A SPEECH FRAGMENT APPROACH TO LOCALISING MULTIPLE SPEAKERS IN REVERBERANT ENVIRONMENTS

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

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

ABSTRACT

Sound source localisation cues are severely degraded when multiple acoustic sources are active in the presence of reverberation. We present a binaural system for localising simultaneous speakers which exploits the fact that in a speech mixture there exist spectro-temporal regions or 'fragments', where the energy is dominated by just one of the speakers. A fragment-level localisation model is proposed that integrates the localisation cues within a fragment using a weighted mean. The weights are based on local estimates of the degree of reverberation in a given spectro-temporal cell. The paper investigates different weight estimation approaches based variously on, i) an established model of the perceptual precedence effect; ii) a measure of interaural coherence between the left and right ear signals; iii) a data-driven approach trained in matched acoustic conditions. Experiments with reverberant binaural data with two simultaneous speakers show appropriate weighting can improve frame-based localisation performance by up to 24%.

Index Terms— Multi-source, Binaural Localisation, Spectro-Temporal Processing, Reverberation.

1. INTRODUCTION

This paper develops approaches for emulating the human ability to robustly localise speakers in conditions where both reverberation and additive noise are present. In particular we investigate the situation where two potentially overlapping speakers are competing for the listener's attention. This is a challenging and important problem, solutions to which have application in many fields, including automatic speech recognition, design of 'intelligent' hearing aids and remote meeting technologies.

The human auditory system is remarkably adept at localising sound sources in the presence of reverberation. Despite much study, there is little agreement as to how such robustness is achieved. It is evident that the human head gives rise to two main cues to angular localisation: 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. However, in typical listening conditions, where the acoustic mixture at the ear contains information relating to the directions of all simultaneous sound sources and their reflections, it is neither clear how these cues are robustly extracted nor how they are further processed to create a clear percept of individual source locations.

Basic computational solutions to ITD and ILD estimation assume the signal is dominated by a *single* source. Typical *multi*-source solutions work using a two-stage process that initially segregates the signal into parts dominated by a single source, and then runs the ITD and ILD estimation independently on each part. It is common practice to employ a filterbank prior to the segregation stage: the resulting spectro-temporal grid of time-frequency 'cells' is a convenient domain in which to develop solutions to sound source segregation and localisation [e.g., 1; 2; 3]. For example, Mandel et al. [3] use an EM-based clustering approach to group cells with similar ITD estimates.

In our previous work [4], we presented an approach that also operates in the spectro-temporal domain. In contrast to Mandel et al., pitch information was used as the primary cue for source segregation. In acoustic mixtures of non-stationary sound sources (such as speech), pitch can be employed to identify *contiguous* spectro-temporal regions or 'fragments', in which the energy is dominated by just one of the sources. The ITD cues can then be integrated within a fragment to estimate the location of the source from which the fragment arose.

However, in the previous work no explicit attempt was made to handle the effects of reverberation. One approach is to first estimate the *global* degree of reverberation [5; 6] and then employ the estimate to adapt the signal processing algorithms [e.g., 7]. Such approaches can be vulnerable to conditions changing while in operation. A further complication is that in scenarios with multiple sources, the degree of reverberation is source-dependent – e.g., varying with the distance to the listener.

In contrast, this paper presents an approach based on *local* estimates. We note that not all spectro-temporal cells will be equally affected by reverberation, e.g. signal onsets are known to contain more direct energy. Greater robustness can perhaps therefore be achieved by averaging ITD estimates within a fragment using a *weighted* mean. Success of the technique depends on finding approaches to appropriately set the weights, i.e. estimating the quality of the ITD cues within each specific spectro-temporal cell.

The rest of the paper is structured as follows: Section 2 describes the proposed method for integrating ITD cues across a fragment using cell-dependent weights. Section 3 describes the experimental framework, and Section 4 summarises the significant findings from evaluating the systems on real, reverberant binaural recordings, which have been remixed to model various multi-speaker scenarios.

2. PROPOSED METHOD

The common approach for determining ITDs is to locate peaks in the summed cross-correlogram, (e.g. Jeffress' model [8]). A cross-correlogram is a 2D representation resulting from cross-correlating

This work was funded by the EU Cognitive Systems STReP project POP (Perception On Purpose), FP6-IST-2004-027268.

978-1-4244-2354-5/09/$25.00

the left and right ear signals in the different frequency bands across the range of possible time differences or *lags*. Such techniques may be adequate in situations where one source dominates the acoustic mixture – e.g. 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 their reflections. The fragment-level model of localisation integrates ITDs across spectro-temporal regions identified in the spectrogram using pitch tracks [9; 4]. This enables integration across frequency channels *within* a fragment, but avoids mixing together cues from different sources.

2.1. Data processing

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 bandwidth (ERB) scale [10] between 50 Hz and 8000 Hz. The output of the filterbank is used to generate *cross*-correlograms on lags corresponding to the range -90° to $+90^\circ$ azimuth and *auto*-correlograms corresponding to a pitch period of up to 15 ms.

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.* [11]. 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. From analysis of the auto-correlation delay patterns, multiple local pitch estimates are computed, and a simple rule-based tracker is used to form potentially overlapping pitch track segments that extend through time. Each pitch track is then used to recruit a spectro-temporal fragment.

2.2. Fragment integration method

Assume access to a set of speech fragments $\{p_1, \dots p_P\}$ which corresponds to spectro-temporal regions dominated by a single source. Each fragment contains a number of cells, $c_{i,j}$ containing a *lag profile*, $CC_{i,j}(\tau)$, which is the result of cross-correlating from frame i in frequency band j over a range of lags $\tau \in \{\tau_{min}, \dots, \tau_{max}\}$.

$$CC'_{i,j}(\tau) = L_j(i) \cdot R_j(i-\tau) + K \cdot CC_{i-1,j}(\tau), \qquad (1)$$

where $L_j(i)$ and $R_j(i)$ are the left and right ear output of filterbank j respectively. $K = \exp^{(-t/\lambda)}$ and 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. The cross-correlogram is then normalised as in [12]:

$$CC_{i,j}(\tau) = \frac{CC'_{i,j}(\tau)}{\sqrt{(XXL_{i,j}(\tau)\; XXR_{i,j}(\tau)}} \qquad (2)$$

where XXL and XXR are the auto-correlation functions of the left and right ear signals respectively time-shifted according to τ:

$$XXL_{i,j}(\tau) = L_j(i) \cdot L_j(i-\tau) + K \cdot XXL_{i-1,j}(\tau), \qquad (3)$$

$$XXR_{i,j}(\tau) = R_j(i) \cdot R_j(i-\tau) + K \cdot XXR_{i-1,j}(\tau) \qquad (4)$$

The naïve way of combining the lag profiles from all cells in fragment p is to average the lag profiles of each cell in the fragment and then locate the position of the maximum peak,

$$ITD_{\text{naïve}} = \underset{\tau}{\operatorname{argmax}} \left(\frac{1}{\mathcal{J}} \sum_{i,j \in \mathcal{P}_p} CC_{i,j}(\tau) \right) \qquad (5)$$

where $\mathcal{P}_p$ is the set of i, j indices comprising fragment p, and $\mathcal{J}$ is the number of elements in $\mathcal{P}_p$.

A potentially more effective approach is to multiply lag profiles by a cell-dependent 'reliability' weight, $\phi_{i,j}$, before integrating over the fragment. This weight is set according to an estimate of the quality of cell's lag profile, i.e. cell's believed to display unperturbed ITD information will be more heavily weighted. The fragment integration now takes the form,

$$ITD_{\text{weights}} = \underset{\tau}{\operatorname{argmax}} \left(\frac{1}{\mathcal{J}} \sum_{i,j \in \mathcal{P}_p} \phi_{i,j} \cdot CC_{i,j}(\tau) \right) \qquad (6)$$

2.3. Weight definitions

In this paper we evaluate four different ways of defining the weights,

simple fragment: all cells are given equal weight, i.e. $\phi_{i,j} = 1 \; \forall i, j$, as used in [4].

peak wavefront: based on Heckmann et al.'s (2006) model of the "precedence effect", i.e. human's ability to distinguish the target sound direction from its reflections [13; 14]. Two masking phenomena are modelled: 1) forward masking: a leading sound suppresses the impact of a shortly following sound, and 2) backward masking: a sufficiently loud lagging sound will perceptually mask the leading sound. The model identifies peaks of the smoothed filterbank outputs that remain unmasked when these two effects are considered ('wavefront' peaks) – see Heckmann et al. for details. In our system, the weight $\phi_{i,j}$ is set to either 1 or 0 depending on whether a spectro-temporal cell contains an unmasked wavefront peak or not. These are points which are dominated by the direct sound energy and relatively free of the effects of reverberation.

IC weighted: this system follows Faller and Merimaa's (2004) use of interaural coherence (IC),

$$IC_{i,j} = \max_{\tau} CC_{i,j,\tau}. \qquad (7)$$

However, whereas Faller and Merimaa (2004) select localisation data points with IC above some empirically defined threshold, T, i.e.,

$$\phi_{i,j} = \begin{cases} = 1 & \text{if } \; IC_{i,j} > T \\ = 0 & \text{otherwise} \end{cases} \qquad (8)$$

we avoid having to tune a threshold by simply setting $\phi_{i,j} = IC_{i,j}$.

ST position: The final system is motivated by the observation that we want to heavily weight cells that have a high Q, where Q is defined as the value of CC at the τ corresponding to the correct location. We clearly do not know the correct location at run-time, but it is possible that cells in some spectro-temporal positions within a fragment will have consistently higher Q than others. This indeed turns

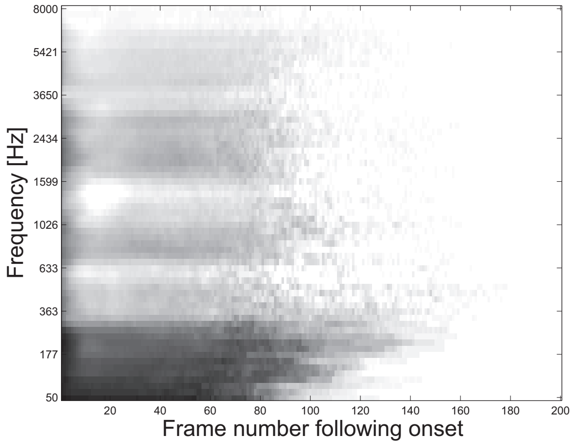

Fig. 1. *The dependence of Q on cell spectro-temporal position within a fragment (darker shading indicates higher average Q values).*

out to be the case, as illustrated in Figure 1, which shows an example of the result of measuring Q on a set of noisy speech training data and displaying the average value as a function of frequency (f) and temporal position relative to the start of the current fragment (t'). As might be expected, we see a degradation in performance as the frame position moves away from the fragment's onset – where the effect of reverberation is greater, – but we also observe a dependence on frequency. In the 'ST position' system we exploit the dependence of Q on f and t' by estimating $Q(f, t')$ using data matching the acoustic conditions on which the system will be tested. The cell weights, $\phi_{i,j}$, are then set to be $Q(f, t')$ where f and t' are the cell's position within its fragment.

3. 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.

3.1. Binaural data

A subset of the TIdigits corpus [15] was rerecorded in a standard, office-style room of dimensions 4.09×3.35×2.35 m using a loudspeaker and a binaural mannequin. No attempt was made to reduce reverberation within the room apart from standard furnishing; the floor was covered with commercial carpet. 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: $0°$, $\pm 5°$, $\pm 10°$, $\pm 20°$ and $\pm 40°$.

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 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 binaural recordings of TIdigits were mixed into one minute segments, each with two talkers where the individual utterances were first normalised to have equal power (where power is measured after first averaging the left and right channels). The results reported in this paper are based on data simulating a conversation between a male and female speaker sitting at $+40$ and -40 azimuths relative to the listener. Further, the segments were mixed according to one of two different styles of speaker interaction:

1. **'Turn taking'** style, where each speaker takes it in turn to speak and there is no speaker overlap, and
2. **'Simultaneous'** style, where both speakers are active all the time and can be considered to be fully overlapping.

For each segment, the choice of speaker identities was randomised, as was the chosen digit strings used to make the segments. For each speaking style a total of 19 segments were mixed.

3.2. Evaluation metric

To evaluate the localisation performance of the system, a frame level metric, denoted $Corr$ 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[1]:

$$Corr = \frac{1}{N} \sum_{n=1}^{N} \delta^*(\theta_n, \hat{\theta}_n) \times 100\% \tag{9}$$

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

$$\delta^*(a, b) = \begin{cases} = 1 & \text{if} \quad |a - b| < \mathcal{B} \\ = 0 & \text{otherwise} \end{cases} \tag{10}$$

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 $3°$.

4. RESULTS AND CONCLUSIONS

The weight definitions proposed in section 2 were used to develop four systems which have been tested on the 'turn taking' and 'simultaneous' speaking style data described above. These systems are compared to two baseline, non-fragment systems. The "frequency int." system integrates ITD cues across the full frequency range within a single time frame, and the "leaky int." system integrates across frequencies and across time by applying a leaky integrator with a mean lifetime chosen to match the average length of a fragment ~ 30 ms.

Figure 2 illustrates the frame correctness scores for the six systems for the two data styles. Comparing the first three systems shows the improvement obtained by integrating ITD cues at the fragment level. For the 'simultaneous' data segregating the spectro-temporal cells into fragments is clearly advantageous (a relative improvement of nearly 95%). For the 'turn taking' sessions using fragments produce a smaller but clearly significant increase in correctness of 31%. This boost in performance can be partly attributed to the fact that at speaker changes, the fragments ensure that cues from two adjacent speakers are not merged together as is the case when a simple leaky integration is performed.

[1] In the simultaneous speaker condition $\hat{\theta}_n$ is scored as correct if it is close to the true angle of *either* speaker.

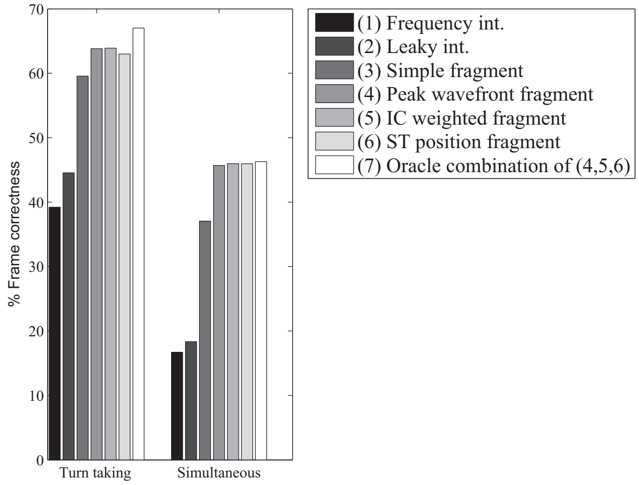

Fig. 2. *Frame correctness for the two baseline systems and the four fragment based systems.*

The effect of the weighted mean integration approach is seen when comparing the 'simple fragment' to systems (4), (5) and (6); a relative improvement of up to 7.3% and 24% for the 'turn taking' and 'simultaneous' data respectively. All weighted systems significantly outperform the simple fragment system; when doing a pair-wise comparison across the systems on the 19 different segments, each of the weighted mean systems perform better than the 'simple fragment' system. However, the picture is less clear when comparing the weighted approaches to each other. Based on the pairwise comparisons on the 'turn taking' data, the 'Peak wavefront' and 'IC weighted' systems perform consistently better than 'ST position' system. For the 'simultaneous' data all weighted systems appear to perform equally well.

An interesting question is whether the three different weight definitions, although clearly related, exhibit complementary information and hence could be advantageously combined. To address this question, we established a final system (7) 'oracle combination' which combines the three weight-based systems in an *oracle* manner, i.e. for each frame, the output of the best performing system is picked. If the errors are uncorrelated the oracle system will be able to perform better than any of the individual systems.

For the 'turn taking' data the oracle system consistently outperforms the other weight-based systems – indicating that these systems are making different errors – but for the 'simultaneous' data the 'oracle' system only shows a very small overall increase in performance.

5. CONCLUSIONS

A fragment-level localisation model has been proposed that integrates the localisation cues within a fragment using a weighted mean. The weights are based on local estimates of the degree of reverberation in a given spectro-temporal cell. Several weight estimation approaches have been investigated emulating the human ability to robustly localise speakers in natural, reverberant environments. Experiments with reverberant binaural data with two simultaneous speakers show appropriate weighting can improve frame-based localisation performance by up to 24%.

The results presented here are a first attempt at obtaining a spectro-temporal 'purity' measure. We are currently investigating ways of applying machine learning techniques to train statistical models for online prediction of individual spectro-temporal cell contamination.

References

[1] O. Yilmaz and S. Rickard, "Blind sepraration of speech mixtures via time-frequency masking," *IEEE Trans. on Sig. Proc.*, vol. 52, pp. 1830–1847, 2004.

[2] N. Roman, D. Wang, and G. J. Brown, "A classification-based cocktail party processor," in *Proc. of Neural Information Processing Systems*, 2003.

[3] M. Mandel, D. Ellis, , and T. Jebara, "An em algorithm for localizing multiple sound sources in reverberant environments," in *Advances Neural Info. Proc. Sys.*, Vancouver, Canada, Dec 2006, vol. 19, pp. 953–960.

[4] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "Integrating pitch and localisation cues at a speech fragment level," in *Proc. of Interspeech'07*, Antwerp, Belgium, August 2007.

[5] S. Vesa and A. Härmä, "Automatic estimation of reverberation time from binaural signals," in *Proc. of ICASSP'05*, 2005.

[6] R. Ratnam, D. L. Jones, and W. D. O'Brien, "Fast algorithms for blind estimation of reverberation time," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 537–540, June 2004.

[7] K. Wilson and T. Darrell, "Improving audio source localization by learning the precedence effect," in *Proc. of ICASSP'05*, 2005.

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

[9] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources.," *Speech Communication*, vol. 49, pp. 874–891, 2007.

[10] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 44, pp. 99–122, 1990.

[11] 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.

[12] 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.

[13] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am*, vol. 106, no. 4, pp. 1633–1654, Oct 1999.

[14] M. Heckmann, T. Rodemann, F. Joublin, C. Goerick, and B. Schölling, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," in *Proc. of Int. Conf. Intelligent Robots and Systems.*, 2006.

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