RECURRENT TIMING NEURAL NETWORKS FOR JOINT F0-LOCALISATION BASED SPEECH SEPARATION

Stuart N. Wrigley and Guy J. Brown

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK

s.wrigley@dcs.shef.ac.uk, g.brown@dcs.shef.ac.uk

ABSTRACT

A novel extension to recurrent timing neural networks (RTNNs) is proposed which allows such networks to exploit a joint interaural time difference-fundamental frequency (ITD-F0) auditory cue as opposed to F0 only. This extension involves coupling a second layer of coincidence detectors to a two-dimensional RTNN. The coincidence detectors are tuned to particular ITDs and each feeds excitation to a column in the RTNN. Thus, one axis of the RTNN represents F0 and the other ITD. The resulting behaviour allows sources to be segregated on the basis of their separation in ITD-F0 space. Furthermore, all grouping and segregation activity proceeds 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 has been evaluated using a source separation task operating on spatialised speech signals.

Index Terms— Speech processing, Speech enhancement, Neural network architecture, Auditory system

1. INTRODUCTION

Bregman [1] has proposed that the human auditory system analyses and extracts representations of the individual sounds present in an environment in a manner similar to scene analysis in vision. Such *auditory scene analysis* (ASA) takes place in two stages. Firstly, the signal is decomposed into a number of discrete sensory elements. These are then recombined into *streams* on the basis of the likelihood of them having arisen from the same physical source in a process termed *perceptual grouping*. The auditory system uses a number of grouping cues such as common periodicity, common onset/offset, proximity in frequency as well as knowledge of commonly experienced acoustic stimuli.

1.1. Harmonicity and location as grouping cues

One of the most powerful grouping cues is harmonicity. Listeners are able to identify both constituents of a pair of simultaneous vowels more accurately when they are on different fundamental frequencies (F0s) rather than on the same F0 (e.g., [2]). On the basis of such studies, it has been proposed that a F0-guided segregation strategy is used to separate, and subsequently identify, simultaneous sounds. A number of computational models of auditory perception exploited this approach (e.g., [3]). For example, it is common to perform bandpass filtering at a number of centre frequencies (to simulate cochlear filtering) followed by periodicity analysis of each channel. Periodicity estimates are merged across frequency to generate an overall estimate of the dominant pitch. Two distinct groups of channels are then created using the dominant pitch estimate: one set consists of all the channels which exhibit a peak at the pitch period and the other set contains the remaining channels.

However, listener performance in such a task may not be due to across-frequency grouping but rather the exploitation of other signal properties such as spectral modulation [4]. Indeed, it has also been shown that although listeners' recognition performance for concurrent speech improves with increasing F0, they only take advantage of across-frequency grouping for separations greater than 5 semitones [5].

There is also mounting evidence that across-frequency grouping does not occur for interaural time difference (ITD) either. ITD is an important cue used by the human auditory system to determine the direction of a sound source [6]. For sound originating from the same location, its constituent energies at different frequencies will share approximately the same ITD. Thus, across-frequency grouping by ITD has been employed by a number of computational models of voice separation (e.g., [7]). However, recent studies have drawn this theory into question; Edmonds and Culling [8] 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.

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

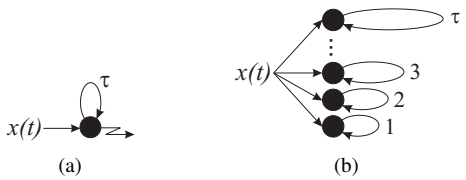

Fig. 1. (a) Coincidence detector with recurrent delay loop. (b) A group of coincidence detectors with recurrent delay loops of increasing length form a recurrent timing neural network (RTNN). Note that all nodes in the RTNN receive the same input.

1.2. Neural mechanisms

Despite strong evidence that harmonicity and ITD are exploited by the auditory system for grouping and segregation (see above), it remains unclear as to the precise mechanism (the 'neural code') by which this occurs. Recently, Cariani has shown that recurrent timing neural networks (RTNNs) can be used as neurocomputational models of how the auditory system processes temporal information to produce stabilised auditory percepts [9, 10]. Indeed, [10] showed that such a relatively simple network was able to successfully separate up to three concurrent synthetic vowels. 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 a mixture of two or more speech signals to be separated on the basis of a joint F0-location cue without need for across-frequency grouping.

2. RECURRENT TIMING NEURAL NETWORKS

The building block of an RTNN is a coincidence detector in which one input is the incoming stimulus response and the other input is from a recurrent delay line (Figure 1(a)). The output of the coincidence detector is fed into the delay line and re-emerges τ milliseconds later. If a coincidence between the incoming signal and the recurrent signal is detected, the amplitude of the circulating pulse is increased by a certain factor.

Pitch analysis approaches employ a one dimensional network, similar to the one shown in Figure 1(b), in which each node has a recurrent delay line of increasing length. As periodic signals are fed into the network, activity builds up in nodes whose delay loop lengths are the same as that of the signal periodicity; activity remains low in the other nodes. Furthermore, multiple repeating patterns with different periodicities can be detected and encoded by such networks: a property exploited by Cariani to separate concurrent synthetic vowels [10] (see also [11]).

We develop this type of network in two ways. Firstly, the network is extended to be two dimensional and, secondly, an additional layer of coincidence detectors are placed between the incoming signal and the RTNN nodes. This allows the network to produce a simultaneous estimate of ITD and F0. Figure 2 shows a schematic of the new network.

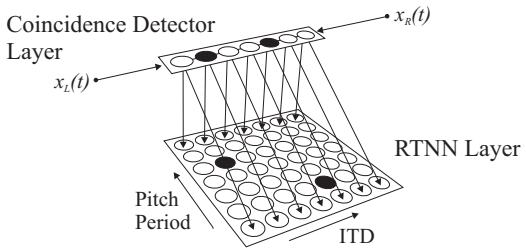

Fig. 2. RTNN (bottom) with coincidence detector layer (top) allowing joint estimation of pitch period and ITD. Each node in the coincidence detector layer is connected to every node in the corresponding RTNN column. Downward connections are only shown for the front and back rows. Recurrent delay loops 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 circles represent activated coincidence detectors.

The first layer receives the stimulus input (with each ear's signal fed into opposite sides of the grid) and is equivalent to the neural coincidence model of Jeffress [12]. This layer acts as the first stage of stimulus separation: the outputs of each node represent each of the constituent, spatially separated, mixture sources. The RTNN layer is expanded to be two dimensional to allow the output of every ITD sensitive node from the top layer to be subject to the pitch analysis of a standard one-dimensional RTNN such as the one shown in Figure 1(b). The activity of the RTNN layer, therefore, is a two-dimensional map with ITD on one axis and pitch period on the other. The example in Figure 2 shows the RTNN (bottom) indicating 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. Note that a source to the left of the head will exhibit a response in the right side of the ITD coincidence detector layer and, hence, also the RTNN (and *vice versa*).

The advantage of this approach is the joint representation of F0 and ITD within the same 'feature map'. Multiple sources tend to be separated on this map since it is unlikely that two sources will exhibit the same pitch and location simultaneously. Indeed, given a static spatial 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.

3. THE MODEL

3.1. Auditory periphery

The frequency selectivity of the basilar membrane is modelled by a bank of 20 gammatone filters [13] whose centre frequencies are equally spaced on the equivalent rectangular bandwidth (ERB) scale [14] between 100 Hz and 8 kHz. Since the RTNN is only used to extract pitch information, each gammatone filter output is low-pass filtered with a cutoff frequency of 300 Hz using a 8th order Butterworth filter.

3.2. RTNN

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:

$$C(t) = \alpha x_\theta(t) + \beta x_\theta(t) C(t-\tau) \quad (1)$$

Here, $C(t)$ is the response which is just about to enter the recurrent delay loop and $C(t-\tau)$ is the response which is just emerging. The weight α is an attenuator for the incoming signal which ensures some input to the recurrent delay loop required for later coincidence detection but is sufficiently small that it does not dominate the node's response ($\alpha = 0.2$). The second weight β determines the rate of adjustment when a coincidence is detected and is dependent on τ such that coincidences at low pitches are de-emphasized [10]. Here, β increases linearly from 3 at the smallest recurrent delay loop length to 10 at the largest.

In the complete system, there are 20 independent networks, each consisting of an ITD coincidence layer coupled to a RTNN layer (as shown in Figure 2), for each frequency channel. The state of each channel's RTNN is assessed every 5 ms using the mean activity over the previous 25 ms; this is used to make an estimate of source activity: highly active nodes indicate that the talker at that F0-ITD combination is active. For example, if the node representing 90 Hz and an ITD of -200 μs is active, it is likely that there is a low-pitched voice situated to the left of the listener. 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).

4. EVALUATION

The system was evaluated on a number of speech mixtures drawn from the TIdigits Corpus [15]. From this corpus, a set of 100 randomly selected utterance pairs were created, all of which were from male talkers. For each pair, three target+interferer separations were generated at azimuths of -40°+40°, -20°+20° and -10°+10°. Note 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 [16]. 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.

Once the system had processed each mixture, a time - frequency binary mask for the target talker was created from the RTNN output. A time-frequency mask unit was set to 1 if the target talker's activity was greater than the mean activity for that frequency channel, otherwise it was set to 0. However, RTNNs cannot represent nonperiodic sounds; in order to segregate unvoiced speech, a time-frequency unit was also set to 1 if there was high energy at the location of the target but no RTNN activity.

The evaluation techniques described below require a number of resynthesized signals. To achieve this, the gammatone filter outputs are divided into 25 ms time frames with a shift of 5 ms to yield a time-frequency decomposition corresponding to that used when generating the mask. These signals are weighted by the binary mask and each channel is recovered using the overlap-and-add method. 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 [17, p. 1146]:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \quad (2)$$

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)} \quad (3)$$

Here, $I(n)$ is the clean target signal which has been resynthesized using an *a priori* binary 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. $O(n)$ 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).

An alternative approach for evaluating separation performance was also employed which involved resynthesizing the target and noise signals using the RTNN generated target mask. This allows the calculation of SNR (an easily understood metric) before and after processing.

Table 1. Separation performance for concurrent speech at different interferer azimuth positions in degrees; 'pre' denotes SNRs before processing; 'RTNN' denotes SNRs after processing and '*a priori*' denotes 'optimal' SNR improvement.

	$\pm$**10°**	$\pm$**20°**	$\pm$**40°**	**AVERAGE**
SNR (dB) pre	1.64	3.13	5.19	3.32
SNR (dB) RTNN	7.60	9.45	12.51	9.90
SNR (dB) *a priori*	12.35	13.27	14.49	13.37
Mean P_{EL} (%)	7.21	7.58	6.91	7.24
Mean P_{NR} (%)	11.13	9.32	6.83	9.09

Table 1 shows SNR before and after processing, P_{EL} and P_{NR} for each separation. For comparison, the SNR achieved using the *a priori* binary mask is also shown. 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. The SNR metric shows a significant improvement at all interferer positions (on average, a threefold improvement). This is supported by low values for P_{NR} indicating good levels of interferer rejection. We note that performance approaches the *a priori* SNR values at wider separations. Furthermore, we predict that an increased sampling rate would produce improvements in performance at smaller separations due to the higher resolution of the ITD sensitive layer (see Discussion).

5. DISCUSSION

A novel extension to Cariani's original pitch analysis recurrent timing neural networks has been described that allows the incorporation of ITD information to produce a joint F0-ITD cue space. Unlike Cariani's evaluation using synthetic static vowels, our approach has been evaluated using a much more challenging paradigm: concurrent real speech mixed at an SNR of 0 dB. The results presented here indicate good separation and the low P_{NR} values confirm high levels of interferer rejection, even for periods of unvoiced target activity. Informal listening tests found that target speech extracted by the system was of good quality.

Relatively wide spatial separations were employed here 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°. To address this issue, we have collected a new corpus of signals using a binaural manikin at a sampling rate of 48 kHz and work is currently concentrating on adapting the system to this much higher sampling rate (and hence significantly larger networks). In addition, we will test the system on a larger range of SNRs and larger set of interferer positions. Our eventual goal is to use the system as a front-end for automatic speech recognition (ASR).

6. REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis. The Perceptual Organization of Sound*, MIT Press, 1990.

[2] P. F. Assmann and A. Q. Summerfield, "Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 88, pp. 680–697, 1990.

[3] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley-IEEE Press, 2006.

[4] J. F. Culling and C. J. Darwin, "Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1559–1569, 1994.

[5] J. Bird and C. J. Darwin, "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and physiological advances in hearing*, A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, Eds., pp. 263–269. Whurr, 1997.

[6] J. Blauert, *Spatial Hearing — The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[7] D. Wang N. Roman and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[8] B. A. Edmonds and J. F. Culling, "The spatial unmasking of speech: evidence for within-channel processing of interaural time delay," *J. Acoust. Soc. Am.*, vol. 117, pp. 3069–3078, 2005.

[9] P. A. Cariani, "Neural timing nets," *Neural Networks*, vol. 14, pp. 737–753, 2001.

[10] P. A. Cariani, "Recurrent timing nets for auditory scene analysis," in *Proc. Intl. Conf. on Neural Networks (IJCNN)*, 2003.

[11] A. de Cheveigné, "Time-domain auditory processing of speech," *J. Phonetics*, vol. 31, pp. 547–561, 2003.

[12] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, pp. 35–39, 1948.

[13] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Tech. Rep. 2341, Applied Psychology Unit, University of Cambridge, UK, 1988.

[14] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, 1990.

[15] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1984, vol. 3.

[16] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1995.

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