Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 874–891
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Exploiting correlogram structure for robust speech recognition with multiple speech sources

Ning Ma *, Phil Green, Jon Barker, André Coy

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK

Received 19 January 2007; received in revised form 11 May 2007; accepted 11 May 2007

Abstract

This paper addresses the problem of separating and recognising speech in a monaural acoustic mixture with the presence of competing speech sources. The proposed system treats sound source separation and speech recognition as tightly coupled processes. In the first stage sound source separation is performed in the correlogram domain. For periodic sounds, the correlogram exhibits symmetric tree-like structures whose stems are located on the delay that corresponds to multiple pitch periods. These pitch-related structures are exploited in the study to group spectral components at each time frame. Local pitch estimates are then computed for each spectral group and are used to form simultaneous pitch tracks for temporal integration. These processes segregate a spectral representation of the acoustic mixture into several time–frequency regions such that the energy in each region is likely to have originated from a single periodic sound source. The identified time–frequency regions, together with the spectral representation, are employed by a 'speech fragment decoder' which employs 'missing data' techniques with clean speech models to simultaneously search for the acoustic evidence that best matches model sequences. The paper presents evaluations based on artificially mixed simultaneous speech utterances. A coherence-measuring experiment is first reported which quantifies the consistency of the identified fragments with a single source. The system is then evaluated in a speech recognition task and compared to a conventional fragment generation approach. Results show that the proposed system produces more coherent fragments over different conditions, which results in significantly better recognition accuracy.

Keywords: Speech separation; Robust speech recognition; Multiple pitch tracking; Computational auditory scene analysis; Correlogram; Speech fragment decoding

1. Introduction

In realistic listening conditions, speech is often corrupted by competing sound sources. The presence of acoustic interference can cause the quality or the intelligibility of speech to degrade; the performance in automatic speech recognition (ASR) often drops dramatically. Many systems have been proposed to separate noise from speech using cues from multiple sensors, e.g. blind source separation by independent component analysis (Parra and Spence, 2000), but separating and recognising speech in single-channel signals, the problem considered in this article, still remains a challenging problem. Human listeners, however, are adept at recognising target speech in such noisy conditions, making use of cues such as pitch continuity, spacial location, and speaking rate (Cooke and Ellis, 2001). They are able to effectively extract target audio streams from monaural acoustic mixtures with little effort, e.g. listening to speech/music mixtures on a mono radio program. It is believed that there are processes in the auditory system that segregate the acoustic evidence into perceptual streams based on their characteristics, allowing listeners to selectively attend to whatever stream is of interest at the time (Bregman, 1990; Cooke and Ellis, 2001). This offers an alternative to techniques which require the noise to be effectively removed from the speech, e.g. spectral subtraction based methods (Lim et al., 1979), and allows the noise

* Corresponding author. Tel.: +44 114 222 1878; fax: +44 114 222 1810.
E-mail addresses: n.ma@dcs.shef.ac.uk (N. Ma), p.green@dcs.shef.ac.uk (P. Green), j.barker@dcs.shef.ac.uk (J. Barker), a.coy@dcs.shef.ac.uk (A. Coy).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.05.003

to be treated as streams that can be ignored while the target speech is attended to.

This ability of listeners has motivated extensive research into the perceptual segregation of sound sources and has resulted in much theoretical and experimental work in *auditory scene analysis* (ASA) (Bregman, 1990). Auditory scene analysis addresses the problem of how the auditory system segregates the mixture of sound reaching the ears into packages of acoustic evidence in which each package is likely to have been produced from a single source of sound. The analysis process, described by Bregman (1990), is interactively governed by 'primitive' bottom-up grouping rules, which are innate constraints driven by the incoming acoustic data and the physics of sound, and 'schema-based' top-down constraints, which employ the knowledge of familiar patterns that have been learnt from complex acoustic environments. Computational auditory scene analysis (CASA) aims to develop computational models of ASA. Many researchers have proposed automatic sound separation systems based on the known principles of human hearing and have achieved some success (Brown and Cooke, 1994; Wang et al., 1999; Ellis, 1999). A good review of CASA development is reported in (Brown and Wang, 2005).

1.1. Correlogram-based CASA models

One important representation of auditory temporal activity that combine both spectral and temporal information is the autocorrelogram (ACG). The autocorrelogram, or simply *correlogram*, is a three-dimensional volumetric function, mapping a frequency channel of an auditory periphery model, temporal autocorrelation delay (or lag), and time to the amount of periodic energy in that channel at that delay and time. Correlograms are normally sampled across time to produce a series of two-dimensional graphs, in which frequency and autocorrelation delay are displayed on orthogonal axes. Fig. 1 shows three correlograms of a clean speech signal uttered by a female speaker, taken at time frames of 300 ms, 700 ms and 2100 ms. Each correlogram has been normalised and plotted as an image for illustration. The periodicity of sound is well represented in the correlogram. If the original sound contains a signal that is approximately periodic, such as voiced speech, then each frequency channel excited by that signal will have a high similarity to itself delayed by the period of repetition. The ACG frequency channels also all respond to the fundamental frequency ($F0$) and this can be emphasised by summing the ACG over all frequency channels, producing a 'summary ACG' (see the bottom panel in Fig. 1). The position of the largest peak in the summary ACG corresponds to the pitch of the periodic sound source. Primarily because it is well-suited to detecting signal periodicity, the correlogram is widely considered as the preferred computational representation of early sound processing in the auditory system.

The correlogram was first suggested as a model for pitch perception by Licklider (1951) in his neural auto-coincidence model, where the concept of subband periodicity detection was discussed. The model was then reintroduced by Slaney and Lyon (1990), among others (e.g. Meddis and Hewitt, 1991), as a computational approach to pitch detection. Slaney and Lyon employed the correlogram computed from the output of a cochlear model to model how humans perceive pitch. The pitch was estimated based on locating the peaks in the summary correlogram. The ACG model has subsequently been extended as a popular mechanism for segregating concurrent periodic sounds and the primary methods have been based on inspection of the summary correlogram. Assmann and Summerfield (1990) reported a place–time model on a concurrent vowel

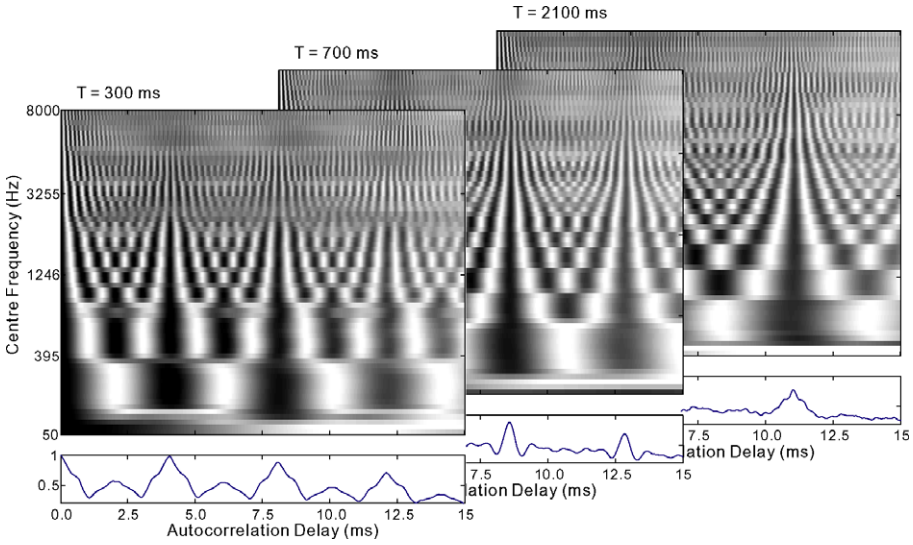

Fig. 1. Three correlograms of a clean speech signal uttered by a female speaker, taken at time frame 300, 700 and 2100 ms, respectively. Each correlogram has been normalised and plotted as an image. A corresponding summary ACG is shown at the bottom of each correlogram.

segregation task. The model estimated the pitch of each vowel as corresponding to the autocorrelation delays with the two largest peaks in the summary correlogram. Meddis and Hewitt (1992) proposed a residual-driven approach. They first selected the largest peak in the summary ACG, the delay of which corresponds to the $F0$ of the stronger sound source. Frequency channels that respond to this $F0$ were grouped and removed from the correlogram. The rest of the channels were integrated together and the largest peak in the residue corresponds to the $F0$ of a second (and weaker) source. Recently, neural oscillator models have been successful at providing accounts of the interaction of cue combinations, such as common onset and proximity (Brown and Cooke, 1994; Wang et al., 1999), in which the summary correlogram model was also employed as a front end.

One limitation of the methods which are based on the summary correlogram is that when speech is corrupted by competing sounds, locating peaks in the summary is often difficult. The position of the largest peak in the summary would not always correspond to the pitch of the target speech and peaks indicating pitches of different sound sources may be correlated. Another limitation is that these models cannot account for the effect of harmonic components of the weaker source being dominated by the stronger source, where all correlogram channels will be assigned to the stronger source (de Cheveigné, 1993). To address these limitations, Coy and Barker (2005) proposed to keep the four largest peaks in the summary ACG as pitch candidates for each time frame and then employed a multi-pitch tracker to form smooth pitch tracks from these candidates. Frequency channels that respond to pitch values in the same pitch track are grouped together. By keeping multiple pitch candidates they show that better sound segregation can be achieved. However, their system relies on a robust multi-pitch tracker and keeping an arbitrary number of pitch candidates is not effective when dealing with different competing sources.

The summary ACG is not the only way to reveal pitch information. The methods based on the summary ACG discard the rich representation of the spectral content and time structure of a sound in the original correlogram. Visually there are clear pitch-related 'dendritic structures' in the correlogram. The 'dendrites' are tree-like structures whose stems are centred on the delay of multiple pitch periods across frequency channels. Slaney and Lyon (1990) discussed this dendritic structure in their perceptual pitch detector. They convolved the correlogram with an operator to emphasise the structure before integrating all ACG channels together. Summerfield et al. (1990) also proposed a convolution-based strategy for the separation of concurrent synthesised vowels with $F0$ not harmonically related in the correlogram. By locating the dendritic structure in the correlogram they demonstrated that multiple fundamentals can be recognised.

1.2. Linking CASA with speech recognition systems

The success of CASA has inspired research into developing a new generation of automatic speech recognition systems for natural listening conditions where competing sounds are often present. In these adverse conditions not all the acoustic evidence from the target source will be recovered. One successful approach to this problem is 'missing data ASR' (Cooke et al., 2001), which adapts the conventional probabilistic formalism of ASR to deal with the 'missing data'. The missing data approach assumes that some acoustic data in the mixture will remain uncorrupted and can be identified as reliable evidence for recognition. Cooke et al. (1997) demonstrated that recognition can indeed be based on a small amount (10% or less) of the original time–frequency 'pixels' if they can be correctly identified.

The limitation of the missing data approach is that accurate identification of target acoustic evidence is a challenging problem and recognition performance is poor if the 'missing data' is not correctly identified. There is evidence that listeners make use of both primitive and schema-based constraints when perceiving speech signals (Bregman, 1990). Barker et al. (2005) proposed a speech fragment decoding (SFD) technique which treats segregation and recognition as coupled problems. Primitive grouping processes exploit common characteristics to identify sound evidence arising from a single source. Top-down search utilises acoustic models of target speech to find the best acoustic combinations which jointly explain the observation sequence without deciding the identity of different sources. The SFD technique therefore provides a bridge

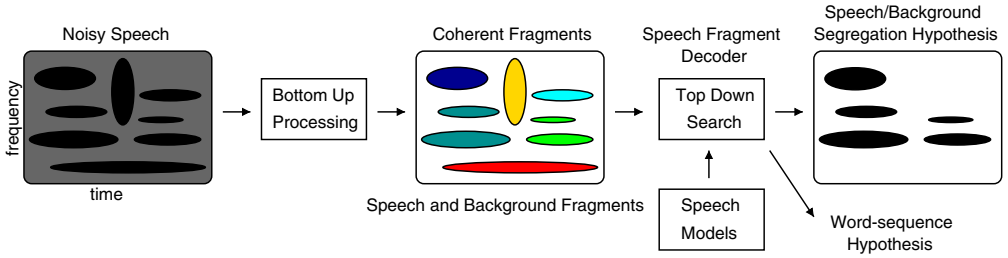

Fig. 2. An overview of the speech fragment decoding system (after Barker et al., 2005). Bottom-up processes are employed to identify spectro-temporal regions where each region is likely to have originated from a single source (coherent fragments). A top-down search with access to speech models is then used to search for the most likely combination of fragment labelling and speech model sequence.

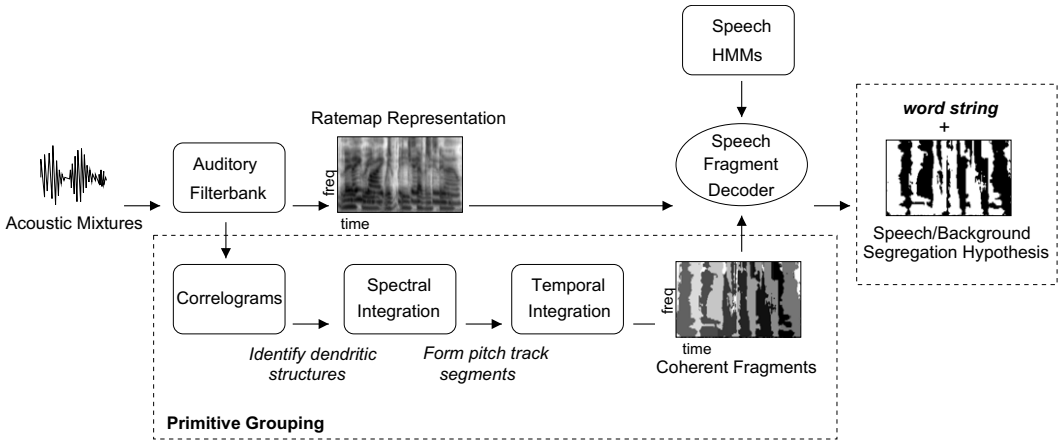

Fig. 3. Schematic diagram of the proposed system.

that links auditory scene analysis models with conventional speech recognition systems. An overview of the SFD system is provided in Fig. 2.

1.3. Summary of the paper

In this article we are concerned with the use of primitive CASA models to address the problem of separating and recognising speech in monaural acoustic mixtures. Some of this work was reported in (Ma et al., 2006). The dendritic correlogram structure is exploited to separate a spectrogram representation of the acoustic mixture into spectro-temporal regions such that the acoustic evidence in each region is likely to have originated from a single source of sound. These regions are referred to as 'coherent fragments' in this study. Some of these fragments will arise from the target speech source while others may arise from noise sources. These coherent fragments are passed to the speech fragment decoder to identify the best subset of fragments as well as the word sequence that best matches the target speech models. We evaluate the system using a challenging simultaneous speech recognition task.[1]

The remainder of this article is organised as follows: in the next section, the overall structure of our system is briefly reviewed. Section 3 describes the techniques used to integrate spectral components in each frame based on the ACG. Section 4 presents methods which produce coherent fragments. Section 5 introduces a confidence map to soften the discrete decision of assigning a pixel to a fragment. In Section 6 we evaluate the system and discuss the experimental results. Section 7 concludes and presents future research directions.

2. System overview

Fig. 3 shows the schematic diagram of our system. The input to the system is a mixture of target speech and interfering sounds, sampled at a rate of 25 kHz. In the first stage of the system, cochlear frequency analysis is simulated by a bank of 64 overlapping bandpass Gammatone filters, with centre frequencies spaced uniformly on the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990) between 50 Hz and 8000 Hz. Gammatone filter modelling is a physiologically motivated strategy to mimic the structure of peripheral auditory processing stage (Cooke, 1991). The gains of the filters are chosen to reflect the transfer function of the outer and middle ears. Having more filters (e.g. 128) can offer a higher frequency resolution but bring more computational cost. The output of each filter is then half-wave rectified.

The digital implementation of the Gammatone filter employed here was based on the implementation of Cooke (1991) using the impulse invariant transformation. The sound is first multiplied by a complex exponential $e^{-j\omega t}$ at the desired centre frequency ω, then filtered with a baseband Gammatone filter, and finally shifted back to the centre frequency region. The cost of computing the complex exponential $e^{-j\omega t}$ for each sound sample t is a significant part of the overall computation. In our implementation, the exponential computation is transformed into simple multiplication to reduce the cost by rearranging $e^{-j\omega t}$ to

$$e^{-j\omega t} = e^{-j\omega} e^{-j\omega(t-1)} \quad (1)$$

The term $e^{-j\omega}$ can be pre-computed and $e^{-j\omega(t-1)}$ is the result of the previous sample $t-1$. Therefore only one complex exponential calculation is needed for the first sample and for the rest of samples the exponentials can be computed by simple multiplication. Experiments showed that using this implementation the Gammatone computation speed can be increased by a factor of 4.

Spectral features are then computed in order to employ the 'speech fragment decoder' (Barker et al., 2005). The instantaneous Hilbert envelope is computed at the output of each Gammatone filter. This is smoothed by a first-order low-pass filter with an 8 ms time constant, sampled at 10 ms intervals, and finally log-compressed to give an

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

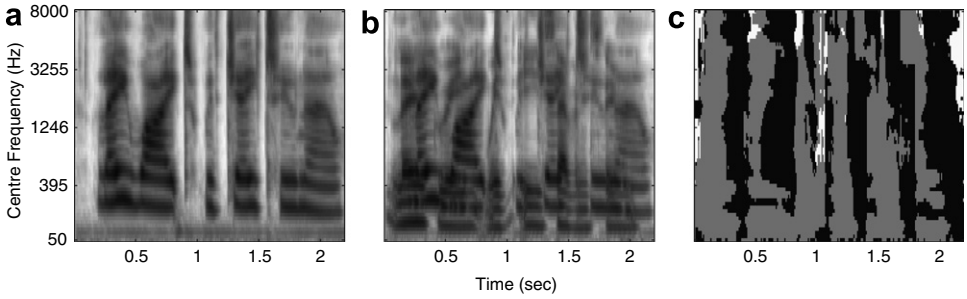

Fig. 4. (a) A 'ratemap' representation for the utterance 'lay white with j 2 now' (target, female) without added masker. (b) Ratemap for the same utterance plus 'lay green with e 7 soon' (masker, male) with a TMR of 0 dB. (c) The 'oracle' segmentation: dark grey – the value in the mixture is close to that in the target female speech; light grey – the mixture value is close to that in the male speech; white pixels – low energy regions.

approximation to the auditory nerve firing rate – a 'ratemap' (Brown and Cooke, 1994). Fig. 4 gives an example of a ratemap representation,[2] for (a) a female utterance 'lay white with j 2 now', and (b) the same utterance artificially mixed with a male utterance 'lay green with e 7 soon', with a target-to-masker ratio (TMR) of 0 dB. Panel c shows the 'oracle' segmentation, obtained by making use of the pre-mix clean signals. Dark grey represents pixels where the value in the mixture is closer to that in the target female speech; light grey represents pixels where the mixture value is closer to that in the male speech; white pixels represent low energy regions. These representations are called 'missing data masks'.

The output of the auditory filterbank is also used to generate the correlograms. A running short-time autocorrelation is computed on the output of each cochlear filter, using a 30 ms Hann window. At a given time step t, the autocorrelation $A(i,t,\tau)$ for channel i with a time lag τ is given by

$$A(i,\tau,t)=\sum_{k=0}^{K-1} g(i,t+k)w(k)g(i,t+k-\tau)w(k-\tau) \qquad (2)$$

where g is the output of the Gammatone filterbank and w is a local Hann window of width K time steps. Here $K=750$ corresponding to a window width of 30 ms. The autocorrelation can be implemented using the efficient fast Fourier transform (FFT), but has the disadvantage that longer autocorrelation delays have attenuated correlation owing to the narrowing of the effective window. We therefore use a scaled form of Eq. (2) with a normalisation factor to compensate for the effect:

$$A(i,\tau,t)=\frac{1}{K-\tau}\sum_{k=0}^{K-1} g(i,t+k)w(k)g(i,t+k-\tau)w(k-\tau) \qquad (3)$$

The autocorrelation delay τ is computed from 0 to $L-1$ samples, where $L=375$ corresponding a maximum delay of 15 ms. This is appropriate for the current study, since the $F0$ of voiced speech in our test set does not fall below 66.7 Hz. We compute the correlograms with the same frame shift as when computing the ratemap features (10 ms), hence each one-dimensional (frequency) ratemap frame has a corresponding two-dimensional (frequency and autocorrelation delay) correlogram frame.

In the stage of spectral integration the dendritic structure is exploited in the correlogram domain to segregate each frame of the mixture into spectral groups, such that the partial spectra in each group is entirely due to a single sound source in that frame. In the next stage local pitch estimates are computed for each group and a multi-pitch tracker links these pitch estimates to produce smooth pitch tracks. Spectral groups are integrated temporally based on these pitch tracks. The processes separate the spectro-temporal representation of the acoustic mixture into a set of coherent fragments, which are then employed in the 'speech fragment decoder', together with clean speech models, to perform automatic speech recognition.

3. Spectral integration based on the ACG

3.1. The dendritic ACG structure

For a periodic sound source all autocorrelation channels respond to $F0$ (i.e. the energy reaches a peak at the same frequency), forming vertical stems in the correlogram centred on the delays corresponding to multiple pitch periods. Meanwhile, because each filter channel also actively responds to the harmonic component that is closest to its centre frequency (CF), the filtered signal in each channel tends to repeat itself at an interval of approximately 1/CF, giving a succession of peaks at approximately the frequency of the CF of each channel in the correlogram. This produces symmetric tree-like structures appearing at intervals of the pitch period in the correlogram (dendritic structures). When only one harmonic source is present, the stem of each dendritic structure extends across the entire frequency range (see the left panel in Fig. 5). The one with the shortest autocorrelation delay is located at the position of the pitch period of the sound source. When a competing sound source

[2] All examples used in this study are utterances from the GRID corpus (Cooke et al., 2006). See Section 6 for detailed explanation.

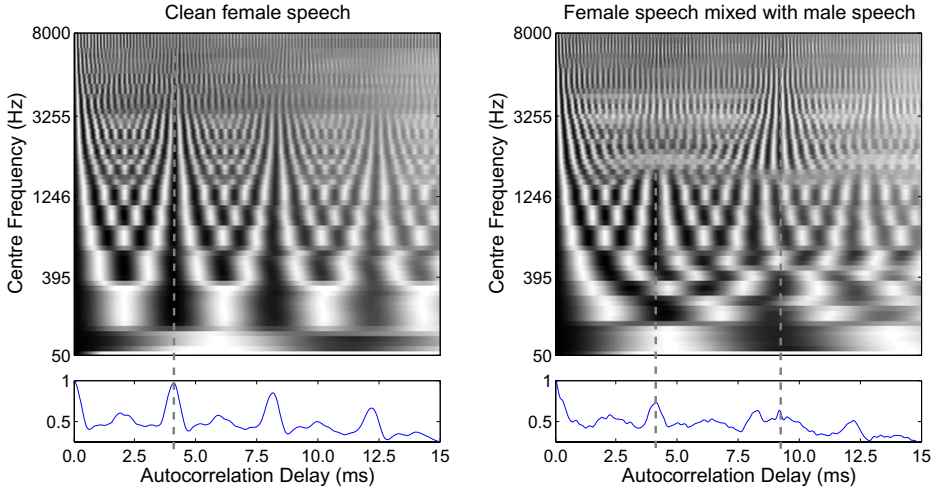

Fig. 5. A comparison of correlograms in clean and noisy conditions. Left – a correlogram and its summary of clean female speech, taken at time frame 60; right – taken at the same frame when the female speech is mixed with male speech at a TMR of 0 dB. The dendritic structures that correspond to the *F*0 of different speech sources are marked using vertical dashed lines. It can be clearly seen that in the noisy condition the dendrites do not extend across the entire frequency range.

is also present, some ACG channels may be dominated by the energy that has arisen from the competing source, causing a gap in the stem of the dendritic structure corresponding to the target source's pitch. If the competing source is also periodic, channels dominated by its energy may also form part of a dendritic structure on the delay of its pitch period.

Fig. 5 compares two correlograms taken at the same time frame of a female speech utterance in either a clean condition (left panel) or when mixed with male speech at a target-to-masker ratio of 0 dB (right panel). The summary ACGs are also shown correspondingly. The dendritic structures which correspond to the *F*0 of sound sources are marked using dashed lines. In the clean condition it is visually clear that the dendritic structure extends across the entire frequency range except those ACG channels whose centre frequency is much below the female speaker's *F*0 (the bottom 5 channels). In the ACG on the right, the dendritic structures corresponding to the two competing speech sources both fail to dominate the whole frequency range. The one extending from 400 Hz to 1300 Hz on the delay of 3.9 ms indicates that there exists a harmonic source with an *F*0 of 256 Hz and the energy of the channels within this range has originated from the female speaker source. The rest of channels form part of another dendritic structure on the delay of 9.0 ms which indicates a second harmonic source with an *F*0 of 111 Hz (the male speaker). This information can be used to separate the two sound sources but is lost in the summary ACG.

3.2. Pre-grouping

ACG channels are pre-grouped before the dendritic structures are extracted in the correlogram. Gammatone filters have overlapping bandwidth and respond to the harmonic with the highest energy. Therefore, ACG channels which are dominated by the same harmonic share a very similar pattern of periodicity (Shamma, 1985). Fig. 5 illustrates this phenomenon. For example, in the left panel channels with a CF between 100 Hz and 395 Hz demonstrate a very similar pattern of periodicity. This redundancy can be exploited to effectively pre-group ACG channels. We employ a cross-channel correlation metric (Wang et al., 1999) where each ACG channel is correlated with its adjacent channel as follows:

$$C(i,t) = \frac{1}{L}\sum_{\tau=0}^{L-1} \widehat{A}(i,\tau,t)\widehat{A}(i+1,\tau,t) \tag{4}$$

where L is the maximum autocorrelation delay and $\widehat{A}(i,\tau,t)$ is the autocorrelation function of Eq. (3) after normalisation to zero mean and unit variance. The normalisation ensures that the cross-channel correlation is sensitive only to the pattern of periodicity of ACG channels, and not to their energy. Channel i and $i+1$ are grouped if $C(i,t) > \theta$. We choose $\theta = 0.95$ to ensure that only ACG channels with a highly similar pattern are grouped together.

A 'reduced ACG' is obtained by summing pre-grouped channels across frequency. Each set of grouped channels is referred to as a 'subband' in the reduced ACG. The pre-grouping significantly reduces computational cost as the average number of ACG subbands is 39 compared to 64 ACG channels originally. Preliminary experiments also show that the process can effectively reduce grouping errors in the later stages.

3.3. Extracting the dendritic structure

The essential idea in this study is to make use of the dendritic structure in the full correlogram for the separation of sound sources. The technique of extracting the pitch-related structure used here is derived from work by

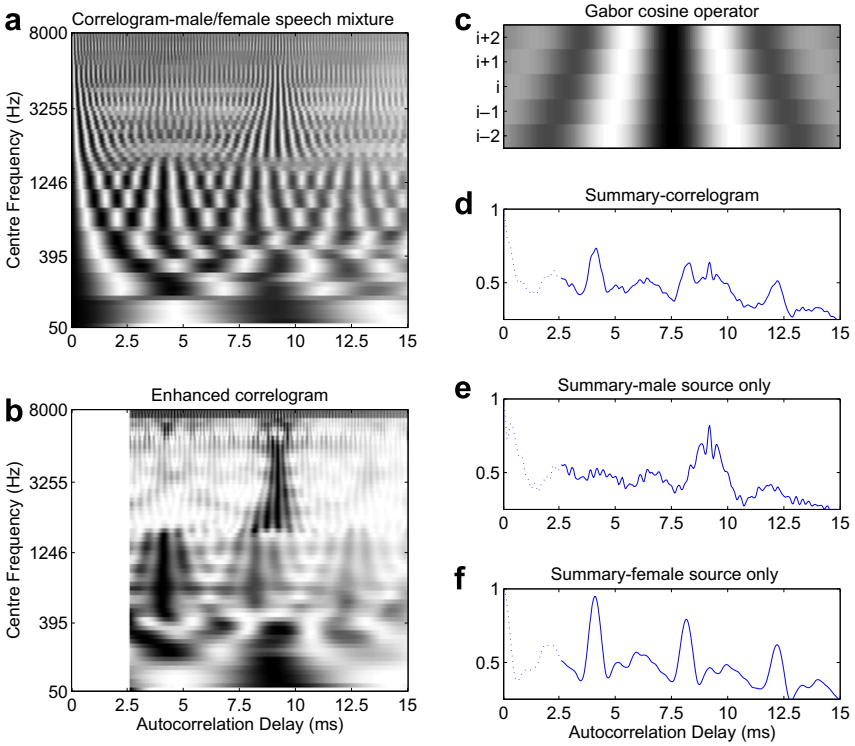

Fig. 6. (a) A correlogram of a mixture of male and female speech. (b) Enhanced correlogram after the 2-D convolution. The region with delays less than 2.5 ms (corresponding to regions with $F0$s higher than 400 Hz) is not computed. (c) An example of a Gabor cosine operator. (d) Summary correlogram. (e and f) Summaries of spectral components in the correlogram dominated by energy from respective speaker sources (male or female). Dotted line represents the high $F0$ region which is not computed.

Summerfield et al. (1990). For each subband in the reduced ACG, a two-dimensional cosine operator is constructed, which approximates the local shape of the dendritic structure around the subband. The operator consists of five Gabor functions applied to adjacent reduced ACG subbands, in which the middle Gabor function is aligned with the subband it operates on (see Fig. 6c). The Gabor function is a sinusoid weighted by a Gaussian. If the sinusoid is a cosine, the Gabor function is defined as

$$\mathrm{gabor}_c(x;T,\sigma) = \mathrm{e}^{-x^2/2\sigma^2}\cos(2\pi x/T) \tag{5}$$

where T is the period of the sinusoid and σ is the standard deviation of the Gaussian. The frequency of each sinusoid used by Summerfield et al. is the centre frequency 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 in their study. However, speech signals are only quasi-periodic and a filter channel responds to a frequency component that is only an approximation to its CF. Therefore the repeating frequency of the filtered signal in each ACG channel is often off its CF depending on how close the nearest harmonic is to the CF, and sometimes the shift is significant. Therefore in our study we compute the actual repeating period p_i in each ACG subband i by locating the first valley (v_i) and the first and second peaks (p_i' and p_i'') of the autocorrelation function. The repeating period p_i of subband i is approximated as

$$p_i = \frac{2v_i + p_i' + p_i''/2}{3} \tag{6}$$

To further enhance the dendrite stem $p_i/2$ is used as the standard deviation in the Gabor function, a value roughly half that used by Summerfield et al. These changes have been very effective with realistic speech signals.

The autocorrelation function $A(i,\tau,t)$ for each subband i, with support of its four adjacent subbands (two above and two below), is convolved with its corresponding two-dimensional cosine operator after zero-padding, producing an initial enhanced autocorrelation function $A_c(i,\tau,t)$:

$$A_c(i,\tau,t) = \sum_{m=-2}^{2}\sum_{n=1}^{L} A(i+m,\tau+n,t)\mathrm{gabor}_c(n;p_{i+m},p_{i+m}/2) \tag{7}$$

where L is the maximum autocorrelation delay. The central part of the convolution is saved for each subband.[3] When the operator is aligned with the stem of a dendrite, the convolution gives a large product, and the product is smaller if misaligned. Unfortunately, ripples will occur as the cosine operator will also align with peaks other than the stem. Following Summerfield et al. (1990), these ripples are

[3] In practice the two-dimensional convolution is computed using the MATLAB function *conv2*.

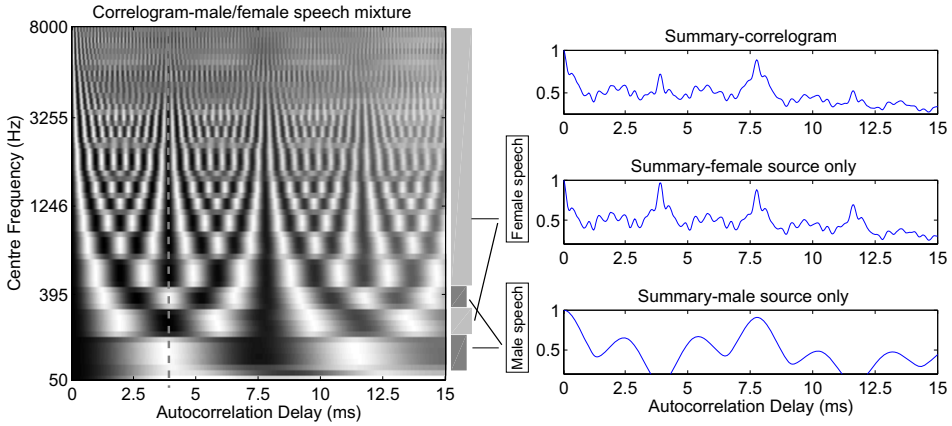

Fig. 7. A correlogram of a mixture of male and female speech. The $F0$ of the male speaker is half of that of the female speaker. Subbands dominated by the energy from different speaker sources are indicated using different shades of grey. The dendritic structure with the shortest delay caused by the female source is marked using a dashed vertical line. Summary of all ACG subbands and those dominated by energy from the female and the male speaker source are shown respectively on the right.

removed using a sine operator constructed by substituting the cosine function in Eq. (5) for a sine function:

$$\text{gabor}_s(x;T,\sigma) = \mathrm{e}^{-x^2/2\sigma^2}\sin(2\pi x/T) \tag{8}$$

The original correlogram is convolved with the sine operators to generate a function $A_s(i,\tau,t)$ in the same manner as in Eq. (7). At each point the results of the two convolutions are squared and summed, producing a final autocorrelation function $A_e(i,\tau,t)$ with the peak in each subband located on the stem of the dendritic structure:

$$A_e(i,\tau,t) = A_c(i,\tau,t)^2 + A_s(i,\tau,t)^2 \tag{9}$$

In the enhanced correlogram, A_e, the stems of dendritic structures are greatly emphasised, as illustrated in Fig. 6b. The correlogram is computed for a frame in which a female speaker source is present simultaneously with a male speaker source. The two black vertical lines in the enhanced correlogram (one around 3.9 ms and the other around 9.0 ms) are the stems of two dendritic structures which correspond to the two speaker sources. To reduce computational cost, regions with autocorrelation delays less than 2.5 ms (corresponding to regions with $F0$ higher than 400 Hz outside the speech $F0$ range) are not computed.

The largest peak in each subband in the enhanced correlogram is selected and a histogram with a bin width equivalent to 3 Hz is computed over these peak positions. The two highest-counting bins indicate the locations of two possible dendrites corresponding to two harmonic sources. A bin is ignored if its count is less than an empirically determined threshold (5 in this study), therefore in each frame 2, 1 or 0 dendritic structures are found.[4]

[4] This technique can be extended to handle more sources provided the maximum number of simultaneous periodic sources at each frame is known.

3.4. Spectral grouping

Once the dendritic structures are extracted from the correlogram, the frequency bands can be divided into partial spectra: the ACG subbands with their highest peak at the same position in the enhanced ACG are grouped together. Each group of subbands therefore form an extracted dendritic structure. The number of simultaneous spectral groups depends on the number of dendrites identified. If no such structures appear in the correlogram (e.g. for an unvoiced speech frame), the system skips the frame and no spectral group is generated.

After this grouping it is still possible that some ACG subbands remain isolated. Although this is rare, it could happen because a subband may respond to a different dendrite from the one formed by its adjacent subbands. Therefore the subband will not be emphasised in the enhanced ACG. When only one spectral group is formed, an isolated subband is assigned to the group only if it matches the periodicity of the subband within a threshold of 5% in the original ACG. When two spectral groups are formed, an isolated subband is assigned to the group which better matches its periodicity within the threshold of 5%.

This spectral integration technique has the ability to deal with the situation where the fundamentals of two competing speakers are correlated. Fig. 7 shows a correlogram computed for a frame in which a male speaker source with a pitch period of 7.8 ms is present simultaneously with a female speaker source with a pitch period approximately half of that (3.9 ms). Since the subbands dominated by the energy from the female source have peaks at an interval of 3.9 ms in the ACG, all the subbands have peaks at the delay of 7.8 ms, causing the largest peak in the summary ACG to occur at that delay. When the summary ACG is inspected, it is difficult to group subbands as they all respond to the largest peak. However, the female speech subbands will form a partial dendritic structure (marked

using a dashed vertical line). The white gaps in its stem clearly indicate that subbands within these gaps do not belong to the female source as otherwise the dendrite would extend across the entire frequency range. Those subbands are actually dominated by the energy from the male speaker source. By exploiting the dendritic structure, a more reliable separation of sources with correlated fundamentals can be performed. Fig. 7 also shows the summary of ACG subbands dominated by female and male speaker sources, respectively. The position of the largest peak in each summary clearly indicates the pitch period of each source.

4. Coherent fragment generation

4.1. Generating harmonic fragments

After the spectral integration in the correlogram domain, spectral groups that are likely to have been produced by the same source need to be linked together across time to form coherent spectro-temporal fragments. In each frame we refer to the source that dominates more frequency channels as the 'stronger' source. If the stronger source were constant from frame to frame, the problem of temporal integration would be solved by simply combining the spectral groups associated with the greater number of channels in each frame. However, due to the dynamic aspects of speech, the dominating source will change as the relative energy of the two sources changes over time. Although a speaker's pitch varies over a considerable range, and pitches from simultaneous speakers may overlap in time, within a short period (e.g. 100 ms) the pitch track produced by each speaker tends to be smooth and continuous. We therefore use this cue to generate harmonic fragments.

4.1.1. Multiple pitch tracking

The original ACG channels grouped in the spectral integration stage are summed and the largest peak in each summary is selected as its local pitch estimate. As shown in Fig. 6e and f, it is easier to locate the largest peak after spectral integration. The peak that corresponds to the pitch period of each source is very clear in each summary, while locating them in the summary of all ACG channels (panel d) is a more challenging problem. For the stronger source the largest peak is selected as its pitch estimate. For the weaker source (if one exists) up to three peaks are selected as its pitch candidates. Although this is rare, there are situations where the position of the largest peak in the summary of the weaker source does not correspond to its pitch period, due to lack of harmonic energy or errors made in the spectral integration stage. In this case the second and third largest peaks may be just slightly lower than the largest peak and it is very likely that the position of one of them represents the pitch period. Keeping three pitch estimates for the weaker source has proved beneficial to reducing this type of error. The pitch estimates are then passed to a multi-pitch tracker to form smooth pitch track segments. The problem is to find a frame-to-frame match for each pitch estimate. Here we compare two different methods.

I. Model-based multi-pitch tracker

Coy and Barker (2006) proposed a model-based pitch tracker which models the pitch of each source as a hidden Markov model (HMM) with one voiced state and one unvoiced state. When in the voiced state the models output observations that are dependent on the pitch of the previous observation. Gender dependent models of pitch dynamics are trained from clean speech by analysing the pitch of the utterances in the Aurora 2 training set (Hirsch and Pearce, 2000). In order to track two sources in a pitch space which contains several candidates, two models are run in parallel along with a noise model to account for the observations not generated by the pitch models. The Viterbi algorithm is employed to return the pitch track segments that both models are most likely to generate concurrently. In this study, the model-based tracker is employed in a manner that does not make assumptions about the genders of the speech sources that were made in (Ma et al., 2006; Barker et al., 2006). In those papers the two simultaneous speakers were always assumed to be different genders and therefore two HMMs for different genders were used. This manner of application is inappropriate as the genders of concurrent speakers are not known. Therefore in this study three different model combinations (male/male, female/female and male/female) are compared and the hypothesis with the highest overall score (obtained using the Viterbi algorithm) is selected.

II. Rule-based multi-pitch tracker

McAulay and Quatieri (1986) proposed a simple 'birth–death' process to track rapid movements in spectral peaks. This method can be adapted to link pitch estimates over time to produce smooth pitch track segments. A match is attempted for a pitch estimate p_t in frame t. If a pitch estimate p_{t+1} in frame $t+1$ is the closest match to p_t within a 'matching-interval' Δ and has no better match to the remaining unmatched pitch estimates in frame t, then it is adjoined to the pitch track associated with p_t. A new pitch track is 'born' if no pitch track is associated with p_t and both p_t and p_{t+1} are added into the new track. Analysis of $F0$ trajectories using clean speech signals show that in 90% of the voiced frames the frame-to-frame (10 ms frame shift) pitch changes do not exceed 5% of the pitch of the preceding frame. Therefore, the matching-interval Δ used here is 5% of the pitch estimate the track is trying to match. This rule-based process is repeated until the last frame.

An example of the output of the rule-based multi-pitch tracker is shown in Fig. 10. Panel c shows the pitch estimates for a female(target)/male(masker) speech mixture. Dots represent pitch estimates of the stronger source in each frame and crosses represent those of the weaker source. The smooth pitch track segments are displayed as

circles in panel d, with ground-truth pitch tracks[5] of the pre-mix clean signals displayed as solid lines in the background. The concurrent pitch track segments produced show a close match to the ground-truth pitch estimates. The model-based tracker gives very similar output.

4.1.2. Temporal integration

Spectral groups produced in the spectral integration stage are combined across time if their pitch estimates are linked together in the same pitch track segment, producing spectro-temporal fragments. Each fragment corresponds to one pitch track segment. This process is illustrated in Fig. 8. The upper-left panel shows integrated spectral groups for five frames. Regions with different shades of grey represent different spectral groups in each frame. Pitch estimates for each group in each frame are shown in the lower-left panel. The lower-right panel shows two smooth pitch track segments that are formed. The two corresponding spectro-temporal fragments are shown in the upper-right panel.

Fig. 10, panel e shows the fragments produced corresponding to the pitch tracks (Fig. 10, panel d) in the example of the female(target)/male(masker) speech mixture. Each fragment is represented using a different shade of grey. It demonstrates a close match between the generated fragments and the 'oracle' segmentation (panel b).

This temporal integration step also has the potential to deal with ambiguous pitch tracks caused by a similar pitch range from different sound sources. Consider the situation where two pitch tracks intersect, as illustrated at the top panel in Fig. 9. The ambiguous pitch tracks will be represented as four pitch track segments by the system and hence four corresponding spectro-temporal fragments can be formed (the middle and bottom panel in Fig. 9). This allows the decision on combining fragments (e.g. $\{AD, BC\}$ or $\{AC, BD\}$) to be deferred to the recognition stage.

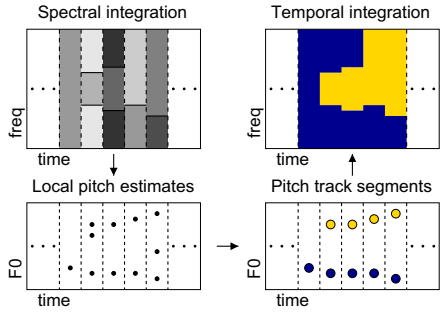

Fig. 8. In anticlockwise sequence: Upper-left panel: regions with different shades of grey represent different spectral groups in each frame. Lower-left panel: dots are local pitch estimates for the spectral groups. Lower-right panel: two pitch track segments are produced by linking the local pitch estimates. Upper-right panel: two spectro-temporal fragments are formed corresponding to the two pitch track segments.

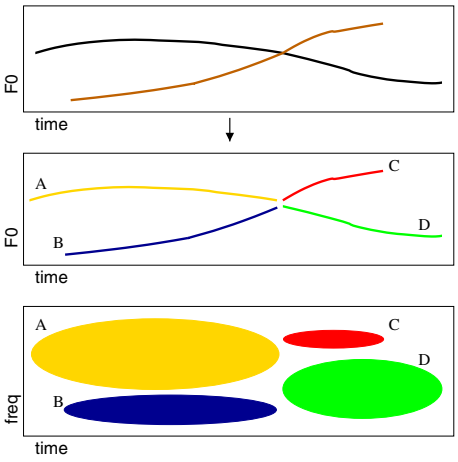

Fig. 9. Top panel: two intersecting pitch tracks. Middle panel: the ambiguous pitch tracks can be represented as four pitch track segments. Bottom panel: four corresponding spectro-temporal fragments can be formed allowing a later decision on fragment combination (e.g. $\{AD, BC\}$ or $\{AC, BD\}$) during the recognition process.

4.2. Adding inharmonic fragments

Unvoiced speech lacks periodicity and thus does not produce dendritic structures in the correlogram domain. The proposed technique which exploits the periodicity cue skips unvoiced regions and as a result spectro-temporal pixels corresponding to these regions are missing (e.g. the white region at about 1.1 s in Fig. 10e). The unvoiced regions of the speech signal are important in distinguishing words which differ only with respect to their unvoiced consonants (e.g. /pi:/ and /ti:/). Therefore it is necessary to include some mechanism that can form coherent fragments for these unvoiced regions.

Hu (2006) gives a systematic study of unvoiced speech segregation. In the current work, as the focus is on separation of *periodic* sounds, we employ a simple inharmonic fragment generation technique reported in (Coy and Barker, 2005). Harmonic regions are first identified in the 'ratemap' representation of the mixture using the techniques described in Section 4.1. The 'ratemap' of the remaining inharmonic regions is then treated as an image and processed by the 'watershed algorithm' (Gonzales et al., 2004). The watershed algorithm is a standard region-based image segmentation approach. Imagine the process of falling rain flooding a bounded landscape. The landscape will fill up with water starting at local minima, forming several water domains. As the water level rises, water from different domains meets along boundaries (*watersheds*). As a result the landscape is divided into regions separated by these watersheds. The technique can be applied to segregate inharmonic sources under the assumption that inharmonic sources generally concentrate their energy in local spectro-temporal regions, and that these concentrations of energy form resolvable maxima in the spectro-temporal domain. The inharmonic fragments produced using this

[5] The pitch analysis is based on the autocorrelation method in the 'Praat' program (www.praat.org).

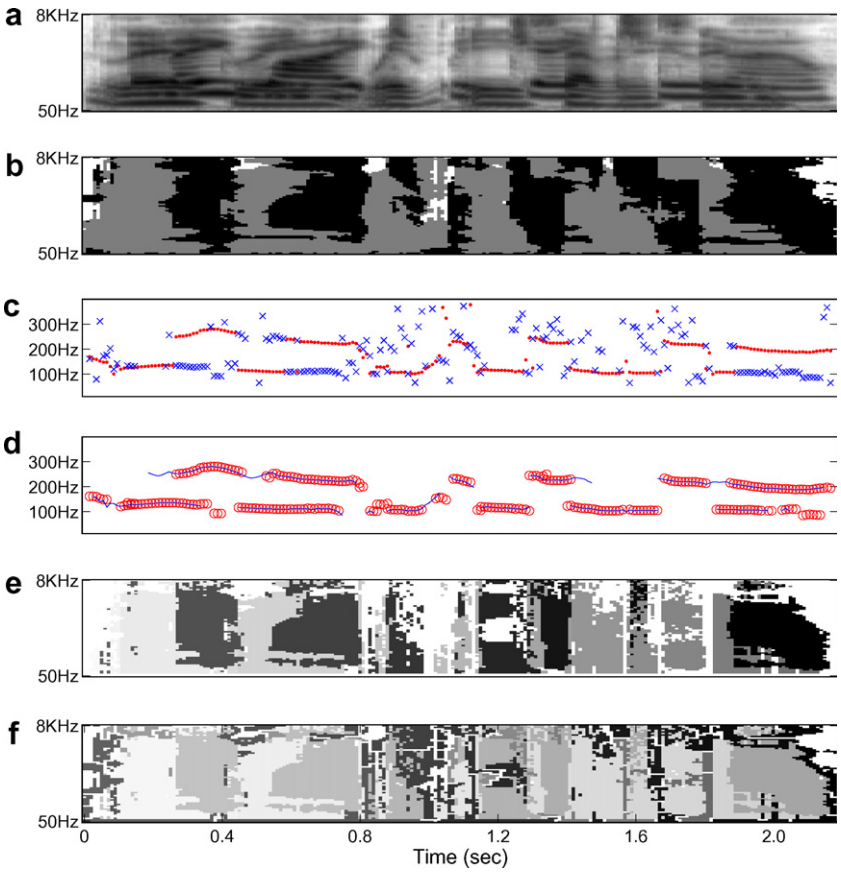

Fig. 10. (a) A 'ratemap' representation of the mixture of 'lay white with j 2 now' (target, female) plus 'lay green with e 7 soon' (masker, male) TMR = 0 dB. (b) The 'oracle' segmentation. Dark grey: pixels where the value in the mixture is close to that in the female speech; light grey: the mixture value is close to that in the male speech; white: low energy regions. (c) Pitch estimates for each source segmentation. Dots represent the pitch of the stronger source in each frame and crosses represent the weaker source at that frame. (d) Circles are pitch tracks produced by the multi-pitch 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.

technique are pooled together with the harmonic fragment as illustrated, for example, in Fig. 10f.

5. Fragment-driven speech recognition

Given the set of source fragments produced by the processes described in Section 4, speech recognition can be performed using the speech fragment decoding (SFD) technique. In brief, the technique works by considering all possible fragment labellings and all possible word sequences. Each fragment may be variously labelled as either being a fragment of the target (foreground) or of the masker (background). A hypothesised set of fragment labels defines a unique target/masker segmentation that can be represented by a 'missing data mask', m_{tf} – a spectro-temporal map of binary values indicating which spectro-temporal elements are considered to be dominated by the target, and which are considered to be masked by the competing sources. Given such a mask, the decoder can use missing data techniques (Cooke et al., 2001) to evaluate the likelihood of each hypothesised word sequence. A Viterbi-like algorithm is then used to find the most likely combination of labelling and word-sequence. A full account of SFD theory is provided in (Barker et al., 2005), and for a detailed description of the application of the technique to simultaneous speech, see Coy and Barker (2007).

One weakness of the SFD technique, in the form described above, is that it produces 'hard' segmentations, i.e. segmentation in which each spectro-temporal element is marked categorically as either foreground or background. If the early processing has incorrectly grouped elements of the foreground and background into a single fragment, then there will be incorrect assignments in the missing data mask that cannot be recovered in later processing. These problems can be mitigated by using missing data techniques that use 'soft masks' containing a value between 0 and 1 to express a degree of belief that the element is either foreground or background (Barker et al., 2000). Such masks can be used in the SFD framework by introducing a spectro-temporal map to express the confidence that the spectro-temporal element belongs to the fragment to which it has been assigned. This confidence map, c_{tf}, uses value in the range 0.5 (low confidence) to 1.0 (high confidence). Given a confidence map, c_{tf}, each hypothesised fragment labelling can be converted into a

soft missing data mask, m_{tf}, by setting m_{tf} to be c_{tf} for time–frequency points that lie within foreground fragments, and to be $1 - c_{tf}$ for time–frequency points within missing fragments. A fuller explanation of the soft SFD technique can be found in (Coy and Barker, 2007).

In harmonic regions, the confidence map is based on a measure of the similarity between a local periodicity computed at each spectro-temporal point, and a global periodicity computed across all the points within each frame in the fragment as a whole. For each spectro-temporal point the difference between its periodicity and the global periodicity of the fragment measured at that time is computed in Hertz, referred to as x. A sigmoid function is then employed to derive a score between 0.5 and 1:

$$f(x) = \frac{1}{1 + \exp(-\alpha(x - \beta))} \tag{10}$$

where α is the sigmoid slope, and β is the sigmoid centre. Appropriate values for these parameters were determined via a series of tuning experiments using a small development data set available in the GRID corpus (see Section 6). It was found that the values of these parameters are not critical to the overall performance and $\alpha = 0.6$ and $\beta = -10$ were used in this study.

Confidence scores for the inharmonic fragments in our study are all set to 1. These confidence scores were used in our coherence evaluation experiment and also employed (as 'soft' masks) along with generated fragments in the SFD system.

6. Experiments and discussion

Experiments were performed in the context of the Interspeech 2006 'Speech Separation Challenge' using simultaneous speech data constructed from the GRID corpus (Cooke et al., 2006). The GRID corpus consists of utterances spoken by 34 native English speakers, including 18 male speakers and 16 female speakers. The utterances are short sentences of the form ⟨*command:4*⟩ ⟨*colour:4*⟩ ⟨*preposition:4*⟩ ⟨*letter:25*⟩ ⟨*number:10*⟩ ⟨*adverb:4*⟩, as indicated in Table 1, e.g. 'place white at L 3 now'. The test set consists of 600 pairs of end-pointed utterances which have been artificially added at a range of target-to-masker ratios. All the mixtures are single-channel signals. In the test set there are 200 pairs in which target and masker are the same speaker; 200 pairs of the same gender (but different speakers); and 200 pairs of different genders. The 'colour' for the target utterance is always 'white', while the 'colour' of the masking utterance is never 'white'.

Table 1
Structures of the sentences in the GRID corpus

Verb	Colour	Preposition	Letter	Digit	Adverb
bin	blue	at	a–z	1–9	again
lay	green	by	(no 'w')	and zero	now
place	red	on			please
set	white	with			soon

Three sets of coherent fragments were evaluated and compared on the same task: 'Fragments-Coy' are fragments generated by the system reported in (Coy and Barker, 2005); 'Fragments-model' and 'Fragments-rule' are coherent fragments generated by the proposed system employing the model-based pitch tracker and the rule-based pitch tracker, respectively.

6.1. Experiment I: Coherence measuring

The fragments are ultimately employed by the speech fragment decoding ASR system and can be evaluated in terms of the recognition performance achieved. However, in addition to ASR performance, a natural criterion for evaluating the quality of fragments is to measure how closely they correspond to the 'oracle' segmentation, obtained with the access to the pre-mix clean signals (see Fig. 10b for an example). To do this we derive the 'coherence' of a fragment as follows. If each pixel in a fragment is associated with a weight, the coherence of the fragment is

$$100 \times \frac{\max(\sum w_1, \sum w_2)}{\sum w_1 + \sum w_2} \tag{11}$$

where w_1 are a set of weights for pixels in the fragment overlapping one source and w_2 are a set of weights for those which overlap the other source. The fragments were compared with the 'oracle' segmentation to identify the pixels overlapping each source. When the decision of each pixel being present or missing in the fragment is discrete (1 or 0), these weights are all simply '1'. In this study we use the confidence scores described in Section 5 as the weights. This choice of weight has the desirable effect that incorrect pixel assignments in regions of low confidence cause less reduction in coherence than incorrect assignments in regions of high confidence. Note that regardless of the confidence score, some spectro-temporal pixels may be more important for speech recognition than others. For instance, pixels with high energy representing vowel regions may be of greater value than low energy pixels. It is less critical that the latter pixels are correctly assigned, and ideally, the coherence score should reflect this. In the current measurement, in the absence of a detailed model of spectro-temporal pixel importance, we make the simple assumption that each pixel has equal importance.

A histogram with a bin width of 10% coherence (hence 5 bins from coherence 50–100%) is computed over the set of fragment coherence values. Both the harmonic and inharmonic fragments are included in the experiment. The fragments are different in size. As smaller fragments are less likely to overlap different sources, their coherence is inherently higher. For example, at one extreme, a single-pixel fragment must always have a coherence of 100%. Although we can get higher coherence scores by generating more small fragments, this would be at the expense of reducing the degree of constraint that the primitive grouping

processes are providing, i.e. a large number of small fragments produces a much greater set of possible foreground/background segmentation hypotheses. Furthermore, the increased hypothesis space leads to an increase in decoding time. This increase can be quite dramatic, especially if fragments are over-segmented across the frequency axis (see Barker et al., 2005). Therefore the aim here is to produce *large and highly coherent fragments*. With these considerations, in the coherence analysis, we reduce the effect of the high coherence contributed by small fragments, by weighting each fragment's coherence value by its size when computing the histogram, i.e., a fragment is counted S times if its size is S pixels. The histograms for the three sets of fragments in all mixture conditions at a TMR of −9 dB are shown and compared in the top three panels of Fig. 11. They have been normalised by dividing the count in each bin by the total number of pixels.

The proposed system with either the model-based pitch tracker or the rule-based pitch tracker produces fragments with very similar quality in terms of coherence. When compared with the fragments generated by Coy and Barker's system, proportionally more fragments with high coherence are produced by the proposed system. This is probably because pitch estimates of each source are computed after the sources are separated. The pitch estimates are thus more reliable and multi-pitch tracking becomes a much less challenging problem. In Coy and Barker's system, however, pitch candidates are formed from the summary of all ACG channels. The multi-pitch tracker possibly finds more incorrect tracks through the noisier pitch data. Furthermore, unlike the proposed system where spectral integration is performed before temporal integration, in Coy and Barker's system spectral integration relies on the less reliable pitch tracks. Therefore it is more likely to produce fragments with low coherence. Within each system, the best results were achieved in the 'different gender' condition, presumably due to the larger difference in the average $F0$s of the sources.

To examine the impact of fragment sizes on the fragment coherence, we also measured the average size of fragments for each coherence histogram bin, shown in the bottom three panels of Fig. 11. Again the two sets of fragments generated by the proposed system give a very similar pattern. In the coherence bins higher than 80% their average fragment size is larger than that of Coy and Barker's system, although in the low coherence bins it is smaller. This is, however, acceptable as there are proportionally less fragments with low coherence in the proposed system.

6.2. Experiment II: Automatic speech recognition

The technique proposed here was also employing within the speech fragment decoding system reported in (Coy and Barker, 2007), and using the experimental set-up developed in (Barker et al., 2006) for the Interspeech 2006 Speech Separation Challenge. The task is to recognise the letter and digit spoken by the target speaker who says 'white'. The recognition accuracy of these two keywords were

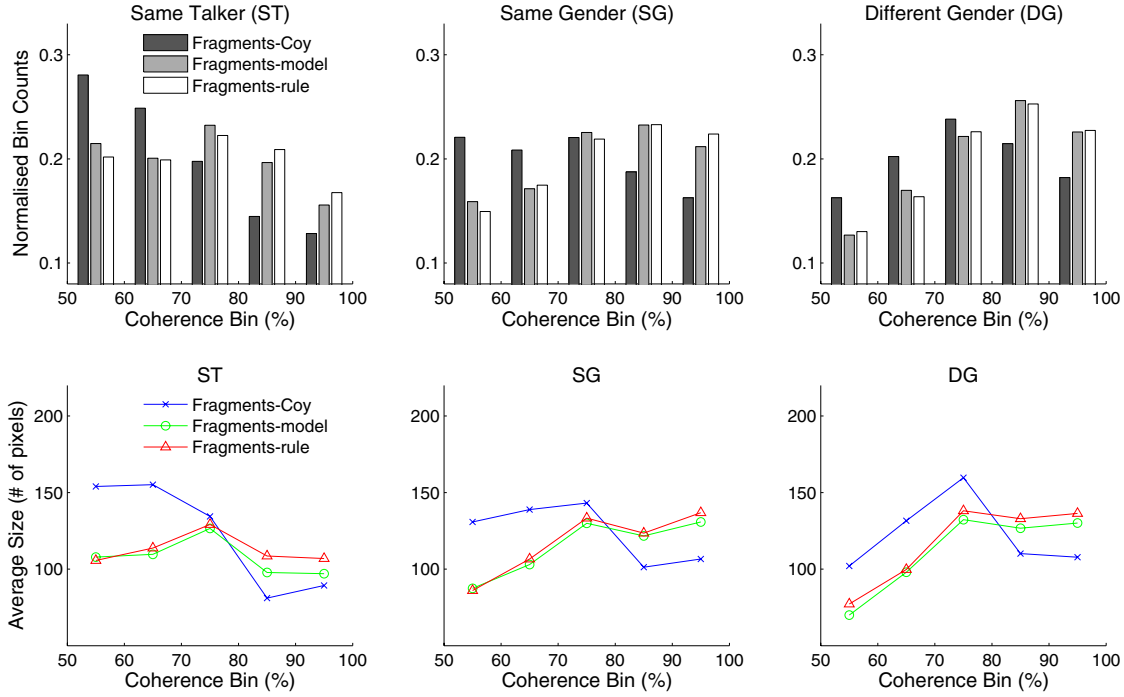

Fig. 11. Coherence measuring results for the three sets of fragments. Top three panels: histograms of fragment coherence after normalisation (TMR = −9 dB). Each fragment's contribution is weighted by its size when computing the histogram. See text for details. Bottom three panels: average size of fragments in each corresponding histogram bin.

averaged for each target utterance. The recogniser employed a grammar representing all allowable target utterances in which the colour spoken is 'white'.

In the SFD system a 64-channel log-compressed 'ratemap' representation was employed (see Section 2). The 128-dimensional feature vector consisted of 64-dimension ratemap features plus their delta features. Each word was modelled using a speaker dependent word-level HMM in a simple left-to-right model topology, with seven diagonal-covariance Gaussian mixture components per state. The number of HMM states for each word was decided based on two states per phoneme. They were trained using 500 utterances from each of the 34 speakers. The SFD system employs the 'soft' speech fragment decoding technique (Coy and Barker, 2007).

The baseline system was a conventional ASR system employing 39-dimensional MFCC features. A single set of speaker independent HMMs with an identical model topology employed 32 mixtures per state. They were trained on standard 13 MFCC features along with their deltas and accelerations.

Following Barker et al. (2006), in all experiments, it is assumed that the target speaker is one of the speakers encountered in the training set, but two different configurations were employed: (i) 'known speaker' – the utterance is decoded using the set of HMMs corresponding to the target speaker, (ii) 'unknown speaker' – the utterance is decoded using HMMs corresponding to each of the 34 speakers and the overall best scoring hypothesis is selected.

We first examine the effect of using soft masks and inharmonic fragments on the recognition performance. The SFD systems with soft masks and inharmonic fragments are then compared to the baseline system and a SFD system using 'Fragments-Coy' with an identical recognition setup.

6.2.1. Effects of soft masks and inharmonic fragments

As discussed in Section 5, an incorrect decision for a spectro-temporal pixel being present in a fragment cannot be recovered when using discrete masks. This also affects the decoding process in automatic speech recognition as the recogniser will try to match speech models with unreliable acoustic evidence. Therefore we compared the recognition performance using the same set of fragments with discrete masks and soft masks. The soft masks described in Section 5 were employed. The discrete masks were produced by simply replacing all the pixels in the soft masks with '1' if their values are greater than 0.5, and with '0' otherwise. The effect of including inharmonic fragments (Section 4.2) on the recognition performance was also examined. Fig. 12 shows recognition results of the SFD system using the set of 'Fragments-rule' in the 'known speaker' configuration. 'All frags + soft masks' represents that both harmonic and inharmonic fragments were used,

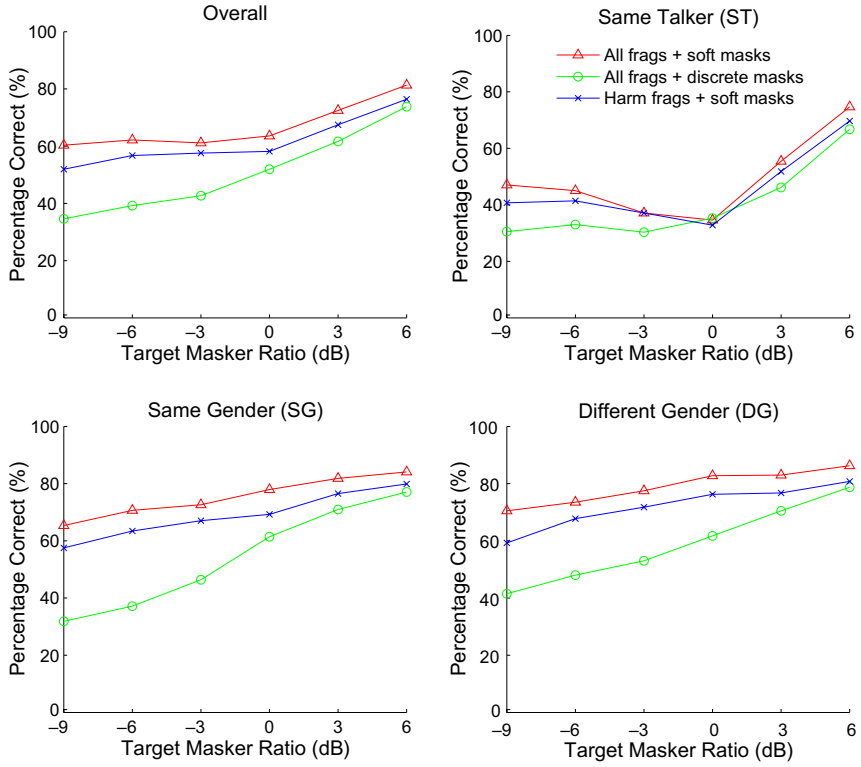

Fig. 12. Recognition accuracy performance of the SFD system using 'Fragments-rule' in 'known speaker' configuration. 'All frags + soft masks': all fragments (both harmonic and inharmonic fragments) with soft masks, 'All frags + discrete masks': all fragments with discrete masks, and 'Harm frags + soft masks': harmonic fragments only with soft masks.

combined with soft masks. 'All frags + discrete masks' represents results using all fragments but with discrete masks. 'Harm frags + soft masks' is the result with harmonic fragments only using soft masks.

Results show that the soft masks had a considerable effect on the recognition performance. With soft masks the system significantly outperformed that with discrete masks across all conditions. As shown in the coherence measuring experiment many fragments have low coherence. Some pixels are unreliable and by assigning a confidence score to each pixel the speech fragment decoder is able to weight the pixel's contribution to the decision. Fig. 12 also shows that in the 'same talker' condition the SFD system using soft masks did not give any recognition accuracy improvement. One possible reason is that in this condition, as shown in Fig. 11, there are more fragments with low coherence and even with soft masks the system could not recover from the errors. Another reason could be that more 'important' pixels were incorrectly assigned in this condition.

Inharmonic fragments also have some impact on the performance in this 'letter + digit' recognition task as many letters are only distinguished by the presence/absence of unvoiced consonants, e.g. letter 'p', 't' and 'e'.

6.2.2. Comparison of different fragment sets

All the recognition results in this section were obtained with the 'soft' SFD system using both harmonic and inharmonic fragments. Fig. 13 shows keyword recognition results of the system using the three sets of coherent fragments discussed before: 'Fragments-Coy', 'Fragments-model' and 'Fragments-rule', in both 'known speaker' and 'unknown speaker' configurations. The 'unknown speaker' results are repeated in Table 2 (model-based pitch tracker) and Table 3 (rule-based pitch tracker). Note the 'known-model' and 'unknown-model' results are essentially the same as those published in (Barker et al., 2006), with minor differences owing to a correction made in the application of the model-based tracker (see Section 4.1.1).

The SFD systems clearly outperform the baseline across all TMRs and across all mixture conditions. They are also able to exploit knowledge of the target speaker identity. The recognition accuracy is significantly higher when the speaker identity is available. Prior knowledge of the speaker identity only fails to confer an advantage in the 'same talker' condition as one would expect. Recognition accuracy results using fragments generated by the proposed system with different pitch trackers are quite similar. This is consistent with the results in the coherence measuring experiment that with different tracks the system produced fragments with similar coherence. The results are significantly better than those produced by Coy and Barker's system, especially at low TMRs. The biggest performance gain was achieved in the 'different gender' condition. This occurs because in this condition the two

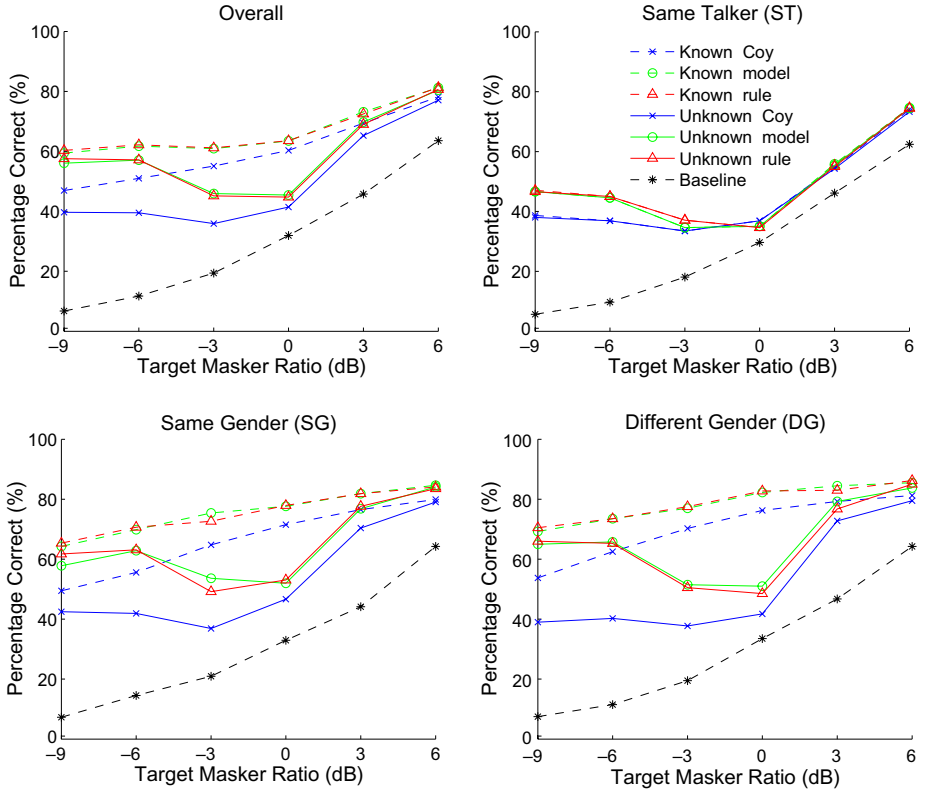

Fig. 13. Keyword recognition results of the proposed system with the model-based/rule-based pitch tracker compared against the system reported in (Coy and Barker, 2005), in both 'known speaker' and 'unknown speaker' configurations. The baseline results are taken from (Barker et al., 2006).

Table 2
Keyword recognition correct percentage (%) for unknown speaker configuration using the model-based pitch tracker

	TMR (dB)					
Condition	−9	−6	−3	0	3	6
Overall	56.08	57.08	45.92	45.42	69.75	80.42
ST	46.61	44.57	34.62	35.07	55.43	74.43
SG	57.82	62.85	53.63	51.96	76.82	84.08
DG	65.00	65.75	51.50	51.00	79.25	83.75

Table 3
Keyword recognition correct percentage (%) for unknown speaker configuration using the rule-based pitch tracker

	TMR (dB)					
Condition	−9	−6	−3	0	3	6
Overall	57.58	57.17	45.17	44.75	69.00	80.67
ST	46.61	45.02	37.10	34.62	54.98	74.43
SG	61.73	63.13	49.16	53.07	77.65	83.52
DG	66.00	65.25	50.50	48.50	76.75	85.00

sources are more likely to have correlated fundamentals, which is difficult to solve purely based on the summary correlogram as discussed in Section 1. The performance improvement in the 'same talker' condition is much less than in the other conditions. This is partially because the target speech and the masker speech are spoken by the same person. With very close *F*0s it is more likely that the pitch-based fragment generation process will group together acoustic evidence from different sources. At low TMRs, same-speaker performance gains may also be reduced because energetic masking is more effective in a same-speaker utterance than in an utterance of different speakers. Many target utterances will be so completely masked at −9 dB that there will be little any system can do to achieve more than chance performance. This effectively reduces the size of the set of utterances on which gains can realistically be made.

As well as examining the ASR performance, it is also instructive to examine the recognition errors. An interesting question is whether the decoder is making errors because it is incorrectly transcribing the target (due to energetic masking), or because it is reporting the masker instead of the target (i.e. a failure to 'attend' to the correct source). To examine this question the recognition output was also scored against the correct transcription for the masker utterance. Table 4 shows the recognition accuracy results at a TMR of 0 dB when scoring against the target speech (as presented in Fig. 13) and when scoring against the masking speech. With the known speaker configuration, the decoder correctly recognised most of the target speech words, without getting confused by the masking speech, in both the 'same gender' and 'different gender' cases. For the artificial 'same talker' condition, however, the reduced performance seems to be explained entirely by the decoder outputting words from the masking utterance. When the simultaneous speech is spoken by the same talker, knowing the identity of the target speaker does not discriminate between fragments of the target and the masker. In fact, at 0 dB there are neither level cues nor speaker identity cues with which to identify the target. For example, when the target speaker says 'a' and the masker (the same speaker) says 'b' concurrently, the two words equally match the known-target speech models and whether 'a' or 'b' is output may be arbitrary.

Table 4
Keyword recognition correct percentage (%) of decoding the target and the masking speech, respectively. TMR = 0 dB

	Known speaker			Unknown speaker		
Condition	Target	Masking	Sum	Target	Masking	Sum
ST	34.62	47.06	81.68	34.62	47.06	81.68
SG	77.93	3.07	81.00	53.07	31.01	84.08
DG	82.75	1.25	84.00	48.50	35.25	83.75

In the unknown speaker configuration the decoder exhibits a performance minimum in the range 0 dB to −3 dB. At these TMR levels the decoder is unable to use level difference cues to distinguish fragments of the target and the masker. As the TMR falls below −3 dB the reintroduction of a level difference between the sources more than compensate for the increased energetic masking and performance initially increases again – at least down to −9 dB.

Although, as discussed earlier, source and target fragments are particularly confusable in the same talker case, the performance dip at 0 dB is also present in the same gender, and even the different gender conditions. It appears that the decoder requires level differences to reliably follow the correct source, and is unable to use speaker differences alone. This is surprising considering the large acoustic differences that exist between the speaker dependent models. Note however that at 0 dB, in the absence of level cues, the only cue for distinguishing target and masker is that the target is the person that says 'white'. So in effect, the system has to solve a speaker identification problem using a single word in the presence of substantial energetic masking. If the word 'white' is not heavily masked it will only fit well to one speaker and decoding paths through that speaker model will be the best overall – hence, the target will most often be correctly identified. However, in utterances where the word 'white' is heavily masked, the fragments masking the word 'white' will be labelled as 'background' and for each speaker there will be a similarly scoring best path. This case is analogous to the word 'white' not being heard, so the cue to the target identity is lost and all speakers become potential targets. In this case, whether the target or the masker is reported may rely on arbitrary factors. In particular, the winning score will depend largely on whether it was the target or the masker who produced an utterance most typical of their average speech patterns, hence leading to the highest likelihood.

7. Conclusions and research directions

7.1. Summary of method

This paper has described a novel approach which exploits the dendritic structure in the correlogram to identify coherent fragments for automatic speech recognition in monaural acoustic mixtures. The use of the full correlogram leads to a more reliable spectral separation and multiple pitch tracking, therefore producing more coherent fragments. The fragments are employed by a speech fragment decoding system which employs missing data techniques and clean speech models to simultaneously search for the set of fragments and the best sequence of words. The recognition accuracy is significantly higher than that of conventional systems.

7.2. Directions for further research

The current system only exploits dendritic structures in a single correlogram (one time frame) for spectral integration. When the decision of assigning a spectral component to a sound source is arbitrary in a frame, it may become more obvious in the next few frames. The correlation between correlograms across time will be examined.

The data set used in this study is artificially mixed simultaneous speech, which therefore lacks some realistic factors (e.g. reverberation, Lombard effect). However, experiments (Cooke et al., submitted for publication) show that this task is challenging even for human listeners. Having the data set artificially made enables us to conduct control experiments. In future we will investigate the robustness of this system to reverberation.

Future work will also aim to develop a statistical model of primitive sequential grouping that will weight segmentation hypotheses according to continuity of primitive properties across fragments through time. e.g. the system fails to make use of the pitch continuity across fragments, which is a useful cue for fragment grouping, specially in the different gender condition.

Acknowledgements

This study was supported by grants from the UK Engineering and Physical Sciences Research Council (GR/R47400/01, GR/T04823/01). We thank Guy Brown for suggesting the idea of using Gabor functions to extract the dendritic structure.

References

Assmann, P., Summerfield, Q., 1990. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. J. Acoust. Soc. Amer. 88 (2), 680–697.

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. ICSLP 2000, Beijing, China, pp. 373–376.

Barker, J., Cooke, M., Ellis, D., 2005. Decoding speech in the presence of other sources. Speech Comm. 45 (1), 5–25.

Barker, J., Coy, A., Ma, N., Cooke, M., 2006. Recent advances in speech fragment decoding techniques. In: Proc. Interspeech 2006, Pittsburgh, pp. 85–88.

Bregman, A., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA.

Brown, G., Cooke, M., 1994. Computational auditory scene analysis. Comput. Speech Lang. 8 (4), 297–336.

Brown, G., Wang, D., 2005. Separation of speech by computational auditory scene analysis. In: Benesty, J., Makino, S., Chen, J. (Eds.), Speech Enhancement: What's New? Springer, New York, pp. 371–402.

Cooke, M., 1991. Modelling auditory processing and organisation. Ph.D. thesis, Department of Computer Science, University of Sheffield.

Cooke, M., Ellis, D., 2001. The auditory organization of speech and other sources in listeners and computational models. Speech Comm. 35 (3–4), 141–177.

Cooke, M., Morris, A., Green, P., 1997. Missing data techniques for robust speech recognition. In: Proc. ICASSP 1997, Vol. 1, Munich, pp. 25–28.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and uncertain acoustic data. Speech Comm. 34 (3), 267–285.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio–visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Amer., 2421–2424.

Cooke, M., Garcia Lecumberri, M., Barker, J. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. J. Acoust. Soc. Amer., submitted for publication.

Coy, A., Barker, J., 2005. Soft harmonic masks for recognising speech in the presence of a competing speaker. In: Proc. Interspeech 2005, Lisbon, pp. 2641–2644.

Coy, A., Barker, J., 2006. A multipitch tracker for monaural speech segmentation. In: Proc. Interspeech 2006, Pittsburgh, pp. 1678–1681.

Coy, A., Barker, J., 2007. An automatic speech recognition system based on the scene analysis account of auditory perception. Speech Comm. 49 (5), 384–401.

de Cheveigné, A., 1993. Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing. J. Acoust. Soc. Amer. 93 (6), 3271–3290.

Ellis, D., 1999. Using knowledge to organize sound: the prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. Speech Comm. 27 (3–4), 281–298.

Glasberg, B., Moore, B., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47, 103–138.

Gonzales, R., Woods, R., Eddins, S., 2004. Digital Image Processing Using MATLAB. Prentice Hall.

Hirsch, H., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ICSLP 2000, Vol. 4, pp. 29–32.

Hu, G., 2006. Monaural speech organization and segregation. Ph.D. thesis, The Ohio State University, Biophysics program.

Licklider, J., 1951. A duplex theory of pitch perception. Experientia 7, 128–134.

Lim, J., Oppenheim, A., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Ma, N., Green, P., Coy, A., 2006. Exploiting dendritic autocorrelogram structure to identify spectro-temporal regions dominated by a single sound source. In: Proc. Interspeech 2006, Pittsburgh, PA, pp. 669–672.

McAulay, R., Quatieri, T., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust. Speech Signal Process. 34 (4), 744–754.

Meddis, R., Hewitt, M., 1991. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. J. Acoust. Soc. Amer. 89 (6), 2866–2882.

Meddis, R., Hewitt, M., 1992. Modeling the identification of concurrent vowels with different fundamental frequencies. J. Acoust. Soc. Amer. 91 (1), 233–245.

Parra, L., Spence, C., 2000. Convolutive blind source separation of non-stationary sources. IEEE Trans. Speech Audio Process., pp. 320–327.

Shamma, S., 1985. Speech processing in the auditory system. I: The representation of speech sounds in the responses of the auditory nerve. J. Acoust. Soc. Amer. 78, 1613–1621.

Slaney, M., Lyon, R., 1990. A perceptual pitch detector. In: Proc. ICASSP 1990. Albequerque, pp. 357–360.

Summerfield, Q., Lea, A., Marshall, D., 1990. Modelling auditory scene analysis: strategies for source segregation using autocorrelograms. In: Proc. Institute of Acoustics, Vol. 12, pp. 507–514.

Wang, D., Brown, G., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks 10 (3), 684–697.