech corpus: at 20 kHz an ITD of one sample is equivalent to an angular separation of approximately 5.4°. Thus, the smallest separation used here corresponds to an ITD of just 3.7 samples. A means of addressing this issue is to upsample the corpus to a higher sampling rate of, for example, 48 kHz. However, this has the effect of significantly increasing the size of the RTNN and thus the computational load—a topic of future work. We will also test the system on a larger range of SNRs and larger set of interferer positions.

Furthermore, an assumption made by the system is that the target is always on the left side of the head. At each frequency, the activity of a source is represented in F0-ITD space. Over the duration of the signal, the position of this source will fluctuate with pitch and location, hence, creating a trace through the 3-dimensional space of F0, ITD and time. This presents a permutation ambiguity problem similar to that encountered in frequency-domain blind source separation approaches [21], which could be solved by combining channels that have a similar temporal structure. Alternatively, an attentional process could be employed which would select one of the sources to be the target based upon some a priori knowledge of the target and track it across time (e.g., [22]).

Acknowledgments. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811).

References

1. Bregman, A.S.: Auditory Scene Analysis. The Perceptual Organization of Sound. MIT Press, Cambridge (1990)
2. Wang, D., Brown, G.J. (eds.): Computational Auditory Scene Analysis: Principles, Algorithms and Applications. IEEE Press / Wiley-Interscience (2006)
3. Cooke, M., Green, P., Josifovski, L., Vizinho, A.: Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34(3), 267–285 (2001)

4. Brokx, J.P.L., Nooteboom, S.G.: Intonation and the perceptual separation of simultaneous voices. J. Phonetics 10, 23–36 (1982)
5. Scheffers, M.T.M.: Sifting Vowels: Auditory Pitch Analysis and Sound Segregation. PhD thesis, Groningen University, The Netherlands (1983)
6. Bird, J., Darwin, C.J.: Effects of a difference in fundamental frequency in separating two sentences. In: Palmer, A.R., Rees, A., Summerfield, A.Q., Meddis, R. (eds.) Psychophysical and physiological advances in hearing, Whurr, pp. 263–269 (1997)
7. Blauert, J.: Spatial Hearing — The Psychophysics of Human Sound Localization. MIT Press, Cambridge (1997)
8. Lyon, R.F.: A computational model of binaural localization and separation. In: Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1148–1151 (1983)
9. Roman, N., Wang, D., Brown, G.J.: Speech segregation based on sound localization. J. Acoust. Soc. Am. 114, 2236–2252 (2003)
10. Edmonds, B.A., Culling, J.F.: The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay. J. Acoust. Soc. Am. 117, 3069–3078 (2005)
11. Cariani, P.A.: Neural timing nets. Neural Networks 14, 737–753 (2001)
12. Cariani, P.A.: Recurrent timing nets for auditory scene analysis. In: Proc. Intl. Conf. on Neural Networks (IJCNN) (2003)
13. Jeffress, L.A.: A place theory of sound localization. J. Comp. Physiol. Psychol. 41, 35–39 (1948)
14. Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P.: An efficient auditory filterbank based on the gammatone function. Technical Report 2341, Applied Psychology Unit, University of Cambridge, UK (1988)
15. Glasberg, B.R., Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data. Hearing Res. 47, 103–138 (1990)
16. Leonard, R.G.: A database for speaker-independent digit recognition. In: Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). vol. 3 (1984)
17. Gardner, W.G., Martin, K.D.: HRTF measurements of a KEMAR. J. Acoust. Soc. Am. 97(6), 3907–3908 (1995)
18. Hu, G., Wang, D.: Monaural speech segregation based on pitch tracking and amplitude modulation. Neural Networks 15(5), 1135–1150 (2004)
19. Cooke, M.P.: Modelling auditory processing and organisation. Cambridge University Press, Cambridge (1991/1993)
20. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.3). Cambridge University Engineering Department (2005)
21. Murata, N., Ikeda, S., Ziehe, A.: An approach to blind source separation based on temporal structure of speech signals. Neurocomputing 41, 1–24 (2001)
22. Wrigley, S.N., Brown, G.J.: A computational model of auditory selective attention. IEEE Trans. Neural Networks 15(5), 1151–1163 (2004)