Integrating pitch and localisation cues at a speech fragment level

Heidi Christensen, Ning Ma, Stuart N. Wrigley, Jon Barker

Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
{h.christensen; n.ma; s.wrigley; j.barker}@dcs.shef.ac.uk

Abstract

This paper proposes a novel speech-fragment based approach for processing binaural data to improve the estimation of speech source locations in reverberant, multi-speaker recordings. The technique employs two stages. First, a robust multi-pitch tracking algorithm is used to locate local spectro-temporal 'speech fragments' – regions where the energy in the mixture is dominated by a single speech source. Second, robust localisation estimates are formed by integrating interaural time difference cues over each speech fragment. The technique is applied to the analysis of more than five hours of two-party meetings that have been constructed from a mixture of binaural mannequin recordings. It is shown that estimating location at the speech fragment level produces better results than conventional location-estimate smoothing techniques leading to a an increase in relative frame accuracy rate of more than 35%.
Index Terms: binaural localisation, pitch cues, speech fragment integration

1. Introduction

In typical listening conditions many different sound sources can be simultaneously active. The task of listening is essentially that of developing a description of the individual sound sources from the mixed acoustic signals arriving at the ears. One of the sound source properties that this unmixing process may exploit is *source location*. Evidence that the auditory system exploits source location when processing acoustic mixtures come from studies of 'spatial unmasking' – the dependency of masking level on the relative position of masker and target sound source [1]. Spatial unmasking arises partly due to monaural energetic masking effects but also largely due to binaural mechanisms. The effects can be particularly strong in situations where significant informational masking is present [2], e.g. when the target and masker are both speech signals with similar characteristics [3].

Despite much study, there is little agreement about how binaural information is used in the processing of simultaneous sources beyond the fact that it exploits two cues to sound source location: interaural time difference (ITD), the direction-dependent delay in the time of arrival of the sound at the ear furthest from the source; and interaural level difference (ILD), the direction-dependent level difference between the two ears caused mainly by the manner in which the head shields the ear which is turned away from the sound source.

A naïve strategy for source separation would be to compute ITD and ILD estimates within narrow frequency bands and over small time windows, and then to cluster time-frequency regions on the basis of these ITD/ILD features. However, ITD estimates in narrow frequency bands are often ambiguous, particularly in quasi-periodic regions of the signal. Furthermore, in most listening environments, reverberation renders instantaneous ITD/ILD estimates highly unreliable. Indeed, there is evidence that the ear is unable to exploit ITD alone to group energy across frequency bands [4]. It appears that only by processing cues over extended spectro-temporal regions can ITD cues provide reliable location estimates.

Another possibility is that location cues are exploited at a more central level of auditory processing. Stern *et al.* suggest that energy is first clustered using more robust grouping cues and ITD is then used to segregate grouped components [5] .This is consistent with the recent 'glimpsing' model of speech perception [6] which suggests that speech perception is built upon reliable 'glimpses' of the speech signal that occur in spectro-temporal regions where the SNR is favourable. In such models, location may be acting as a tag that enables the robust sequential grouping of acoustic fragments to form competing auditory streams. Additionally, the location tags may be used by attentional processes that monitor glimpses from a certain direction.

This paper investigates a speech fragment-based model of source localisation and uses reverberant multi-speaker data for evaluation. A robust multi-pitch tracking algorithm locates pitch-track segments that are used to group spectro-temporal speech fragments. ITD-based source location estimated by systems that exploit these fragments are compared to the estimates of systems that process ITD cues in the absence of pitch information. The rest of the paper is structured as follows: Section 2 describes the extraction of auditory cues and the speech fragment generation which is followed by a discussion of the integration of cues in Section 3. Finally Sections 4, 5 and 6 provide experimental details, results and conclusions.

2. Extraction of auditory cues

Both the pitch and localisation cues are based on an auditory front-end simulating the cochlear frequency analysis of the human ear. The model is implemented using a filterbank consisting of 64 overlapping bandpass gammatone filters, with centre frequencies spaced uniformly on the equivalent rectangular bandwith (ERB) scale [7] between 50 Hz and 8000 Hz. The output of the filterbank is used to generate correlograms. The localisation cues, an estimate of the ITD in each frequency band, are obtained by cross-correlating the filtered left and right ear signals and inspecting the *cross*-correlogram covering time lags corresponding to the range -90° to $+90^\circ$ azimuth. The pitch is estimated on the basis of an *auto*-correlogram of the average left and right ear signals corresponding to a pitch period of up to $15\,ms$.

2.1. Binaural Localisation Cues

A running cross-correlation is computed on the output of the gammatone filter. At a given time step t, the cross-correlation

$CC(i,\tau,t)$ for channel i with a time lag τ is given by:

$$CC(i,\tau,t) = L(i,t) * R(i,t-\tau) + K * CC(i,\tau,t-1), \quad (1)$$

where $L(i,t)$ and $R(i,t)$ are the left and right ear output of filterbank i respectively. $K = \exp^{(-t/\lambda)}$ and for the experiments reported here, the exponential time constant, λ was set to 8 ms, which was found to be a good trade off – long enough to produce robust correlations and short enough to approximately satisfy the assumption of stationarity over the correlation window.

2.2. Pitch-based fragment generation

The pitch-based fragments are generated from a signal produced by averaging the left and right ear signals. After averaging, the fragment generation procedure follows that of the system designed for monaural signals presented by Ma *et al.* [8]. In brief, the system first computes the auto-correlogram for the signal, i.e. a running short-time autocorrelation is computed on the output of each gammatone filter, using a 30 ms Hann window. At a given time step t, the scaled autocorrelation $A(i,\tau,t)$ for channel i with a time lag τ is given by

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

where g is the output of the gammatone filterbank and w is a local Hann window of width K time steps.

For periodic sounds, a frame of the auto-correlogram exhibits symmetrical tree-like structures whose stems are located on delays that correspond to multiples of the pitch period. When multiple sources are present the stems for each source fall at different delays, and the position of the stems within each frequency band will indicate which source dominates at that particular frequency. From analysis of the overlapping stem patterns, multiple local pitch estimates are computed. A simple rule-based tracker is then used to form potentially overlapping pitch track segments that extend through time. Each pitch track is then used to recruit a spectro-temporal fragment. In each time frame the frequency channels that have correlogram peaks corresponding to the pitch track value are recruited into that track's fragment. When more than one pitch track is simultaneously active, channels are assigned to tracks according to which pitch track best explains the auto correlogram peaks. Full details of the pitch-fragment generation system can be found in [8].

3. Integration of auditory cues

The standard approach for determining ITDs is to look for peaks in the summed cross-correlogram (e.g. Jeffress' model [9]). These techniques may be adequate in situations where one source dominates the acoustic mixture – e.g. multiparty conversations in quiet rooms where speakers are taking turns to speak. However, in multi-source scenarios a more sophisticated analysis of the cross-correlogram is needed to distinguish peaks corresponding to the different sources from peaks arising from a fusion of the ITDs from multiple sources and reverberation.

Much recent work has focused on the analysis of simultaneous speech signals, exploring the idea that even in multi-speaker scenarios, small time-frequency regions exists which are dominated by just one speaker [6]. This idea was used by Faller and Merimaa [10] who proposed an interaural coherence measure which can be used to identify individual time-frequency points that are dominated by a single speaker, and hence have reliable ITD information. The approach explored in this paper is related to that of Faller and Merimaa, but instead of identifying individual points, it locates extended spectro-temporal regions of single source dominance. Also related is the recent work of Wrigley and Brown who demonstrated the joint use of pitch and localisation cues to improve sound source separation [11].

The experimental framework incorporates two systems, each making use of the speech fragment information in different ways. The first system uses the fragment grouping on a per-frame basis; i.e. information regarding the allocation of the frequency channels to either source is used, and the cross-correlogram is summed separately for channels believed to be dominated by separate sources (*IntFrame*). If no fragments are identified for a frame, the system reverts back to integrating cues across all frequency channels. The second system, (*IntFullFrag*) makes use of the full spectro-temporal extent of the speech fragment. That is, a single location estimate is obtained for a particular fragment by integrating over the part of the cross-correlogram corresponding to the spectro-temporal regions of the fragment. In this scheme a fragment's location estimate does not become available until the fragment has ended, however, at the expense of a short latency, the technique remains compatible with online processing. Like the *IntFrame* system, in the absence of fragment information, this system integrates over all frequency channels.

For comparison two baseline systems have been tested, both of which are based only on localisation cues. The first baseline system implements the standard approach of estimating the ITD from the largest peak in the cross-correlogram summed over all frequency channels (*MaxSum*). The second baseline system is a leaky-integrated version of the first baseline system (*LeakMaxSum*) with a mean lifetime chosen to match the average length of a fragment ~ 30 ms; hence the system is an ITD-only system with a comparable amount of memory to that of the *IntFullFrag* systems[1].

4. Experimental framework

To simulate a natural environment with spatialised multi-speaker scenarios, a set of binaural recordings of digit strings was mixed into longer segments modelling different speaker interaction behaviours.

4.1. Binaural data recordings

A subset of the TIdigits corpus was rerecorded in a standard, office-style room of dimensions $4.09 \times 3.35 \times 2.35$ m using a loudspeaker and a binaural mannequin [12]. No attempt was made to reduce reverberation within the room apart from standard furnishing; the floor was covered with commercial carpet.

Two Brüel & Kjær (B&K) Type 4190 1/2-inch microphones, each connected to a B&K Type 2669 preamplifier, were mounted within a B&K Type 4128C head and torso simulator. These were attached to a B&K Type 2690-0S2 Nexus conditioning amplifier which was, in turn, attached to an M-Audio Firewire Audiophile Mobile Recording Interface under the control of a laptop computer. Original TIdigit utterances were played using a Denon PMA-250SE amplifier coupled to a Mission 760i 2-way reflex loudspeaker. The playback and recording

[1] In fact, the *LeakMaxSum* will have access to more 'memory' than the *IntFullFrag* system which only makes use of temporally integrated information when fragments are available and otherwise reverts back to the *MaxSum* system.

processes were controlled by inhouse software and the captured audio data (sampled at a rate of 48 kHz) was saved directly to hard disk. The laptop computer and external hard disk were positioned outside of the room to limit noise.

The mannequin was placed on a padded office chair in the middle of the room facing one wall. The loudspeaker was positioned along an arc, which maintained a constant distance of 1.5 m from the mannequin, at each of the following azimuths: 0°, $\pm 5^\circ$, $\pm 10^\circ$, $\pm 20^\circ$ and $\pm 40^\circ$. The loudspeaker was positioned at a height of 0.9 m above the floor. Prior to recording at each azimuth, a calibration recording was made for each microphone using a B&K Type 4231 Sound Calibrator. Recordings were made overnight to limit the amount of external noise.

4.2. Mixing multi-speaker segments

The binaural recordings of TIdigits were mixed into 1 minute segments, each with two speakers. Different 'configurations' enabled the study of specific factors on localisation performance such as speaker gender (male/female, female/female and male/male) and inter speaker distance (either spread out ('broad') or closer ('narrow')).

Further, the segments were mixed according to one of three different styles of speaker interaction: 1) **'Turn taking'** style, where each speaker takes it in turn to speak and there is no speaker overlap. 2) **'Simultaneous'** style, where both speakers are active all the time and can be considered to be fully overlapping, and 3) **'Overlapping'** style in which speakers take it in turns but speaker turns partially overlap. Whereas the 'turn taking' and 'simultaneous' styles described above have 0% and 100% overlap respectively, the 'overlapping' style segments have periods with either one or two active speakers in approximately even distribution.

19 segments were mixed for each combination of style, gender composition and speaker position giving a total of 342 ($19 \times 3 \times 3 \times 2$) minutes of data. For each segment, the choice of speaker identities was randomised, as was the chosen digit strings used to make the segments.

4.3. System evaluation and methodology

To evaluate the localisation performance of the system, a frame level metric, denoted Acc is used. It is the number of frames, where the estimated localisation angle, $\hat{\theta}_n$, is close enough to the true angle, θ_n, to be considered correct:

$$Acc = \frac{1}{N} \sum_{n=1}^{N} \delta^*(\theta_n, \hat{\theta}_n) \tag{3}$$

where N is the number of frames. δ^* is defined as

$$\delta^*(a,b) = \begin{cases} = 1 & \text{if } |a-b| < \mathcal{B} \\ = 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathcal{B}$ is a grace boundary around the true angle within which the estimated angle is considered correct. This grace boundary was fixed at 3°.

5. Results and analysis

Figure 1 shows the results of testing the full range of systems described above on the 'turn taking', 'simultaneous' and 'overlap' data in the condition with a male and a female speaker placed at -40 and 40 azimuth respectively. The bars on the far left represent the Acc score for the 'turn taking' data. The most basic baseline system (MaxSum) has a frame accuracy of just over 26%. However, when this system is augmented with the frame-wise fragment cues, the accuracy drops to 24% (IntFrame). Although the IntFrame system is using fragment definitions to provide cross-frequency grouping, this information does not help to improve the ITD estimates. The reason is that there is only ever one active speaker in this condition. Compare this to the 'simultaneous' segment case. Here the more difficult task of assigning location estimates to two simultaneous speakers leads to a reduced baseline frame accuracy for the MaxSum system of just under 17%, but using the frame-level fragment constraints now improves the accuracy significantly to above 28%. In the 'overlapping' data, the IntFrame system produces significant but somewhat smaller improvements.

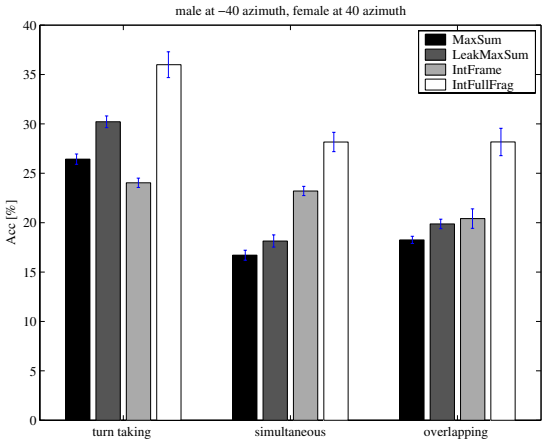

Figure 1: *Acc results from applying the different localisation methods to the 'turn taking', 'simultaneous' and the 'overlapping' segments for the male/female condition at -40/40 azimuth. All results are averaged over 19 segments for each condition, and the error bars indicate standard error.*

The IntFullFrag system, which makes use of the fragment information in both frequency and time, significantly improves the baseline performances for all data styles. For the 'turn taking' data accuracy increased to 36%. When tested on the 'simultaneous' and 'overlapping' data similar large improvements are observed. This increase in accuracy can not be replicated by smoothing across time in the absence of fragment knowledge, as can be seen by examining the results of the LeakIntMaxSum system. For all the data styles, LeakIntMaxSum outperforms the MaxSum system, but it does not achieve as good a performance as the system which integrates across fragments. There was a consistent benefit of the IntFullFrag system over both the MaxSum and LeakIntMaxSum systems when testing on all data styles and on all gender and speaker position conditions.

Results of testing contrasting conditions were analysed in more detail. A clearer picture of system performance can be gained by examining the complete distribution of the localisation estimates for a given condition. Figure 2 shows the histograms of the localisation estimates for two systems on two different conditions: the MaxSum system and the IntFullFrag system on the 'broad' and 'narrow' speaker position conditions; all the histograms are for a male/female mixture.

By comparing the top row of Figure 2 ('broad') to the bottom row ('narrow') it can be seen that when applying both systems to the 'narrow' speaker positions, the localisation estimates become more narrowly distributed around the correct

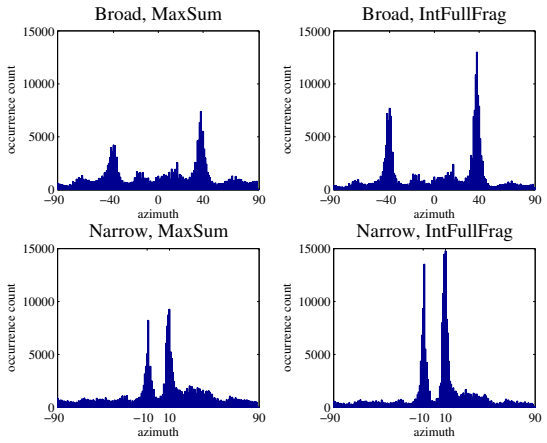

Figure 2: *Histograms for location estimates for broad and narrow locations of speakers. All data from testing on 19 'simultaneous' files for the condition using a male a female speaker are represented.*

positions. The associated frame accuracy scores increase from 22.5% to almost 42%.

The reliability of the pitch fragment generation is highly dependent on the underlying multipitch tracking algorithm. For this, it is important that the pitches of competing sources are resolvable, and that unambiguous track continuations can be found. One might suspect that mixtures with same gender speakers would cause more confusion due to the higher occurrence of overlapping pitch regions. Surprisingly, as can be seen in Figure 3, where the contours of the localisation histograms for all 'simultaneous' segments with different gender compositions are plotted, gender composition does not have a large effect on the distribution of the localisation estimates. The accompanying frame accuracies vary by less than 1%, and the 'turn taking' and 'overlapping' segments display a similar result.

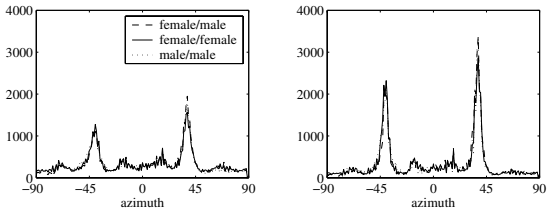

Figure 3: *Histogram contours for location estimates for different gender combinations. The curves are almost identical. All data from testing on 19 'simultaneous' files for the different gender combinations are represented.*

6. Conclusions

A novel speech-fragment based processing of binaural data has been proposed to improve the estimation of speech source locations in reverberant multi-speaker recordings. The technique allows robust location estimates to be produced from noisy cross-correlogram ITD cues by integrating over spectro-temporal regions which are dominated by a single source. We have shown that such speech fragments provide information which, when taken into account when extracting localisation estimates, can improve the frame localisation accuracy for real, reverberant recordings. The systems presented have all operated on low-level cues in a bottom up manner; future work will look at incorporating these techniques into a more sophisticated system with both bottom-up and top-down components.

7. Acknowledgements

This work was funded by the EU Cognitive Systems STReP project POP (Perception On Purpose), FP6-IST-2004-027268. The corpus recording was funded by EU 6th FWP IST IP AMI (FP6-506811).

8. References

[1] A. W. Bronkhurst, "The cocktail party phenomenon: A review of research on speeech intelligibility in multiple speaker conditions." *Acoustica*, vol. 86, pp. 117–128, 2000.

[2] D. Brungart, B. Simpson, M. Ericson, and K. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Amer.*, vol. 100, pp. 2527–2538, 2001.

[3] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, "The role of perceived spatial separation in the unmasking of speech." *J. Acoust. Soc. Amer.*, vol. 106, no. 6, pp. 3578–3588, 1999.

[4] J. F. Culling and Q. S. Summerfield, "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay." *J. Acoust. Soc. Amer.*, vol. 98, no. 2, pp. 785–797, 1995.

[5] R. M. Stern, G. J. Brown, and D. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, algorithms and applications*, Wang and Brown, Eds. Wiley-IEEE press, 2006.

[6] M. P. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, pp. 1562–1573, 2006.

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

[8] N. Ma, P. Green, and A. Coy, "Exploiting dendritic autocorrelogram structure to identify spectro-temporal regions dominated by a single sound source," in *Proceedings of Interspeech 2006*, Pittsburg, USA, 2006.

[9] L. A. Jeffress, "A place theory of sound localization," *Comparative Physiology and Psychology*, vol. 41, pp. 35–39, 1948.

[10] C. Faller and J. Merimaa, "Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence." *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[11] S. N. Wrigley and G. J. Brown, "Recurrent timing neural networks for joint f0-localisation based speech separation," in *Proceedings of ICASSP 2007*, Honolulu, Hawaii, 2007.

[12] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP-84*, vol. 3, 1984, pp. 111–114.