Exploiting dendritic autocorrelogram structure to identify spectro-temporal regions dominated by a single sound source

Ning Ma, Phil Green and André Coy

Department of Computer Science,
University of Sheffield, Sheffield, UK
{n.ma, p.green, a.coy}@dcs.shef.ac.uk

10.21437/Interspeech.2006-236

Abstract

Autocorrelograms exhibit tree-like structures whose spines are located at a delay of 1/F0. This paper exploits the dendritic autocorrelogram structure for the identification of spectro-temporal regions dominated by a single periodic sound source in monaural acoustic mixtures. Each frame of the mixture is first segmented into different sound sources in the autocorrelogram domain. Local pitch estimates are formed for each source and used as a cue for temporal integration. A confidence score is computed for each time-frequency pixel in the grouped regions to determine its probability of belonging to the group. The system is evaluated using simultaneous speech in a coherence measuring experiment and also employed within an ASR system where it produces improved results for the Interspeech 2006 Speech Separation Challenge.

Index Terms: speech separation, correlogram, multipitch tracking.

1. Introduction

In realistic listening conditions speech is often corrupted by other sound sources. Many systems have been proposed to separate noise from the speech (e.g. blind source separation based methods), but they often fail on single-channel signals. However, human listeners are adept at extracting target sound sources from monaural acoustic mixtures. It is believed that there are processes in the auditory system that segregate the acoustic evidence into streams based on their characteristics. This ability has motivated extensive research into the perceptual segregation of sound and has resulted in much theoretical and experimental work in auditory scene analysis (ASA) [1].

Several models have been proposed to separate simultaneous sounds using the autocorrelogram (ACG) in which the periodicity of sound is well represented (e.g. [2, 3]). The autocorrelogram (or correlogram) is a 3-D volumetric function mapping a frequency channel of a periphery model, temporal autocorrelation delay, and time to the amount of periodic energy in that channel at that delay and time. Most methods have been based on inspection of a 'pooled' (or summary) correlogram obtained by summing the correlogram across all frequency channels. It is believed that the position of the strongest peak in the pooled ACG corresponds to the fundamental frequency (F0) of the strongest sound source which dominates the energy in those channels that respond to this F0. Once those channels are removed the strongest peak in the residue suggests the F0 of a second (and weaker) source. One limitation of these methods is that locating peaks in the pooled ACG is often difficult when speech is corrupted by competing sounds. Another limitation is that thy cannot account for the effect that the channels dominated by the weaker source may also respond to the F0 of the stronger source, e.g. when the weaker sound has twice the F0 of the stronger sound; thus all channels will be assigned to one source. There are also methods which make use of the entire correlogram for sound source separation. Summerfield *et.al.* [4] discussed a convolution-based strategy for separating simple synthesised vowels with F0 not harmonically related in the correlogram. By locating a tree-like structure in the correlogram, they demonstrated that multiple fundamentals can be recognised.

The paper presents a sound source separation system which exploits the 'dendritic structure' in the correlogram to identify spectro-temporal regions dominated by a single periodic sound source (referred to as 'coherent fragments' in this work) in a monaural acoustic mixture. The 'dendrites' are tree-like structures whose stems, or 'spines' are centred on the delay of multiple pitch periods across frequency channels in a correlogram (see Fig. 1). The system operates by first identifying the dendritic structures and segregating each 10 ms frame of the mixture into different sound sources in the correlogram domain. Local pitch estimates are then formed for each source after the segmentation and used as cues for temporal grouping. This process results in the spectro-temporal representation of speech mixtures being separated into a set of coherent fragments.

We evaluated the system using a small vocabulary simultaneous speech separation task. A coherence measuring experiment was performed to validate the quality of the fragments generated. Section 2 describes the techniques used in coherent fragments generation. Section 3 introduces a confidence score for each spectro-temporal pixel in a fragment, which represents the probability of the pixel belonging to the fragment. In section 4 we evaluate the system and discuss the experimental results. Section 5 concludes and presents future research directions.

2. Coherent Fragment Generation

2.1. Overview

The summary correlogram is not the only way to extract the pitch period. For periodic sounds all autocorrelation channels respond to the fundamental frequency forming a vertical spine in the correlogram which is centred on the delay corresponding to the pitch period. Meanwhile, because the channels also actively respond to the harmonics closest to their centre frequencies (CF), the filtered signal in each channel tends to repeat itself at intervals of approximately 1/CF, giving a succession of peaks at the frequency of its CF in the correlogram. This demonstrates a symmetric tree-

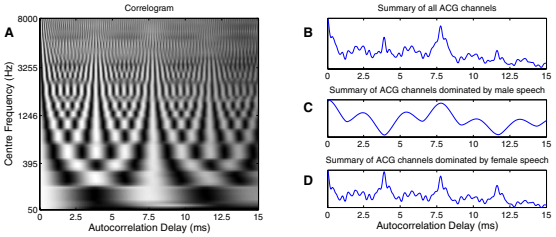

Figure 1: ***(A)*** *Correlogram of a male/female speech mixture.* ***(B)*** *The summary of all ACG channels.* ***(C)*** *The summary of the channels dominated by male speech source.* ***(D)*** *The summary of the channels dominated by female speech.*

like structure centred on the delay of multiple pitch periods in the correlogram, which we refer to as the 'dendritic structure'(see Fig. 1(A)). When only one harmonic source is present, the spine of the dendritic structure is displayed coherently across the entire frequency range. When two (or more) periodic sound sources are present, there may be gaps on a spine as some channels form a spine on the delay of the pitch period of the other source. These cues are employed in this study to separate sound sources.

The correlogram is generated by passing acoustic signals through a 64-channel overlapping gammatone filter bank distributed in their centre frequencies (CF) between 50Hz and 8000Hz on the equivalent rectangular bandwidth (ERB) scale [6]. They are then half-wave rectified and short-time autocorrelation is computed on the output of each filter using a 30ms Hanning window, with a frame shift of 10 ms. This process produces a 2-D correlogram for each frame. Fig. 1(A) illustrates the correlogram of a male/female speech mixture. It has been normalised and plotted as an image for illustration. The pitch periods of the male and female speech are 7.8 ms and 3.9 ms, respectively. The pooled correlogram is displayed in Fig. 1(B). The strongest peak in the summary is on the delay of 7.8 ms which is the pitch period of male speech and all the channels respond to the peak as the female pitch period is half of the male pitch period. When only looking at the strongest peak in the summary all channels may be grouped together.

It is visually clear that there are three dendritic structures in the correlogram centred on the delay of multiple periods of 3.9 ms. This is a strong indication that there is a sound source with a pitch period of 3.9 ms. There are, however, gaps on the spine of the left-most structure, around CF of 100 Hz and 395 Hz, suggesting that some channels belong to another source. These channels are actually dominated by the male speech. Fig. 1(C,D) display the summary of channels dominated by male speech only and female speech only, respectively. If we can locate the left-most dendritic structure the task of separation here is much easier.

2.2. Locating the dendritic structure

The strategy explored here is derived from work reported by Summerfield *et.al* [4]. The correlogram is convolved with a two-dimensional operator which consists of five sinusoids, each weighted by a Gaussian (i.e. five Gabor functions). The Gabor function is defined as:

$$g(x;T,\sigma) = e^{-x^2/2\sigma^2}\cos(2\pi x/T) \qquad (1)$$

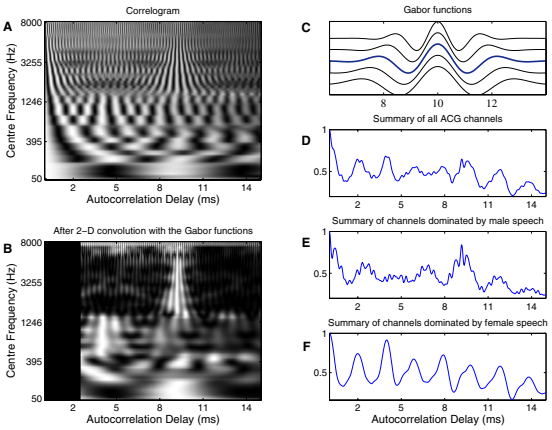

Figure 2: ***(A)*** *Correlogram of a male/female speech mixture.* ***(B)*** *2-D convolution result of the correlogram with Gabor functions.* ***(C)*** *Gabor functions for a particular channel.* ***(D)-(F)*** *Summary of all channels and those dominated by male and female speech source, respectively.*

where $1/T$ is the frequency of the sinusoid and σ is the standard deviation of the Gaussian. The frequency of each sinusoid used by Summerfield *et.al* is the CF of the channel with which it is aligned, and the standard deviation of the Gaussian is 1/CF. This works well with the synthesised vowels used by them. However, speech signals are only quasi-periodic and filter channels only respond to a particular frequency component close to its CF. This means the repeating frequency of the filtered signal in each channel is often off its CF and sometimes the shift is significant. Therefore we measure the actual repeating period p_i in each channel i. This is done by locating the first valley (v_i), the first and second peaks (p_i' and p_i'') of the correlation function in a channel. The actual repeating period p_i is approximated as:

$$p_i = \frac{2v_i + p_i' + p_i''/2}{3} \qquad (2)$$

We use $p_i/2$ as the standard deviation of the Gaussian. These changes have been very effective for realistic speech signals.

The operator approximates the local shape of the dendritic structure at the channel with which the middle Gabor is aligned (see Fig. 2(C)). For each channel c, a 2-D convolution of its operator and corresponding sub-band correlogram (five channels) is performed:

$$A(i,\tau) \otimes g(x;p_i,p_i/2) \quad \text{for } c-2 \le i \le c+2, 1 \le \tau \le L \quad (3)$$

where $A(i,\tau)$ is the autocorrelation function and L is the maximum autocorrelation delay. The central part of the convolution is saved as the result of channel c. When the operator is aligned with the spine of a dendrite, the 2-D convolution gives a large product, and the product is smaller if misaligned. To remove ripples produced when the cosine operator is aligned with other peaks rather than a spine, we also convolve the correlogram with an operator composed of sine functions [4]. At each point the results of the two convolutions are squared and summed, and the relationship $\sin^2 + \cos^2 = 1$ ensures a smooth function with peaks located on the spines of the dendritic structure. The final result of these 2-D convolutions is a simplified correlogram in which the

spines of dendritic structures are greatly enhanced, as illustrated in Fig. 2(B). The two white vertical lines (one around 4ms and the other around 9ms) are the spines of major dendritic structures in the correlogram.

A simple peak-picker selects the strongest peak in each channel. A histogram of the delay periods is computed and the two periods with most counts are selected as the locations of two dendrites. We also make sure that each dendritic structure identified is across a minimum number of channels (4 channels in this study). When the two dendrites are harmonically related (e.g. one pitch period is half of the period of the other one) the one with longer pitch period is removed. Here we assume the maximum number of periodic sources in each frame is two, but this technique can be easily extended to handle more sources.

2.3. Grouping channels across frequency

Once the dendritic structures are found, the correlogram is naturally separated into two sources by grouping channels associated with each spine. However, it is possible that some channels remain isolated. This happens when the energy in a channel is equally dominated by both sources. We assign each isolated channel to a source if it more closely matches the periodicity of that channel. When only one dendritic structure is identified in the correlogram, an isolated channel is only assigned to the source if its period matches within 5% of the source pitch period. In this case the remaining channels are grouped together. Fig. 2(E) and (F) show the summaries of grouped correlogram channels after the separation. The F0 peak is very clear in each summary, while locating them is more difficult in the pooled correlogram (Fig. 2(D)).

2.4. Temporal integration

After grouping frequency channels in each frame, segmentations are integrated across time to form coherent fragments. The stronger source can swap from frame to frame as speech energy varies over time. We use pitch continuity as a cue for integration [1]. At each frame local pitch estimates are formed by summing the ACG channels of each source. For the stronger source only the highest peak is picked. For the weaker source (if there is one) up to two peaks are picked. Fig 3(C) shows the pitch estimates for a female(target)/male(masker) speech mixture. The dots represent the pitch of the stronger source at each frame and the crosses represent the weaker source. It can be seen that the sources swap position across time.

The pitch estimates are then passed to a multipitch tracker to form smooth tracks. The tracker models the pitch of each source as an HMM with one voiced state and an unvoiced one [7]. The smooth pitch tracks are displayed as circles in Fig 3(D), while the ground-truth pitch tracks on the pre-mix clean signals (using the 'Praat' package) are displayed as solid lines. Our system correctly tracks most pitch points. Segmentations with a same pitch track are given a same label. Fig 3(E) shows the final formed fragments, each of which is represented using a different shade of grey. The target-to-masker ratio (TMR) here is 0dB.

2.5. Adding inharmonic fragments

The missing pixels in Fig 3(E) are either low energy regions or inharmonic regions which the proposed system currently does not handle. There are no dendritic structures in a correlogram if the frame is dominated by inharmonic sound. Following [8] we employ the Watershed algorithm to process the spectrogram as an image after the harmonic regions are removed. The inharmonic fragments are then combined with harmonic fragments (Fig. 3(F)).

3. Confidence Measures

One flaw in this system is that the fragment labelling is discrete. This means that each spectro-temporal pixel in a fragment is treated either wholly missing or wholly present. To 'soften' the discrete decision, for each pixel we introduce as a confidence score its probability of belonging to the fragment. This produces a soft mask

The difference between the period of each corresponding ACG channel and the global pitch period measured for that fragment at that time is computed and converted into a score between 0.5 and 1 using a sigmoid function. Channels with a period fully aligned with the pitch period are given a value close to 1; a value close to 0.5 if fully misaligned. Confidence scores for the inharmonic fragments in our study are all set to 1. These confidence scores (as 'soft' masks) are employed along with the discrete fragments in the 'speech fragment decoding' system [5]. They are also used in our coherence evaluation experiment.

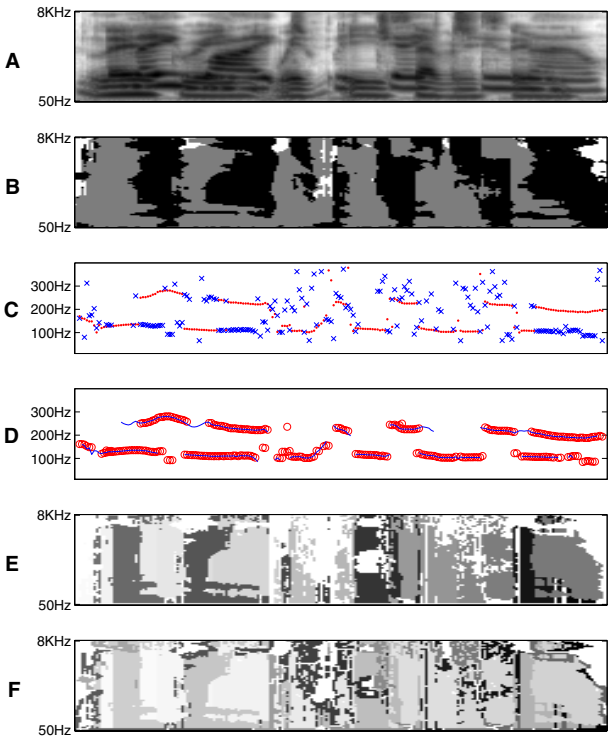

Figure 3: ***(A)*** *A 'ratemap' representation of the mixture of 'lay white with j 2 now' (target, female) plus 'lay green with e 7 soon' (masker, male) TMR = 0dB.* ***(B)*** *The 'oracle' segmentation: dark grey: the value in the mixture is close to that in the female speech; light grey: the mixture value is close to that in the male speech.* ***(C)*** *Pitch estimates for each source segmentation. Dots represent the pitch of the stronger source of each frame and crosses represent the weaker source of that frame.* ***(D)*** *Circles are pitch tracks produced by the multipitch tracking algorithm; solid lines are the ground-truth pitch tracks.* ***(E)*** *Fragments after temporal integration based on the smooth pitch tracks.* ***(F)*** *Combining inharmonic fragments.*

4. Experiments and Results

4.1. Coherence measuring

A natural criterion for evaluating the quality of fragments is to measure their coherence. The coherence of a fragment is referred to as its consistency with a single source. If each pixel is associated with a weight w, we define the coherence as:

$$100 \times \frac{\sum w_1}{\sum w_1 + \sum w_2} \quad (4)$$

where w_1 are a set of weights for pixels in the fragment overlapping the majority source and w_2 are a set of weights for those which overlap the minority source. In this study we use the confidence scores described in Sec. 3 as the weights. The fragments are compared with the 'oracle' segmentation obtained from pre-mix signals. Note that small fragments are less likely to have pixels that belong to different sources than large fragments. To reduce the effect of small fragments contributing high coherence to the overall scores, we did not include fragments with less than 20 pixels. The average coherence of these small fragments is higher than 82%.

Table 1: *Average sizes of fragments with different coherence*

coherence	lower than 80%	higher than 80%
Proposed system	209 pixels	285 pixels
Previous system	348 pixels	311 pixels

4.2. Results and discussion

The experiment was performed using simultaneous speech constructed from the Grid corpus [9]. The test set consists of 600 pairs of endpointed utterances by 34 speakers that have been artificially added together at a range of TMRs; 200 pairs in which target and masker are the same speaker, 200 pairs of the same gender (but different speakers), and 200 pairs of mixed gender.

The histograms of coherence scores of fragments generated by the proposed system and the system reported in [8] are displayed in Fig 4. They have been normalised by dividing the counts in the bins by the total number of fragments. It shows that the coherence of fragments generated by the system reported here is significantly higher than the previous system throughout different conditions. To examine the effect of fragment sizes on the fragment coherence, we also measured the average size of fragments with coherence lower and higher than 80% (Tab. 1). The sizes of fragments whose coherence is higher than 80% are not significantly different in both systems while the proposed system produces smaller fragments with low coherence. This is acceptable as it also produces proportionally less fragments with low coherence. The aim is to produce more large fragments with high coherence.

The technique proposed here was also employed within the ASR system reported in [5], producing improved results for the Interspeech 2006 Speech Separation Challenge[1].

5. Conclusions

This paper presents a novel approach which exploits the dendritic structure in correlograms to identify spectro-temporal regions dominated by a single sound source in monaural acoustic mixtures. The fragments generated in this way are more coherent than a previous method. Future work includes comparing correlograms across time to produce better segmentation. We will also investigate a more robust multipitch tracker.

[1] http://www.dcs.shef.ac.uk/∼martin/SpeechSeparationChallenge.htm

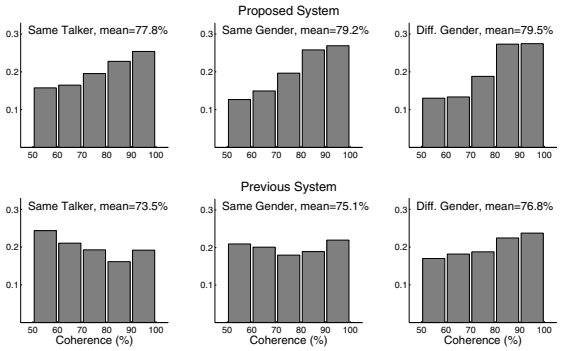

Figure 4: *Histograms of fragments coherence scores after normalisation (TMR = 0dB). The top three are for fragments generated by the proposed system while the bottom three are for fragments generated by the system reported in [8].*

6. Acknowledgements

We thank Guy Brown for discussion on the dendritic correlogram structure. Ning Ma's doctoral studies were supported by EPSRC grand GR/R47400.

7. References

[1] A.S. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge MA, 1990.

[2] R. Meddis and M.J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Amer.*, vol. 91, no. 1, pp. 233–245, 1992.

[3] Wang D. and G. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, May 1999.

[4] Q. Summerfield, A.P. Lea, and D. Marshall, "Modelling auditory scene analysis: strategies for source segregation using autocorrelograms," in *Proc. Institute of Acoustics*, 1990, vol. 12, pp. 507–514.

[5] J. Barker, A. Coy, N. Ma, and M Cooke, "Recent advances in speech fragment decoding techniques," in *Proc. Interspeech'06*, accepted.

[6] M.P. Cooke, *Modelling auditory processing and organisation*, Ph.D. thesis, Department of Computer Science, University of Sheffield, 1991.

[7] A. Coy and J. Barker, "A multipitch tracker for monaural speech segmentation," in *Proc. Interspeech'06*, accepted.

[8] A. Coy and J. Barker, "Soft harmonic masks for recognising speech in the presence of a competing speaker," in *Proc. Interspeech'05*, Lisbon, 2005, pp. 2641–2644.

[9] M.P. Cooke, J. Barker, S.P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, submitted.