Joint F0-localisation estimation using recurrent timing neural networks

Stuart N. Wrigley and Guy J. Brown

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

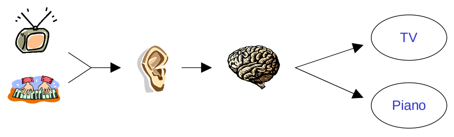
{s.wrigley,g.brown}@dcs.shef.ac.uk

http://www.amiproject.org

Introduction – auditory scene analysis

The ear is bombarded with energy from multiple sound sources, some of which exhibit very similar characteristics (pitch, location, etc.).

Despite being mixed together, the human auditory system has the ability to analyse and extract cognitive representations for the individual sounds.



Achieved by auditory scene analysis (ASA) - a two step process:
1. decomposition into discrete sensory elements
2. perceptual grouping forms streams (one per sound source)

Goals:

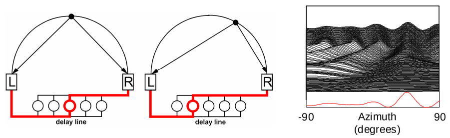
1. Separate concurrent speech using joint F0-ITD cue
2. Demonstrate RTNNs can be applied to real signals
3. Perform all processing within-channel

Interaural time difference (ITD)

ITD is an important cue used by the human auditory system to determine the direction of a sound source.

Bandpass filtering at a number of centre frequencies simulates cochlear filtering.

Conventionally, estimated by cross-correlation of the left and right auditory nerve response approximations at each frequency channel.



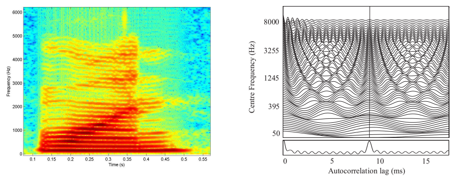
Increasing evidence that across-frequency grouping does not occur for interaural time difference (ITD).

Rather, differences in ITD are exploited independently within each frequency channel[1].

Harmonicity

One of the most powerful grouping cues.

Many natural sounds, including speech, are caused by the vibration of some physical structure followed by filtering and resonance. Exhibit a fundamental (F0) and a number of related harmonics.



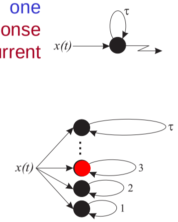
Conventionally, periodicity estimates are merged across frequency to generate an overall estimate of the dominant pitch. Channels which agree with this pitch are then grouped together.

However, doubt over physiological use of global pitches[2]. Our aim is to segregate speech using only within-channel mechanisms.

Recurrent timing neural network (RTNN)

Coincidence detectors in which one input is the incoming stimulus response and the other input is from a recurrent delay line.

Pitch analysis: 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.

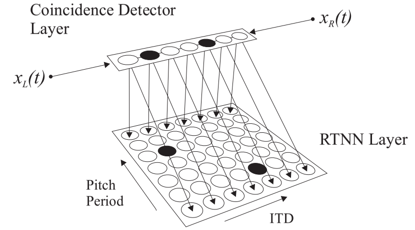


Used by Cariani to separate 3 concurrent synthetic vowels[3].

We expand the system to deal with the ITD cue. Extra layer of delay line coincidence detectors performs cross-correlation analysis.

RTNN becomes 2D: each ITD lag node feeds information to one column; each column is a standard 1D RTNN.

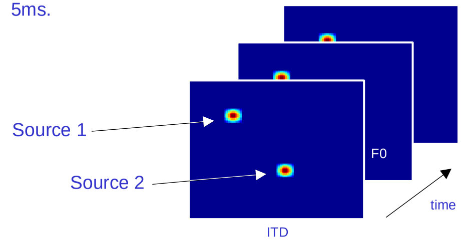
RTNN for joint F0-ITD



2 gammatone filter banks simulate cochlear filtering for for each ear. Filter outputs lowpass filtered at 300Hz, HWR and $\sqrt[3]{}$ compressed to give $x_L(t)$ and $x_R(t)$.

This processing occurs for every frequency channel.

Average of previous 25ms activity calculated every 5ms.



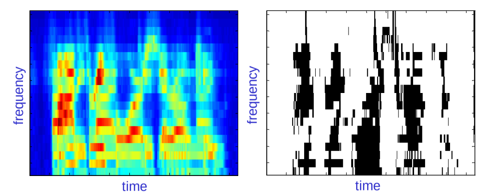
Binary mask generation

A time-frequency unit is set to 1 if the target talker is active in that frequency channel and time frame, otherwise it was set to 0. Target talker is active if RTNN activity found in expected region.

Assumption: target is always on the left.

However, RTNNs can only segregate periodic speech; in order to segregate unvoiced speech, a time-frequency unit is set to 1 if there is high energy at the previous location of the target but no RTNN activity.

Ratemap and mask for *534z2* (male speaker) at -40°; interfering speech at +40°.



Evaluation data and metrics

Speech mixtures drawn from the *TIdigits Studio Quality Speaker-Independent Connected-Digit Corpus*.

100 randomly selected male utterance pairs; 3 types of pairing: -40°+40°, -20°+20° and -10°+10°. TIR of 0dB (prior to spatialisation). The signals were spatialised by convolving them with HRTFs.

Three evaluation metrics:
1. percentage of target speech excluded from the segregated speech (P_{EL}) and percentage of interferer included (P_{NR})
2. SNR improvement
3. ASR performance improvement

1 & 2 use resynthesised target speech using the binary mask.

ASR used 'missing data' technique; RTNN mask used to specify reliable and unreliable spectral regions.

Trained on whole TIdigits training set using HTK; segregated target recognised using CTK (a missing data recogniser).

Evaluation results

	± 10°	± 20°	± 40°	Average
SNR (dB) pre processing	1.64	3.13	5.19	3.32
SNR (dB) post (higher better)	10.03	11.55	15.01	12.20
SNR (dB) *a priori* (higher better)	12.35	13.27	14.49	13.37
Mean P_{EL} (%) (lower better)	10.62	12.74	10.22	11.19
Mean P_{NR} (%) (lower better)	9.99	8.42	6.02	8.14
ASR Acc. (%) pre processing	15.00	22.20	28.20	21.80
ASR Acc. (%) post	71.60	74.60	83.40	76.53
ASR Acc. (%) *a priori*	93.40	94.00	94.60	94.00

Conclusions

Novel form of RTNN to exploit joint F0-ITD cue for speech separation performs well and operates strictly within-channel.

Challenging evaluation paradigm: concurrent real speech mixed at an SNR of 0 dB.

Good segregation: minimal loss of target energy; SNR improved by a factor of 3; high ASR accuracy on target

Informal listening tests found that target speech extracted by the system was of good quality.

AMI is an EU 6th FWP IST Integrated Project (project FP6-506811).

[1] 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.*, 117:3069–3078, 2005.
[2] J. Bird and C. J. Darwin, "Effects of a difference in fundamental frequency in separating two sentences," in Psychophysical and physiological advances in hearing, Palmer *et al.* Eds., pp. 263–269. Whurr, 1997.
[3] P. A. Cariani. Recurrent timing nets for auditory scene analysis. In *Proc. Intl. Conf. on Neural Networks (IJCNN)*, 2003.