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

www.dcs.shef.ac.uk/~stu

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

The University Of Sheffield.

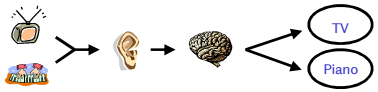
AMI Consortium

www.amiproject.org

Making sense of sounds.

Human separation of multiple sound sources 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). Grouping uses cues (e.g., periodicity) to fuse discrete sensory elements.



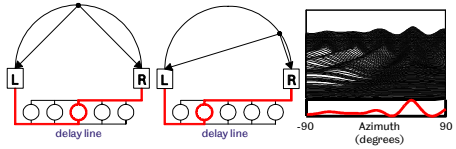
Computational approaches commonly represent cues in distinct feature spaces. How are these associated?

In our approach, cues are inherently associated. Hence, it's easy to represent (and separate) multiple sounds.

Acoustic cues.

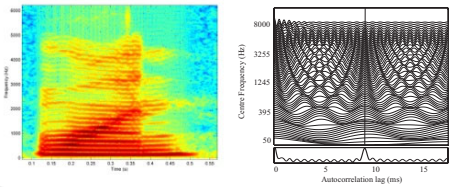
Interaural time difference (ITD)

- Determines the direction of a sound source.
- 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 ITD.
- Rather, differences in ITD are exploited independently within each frequency channel[1].



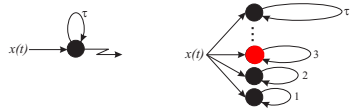
Harmonicity

- Fundamental (F0) and a number of related harmonics.
- Auto-correlation at each frequency channel.
- Merge across frequency: 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].

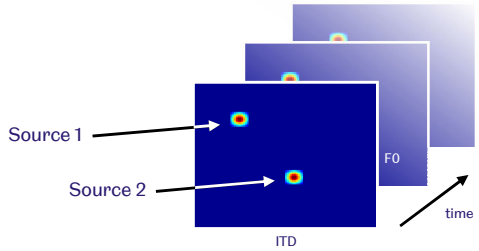


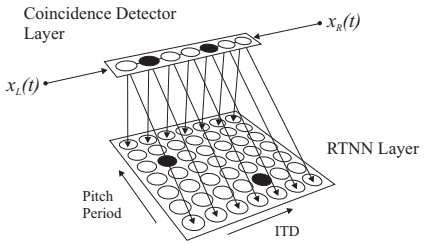
Recurrent timing neural network (RTNN).

- Coincidence detectors: one input is incoming stimulus response, 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 add extra layer of delay line coincidence detectors: ITD cue.
- RTNN becomes 2D: each ITD lag node feeds information to one column; each column is a standard 1D RTNN.

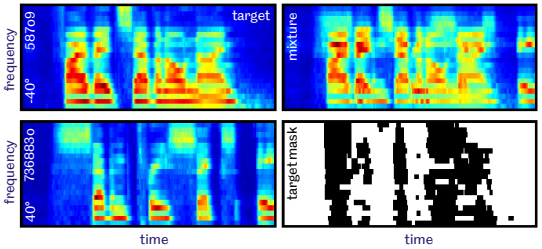




- This processing occurs for every frequency channel.
- Average of previous 25ms activity calculated every 5ms.

Mask generation for resynthesis and missing data ASR.

- 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.
- 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.
- Mask used to resynthesise separated target for energy-based evaluation or for use directly with missing data ASR.



Evaluation.

- 100 randomly selected male utterance pairs from *TIdigits*; 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.

1. Percentage of target speech excluded from the segregated speech (P_{EL}) and percentage of interferer included (P_{NR}).
2. Target SNR improvement.
3. Missing data ASR performance improvement.

- All approaches assumed target was on left.
- 1 & 2 use resynthesised target speech using the binary mask.

	10°	20°	40°	Average
SNR (dB) pre processing	1.64	3.13	5.19	3.32
SNR (dB) RTNN (higher better)	10.03	11.55	14.49	12.02
SNR (dB) *a priori* (higher better)	12.35	13.27	15.01	13.54
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

Missing data ASR

- 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).

- Features: auditory rate maps.
- 12 word-level HMMs (silence, 'oh', 'zero' and '1' to '9').
- Each HMM: 18 no-skip, straight-through states with observations modelled by a 12 component diagonal Gaussian mixture.

	10°	20°	40°	Average
ASR Acc. (%) pre processing	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

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 0dB.

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.

[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. IJCNN*, 2003.

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