INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech localisation in a multitalker mixture by humans and machines

Ning Ma and Guy J. Brown

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{n.ma, g.j.brown}@sheffield.ac.uk

Abstract

Speech localisation in multitalker mixtures is affected by the listener's expectations about the spatial arrangement of the sound sources. This effect was investigated via experiments with human listeners and a machine system, in which the task was to localise a female-voice target among four spatially distributed male-voice maskers. Two configurations were used: either the masker locations were fixed or the locations varied from trial-to-trial. The machine system uses deep neural networks (DNNs) to learn the relationship between binaural cues and source azimuth, and exploits top-down knowledge about the spectral characteristics of the target source. Performance was examined in both anechoic and reverberant conditions. Our experiments show that the machine system outperformed listeners in some conditions. Both the machine and listeners were able to make use of *a priori* knowledge about the spatial configuration of the sources, but the effect for headphone listening was smaller than that previously reported for listening in a real room.

Index Terms: Speech localisation, multitalker, human-machine comparison, deep neural networks

1. Introduction

In the 1950s, Cherry [1] noted the ability of listeners to attend to one speaker in the presence of others, and called this the 'cocktail party problem'. Since then, this aspect of human hearing has been the subject of much psychophysical investigation [2], and has also motivated computational work which aims to build voice separation systems. However, developing a system which matches human performance in the cocktail party problem has proven to be very challenging.

Both bottom-up and top-down systems are at play in the perceptual organisation of sound, via a process termed 'auditory scene analysis' (ASA) by Bregman [2]. In the cocktail party scenario, the voice of the target speaker and interfering (masker) voices will originate from different locations. Hence, binaural cues – interaural time difference (ITD) and interaural level difference (ILD) – will differ for the target and maskers, providing a means to identify them. In addition to this bottom-up cue, top-down knowledge can also be applied. In the cocktail party, this includes information about the vocal characteristics of the target and masker voices, and also knowledge about their spatial positions. In the latter regard, listeners could potentially exploit the fact that the spatial locations of the masker voices are known.

Indeed, a recent psychophysical study has shown that listeners are able to exploit prior knowledge of the masker locations in a cocktail party scenario. Kopco et al. [3] investigated the ability of listeners to localise a female target voice in the presence of four male masking voices. They found that listeners were better able to localise the target when the spatial locations of the masker voices were cued before the task. Kopco et al.'s experiment was conducted in a natural listening environment, in which voices were played from a loudspeaker array and listeners were free to move their heads during the task.

This paper addresses two main research questions. First, we ask whether listeners are able to exploit prior information about the masker locations in Kopco et al.'s task when listening over headphones, where binaural cues are limited to those present in the head related impulse responses (HRIRs) used to spatialise the signals for headphone listening. In headphone listening, head movements are not available and room characteristics can be carefully controlled; hence, we also investigate whether prior knowledge of the masker locations can assist localisation in both anechoic and reverberant conditions. Second, we ask whether the sources of knowledge available to listeners in this scenario – speaker characteristics and masker locations – can be successfully exploited in a computational system for sound localisation. Such information is not typically used in machine listening systems for source localisation [4].

The remainder of the paper is structured as follows. First, we describe a listening test which broadly follows the method of Kopco et al., but uses headphone listening and assesses listener performance under both reverberant and anechoic conditions. A computational model is then described, which exploits speaker models and prior information of masker locations within a deep neural network (DNN) architecture. Finally, a comparison of listener and model performance on the same localisation task is presented.

2. Methods

2.1. Participants

Eleven normal-hearing listeners participated in the listening test, including four females and seven males between the ages of 22 and 50 years.

2.2. Stimuli and setup

Speech materials were taken from a corpus of monosyllabic words recorded at Boston University's Hearing Research Centre [5], as used in [3]. The target was a female voice speaking the word 'two'. The four maskers were all male voices speaking non-digit words, drawn randomly from a set of 32 words. All speech material was recorded at a sample rate of 44.1 kHz with an average duration of 0.4 s.

Participants listened to the stimuli via headphones in a simulation of binaural localisation. Two listening sessions were included. For the anechoic session, binaural speech signals were created by convolving monaural signals with HRIRs recorded

This work was supported by the European Union FP7 project TWO!EARS (http://www.twoears.eu) under grant agreement No. 618075.

http://dx.doi.org/10.21437/Interspeech.2016-1149

from the Knowles Electronic Manikin for Acoustic Research (KEMAR) dummy head [6]. For the reverberant session, the binaural room impulse response (BRIR) of Room A from the Surrey BRIR database [7] was used to simulate reverberant room conditions. The Surrey database was captured using a Cortex head and torso simulator (HATS). Room A has a reverberation time (T_{60}) of 0.32 s and a direct-to-reverberant ratio (DRR) of 6.1 dB. The room has dimensions 5.7 × 6.6 × 2.3 m (width × length × height), and the BRIR was measured at a head height of 1.78 m and a distance of 1.5 m between the circular loudspeaker array and the HATS.

Binaural mixtures of five competing talkers (one female target, four male maskers) were created by spatialising each talker separately before adding them together in each of the two binaural channels. For both anechoic and reverberant sessions, each masker was equal in level to the target.

2.3. Procedure

Listeners participated in the experiment in a sound-attenuating booth using a computer running MATLAB. Stimuli were presented over a pair of Sennheiser HD 600 headphones. A graphical user interface (GUI) was used to record participants' responses. Their task was to report the location of the female-voice target either in isolation (control runs), or in the presence of four male-voice maskers (masker runs). Participants indicated their response by selecting a loudspeaker location in a loudspeaker array shown on the computer screen using a computer mouse. There was also a button in the GUI that listeners could press to indicate that no target was heard.

The listening tests were administered across two sessions that were completed on different days. In one session anechoic stimuli were used while in the other session reverberant stimuli were used. At the beginning of each session, a practice run was included in which the participants listened to the female-voice target in isolation from all target locations. After that, 12 runs were presented following a similar procedure adopted in [3]. The first and last of these were no-masker control runs, each of which consisted of 55 trials (5 trials per target location). In the masker runs the maskers were presented in one of five masker patterns (see Figures 3 and 4). There were five runs where the masker pattern was kept fixed for the duration of the run (Fixed), and five runs where the masker pattern was randomly chosen on each trial (Mixed). Each masker run consisted of 60 trials including five catch trials, in which the target was replaced by another random male masker. The catch trials were included in order to monitor false alarm rates [3]. The type of masker runs was indicated at the beginning of each run by presenting a recording of the phrase 'fixed maskers' sequentially at each of the four masker locations for the Fixed runs, and a recording of the phrase 'mixed maskers' for the Mixed runs. The Fixed and Mixed runs were interleaved.

3. Model

We now present a computational model of binaural speech localisation for multitalker scenarios. This system uses DNNs to learn the relationship between binaural cues and source azimuth. It can also exploit top-down knowledge about the spectral characteristics of the target source, and the prior knowledge of masker positions when available.

3.1. Binaural feature extraction

The auditory front-end consisted of a bank of 32 overlapping Gammatone filters, with centre frequencies uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 8 kHz [8]. Inner-hair-cell processing was approximated by half-wave rectification. Following this, the cross-correlation between the right and left ears was computed independently for each frequency channel using overlapping frames of 20 ms duration with a shift of 10 ms.

As in [9], the system used the whole cross-correlation function, instead of ITD, as localisation cues. This approach was motivated by observations that computation of ITD may not be robust in the presence of multiple talkers, and that there are systematic changes in the cross-correlation function with source azimuth. When sampled at 16 kHz, the cross-correlation function with a lag range of ± 1 ms produced a 33-dimensional binaural feature vector for each frequency channel. This was supplemented by the ILD, forming a final 34-d feature vector.

3.2. DNN-based localisation

The relationship between binaural cues and source azimuth in each frequency channel was learned by a DNN. The DNN consists of an input layer, 4 hidden layers, and an output layer. The input layer contained 34 nodes and each node was assumed to be a Gaussian random variable with zero mean and unit variance. The hidden layers had sigmoid activation functions, and each layer contained 128 hidden nodes. The output layer contained 51 nodes corresponding to 51 azimuth angles between -50° and 50° with an azimuth resolution of 2°. The 'softmax' activation function was applied at the output layer.

Given the observed localisation feature set $\boldsymbol{s}_{tf}$ at time frame t and frequency channel f, the 51 'softmax' output values from the DNN for frequency channel f were considered as posterior probabilities $\mathcal{P}(\phi|\boldsymbol{s}_{tf})$, where ϕ is the azimuth angle and $\sum_\phi \mathcal{P}(\phi|\boldsymbol{s}_{tf}) = 1$. The posteriors were then integrated across frequency to yield the probability of azimuth ϕ, given features of the entire frequency range at time t

$$\mathcal{P}(\phi|\boldsymbol{s}_t) = \frac{\prod_f \mathcal{P}(\phi|\boldsymbol{s}_{tf})}{\sum_\phi \prod_f \mathcal{P}(\phi|\boldsymbol{s}_{tf})}. \tag{1}$$

$\mathcal{P}(\phi|\boldsymbol{s}_t)$ was then integrated over the duration of the entire stimulus. The target location was given by the azimuth that maximises the probability.

The DNNs were trained using speech signals from the GRID corpus [10], spatialised using an anechoic HRIR measured with a KEMAR dummy head [6]. Diffuse noise were added during training and there was no retraining using the matching BRIR for this study.

3.3. Exploiting top-down source knowledge

The DNN localisation system indicates the probability of a sound source occurring at each possible azimuth. In the proposed system, a set of parameters ω_{tf} were employed to selectively weight the contribution of binaural cues from each time-frequency bin in order to localise the attended target source in the presence of competing sources [4]:

$$\mathcal{P}(\phi|\boldsymbol{s}_t) = \frac{\prod_f \mathcal{P}(\phi|\boldsymbol{s}_{tf})^{\omega_{tf}}}{\sum_\phi \prod_f \mathcal{P}(\phi|\boldsymbol{s}_{tf})^{\omega_{tf}}}. \tag{2}$$

Here ω_{tf} is introduced as a factor between $[0, 1]$. This allows cues that derive from a frequency channel dominated by the target talker to be emphasised; or conversely, cues that derive from a masking taker to be penalised.

Under the *log-max* approximation [11] of the interaction function between two acoustic sources in log-spectra, i.e. $\boldsymbol{y}_{tf} \approx \max(\boldsymbol{x}_{tf}, \boldsymbol{n}_{tf})$, the localisation weight ω_{tf} can be defined as the probability of $\boldsymbol{y}_{tf}$ being dominated by $\boldsymbol{x}_{tf}$

$$\omega_{tf} = P(\boldsymbol{x}_{tf} = \boldsymbol{y}_{tf}, \boldsymbol{n}_{tf} \leq \boldsymbol{y}_{tf} | \boldsymbol{y}_t, \lambda_x, \lambda_n), \quad (3)$$

where λ_x and λ_n are the models for the target and masking talkers, respectively. Here, the talker models were represented as Gaussian mixture models (GMMs) with diagonal covariances.

Source spectral characteristics were modelled using ratemap features [12]. A ratemap is a spectro-temporal representation of auditory nerve firing rates, extracted from the inner hair cell output of each frequency channel by leaky integration and downsampling. Ratemaps were computed for each ear, averaged across the two ears, and finally log-compressed (cf. the log-max approximation noted above). The stimuli from the practice run were used to estimate source model parameters for the target talker, and the signals of the catch trials were used to estimate the masker model, i.e. the system used the same information that was available to the listeners.

3.4. Exploiting prior knowledge about masker locations

In the Fixed runs prior knowledge about masker locations was available. Such information can be exploited in the system by reducing the probabilities $\mathcal{P}(\phi|\boldsymbol{s}_{tf})$ in Eq. 2 where a masker is located at azimuth ϕ. However, since a target and a masker can be co-located, by doing so the probability at the true target location is also reduced. The proposed system therefore only reduced the probabilities $\mathcal{P}(\phi|\boldsymbol{s}_{tf})$ at a masker location ϕ in the time-frequency region considered to be dominated by the maskers, i.e. where $\omega_{tf} < 0.5$ (Eq. 3).

4. Results and discussion

4.1. Results of listening tests

Figure 1 plots RMS errors averaged across participants in the no-masker control condition as a function of the target location. For the anechoic case, RMS errors grew approximately with target laterality from about 11° to 22°. The V-shape trend is consistent with that reported in [3] where localisation was performed in a real room, but the localisation accuracy is substantially lower for headphone listening and the V shape is also less pronounced.

For the reverberant case, however, control data did not show a dip at the central locations, with RMS errors ranging between 12° to 17°. This is most likely due to the effect of reverberation on perceived location. Many participants reported that the target speech appeared to emirate from above them, presumably due to reflections from the ceiling in which the BRIRs were recorded. Another possible explanation is that the mismatch between the listeners' HRIRs and the one used to simulate binaural listening in this study could disrupt speech localisation more when reverberation is present, as shown in [13].

Figure 3 shows the effect of maskers on RMS errors for each target location and each masker pattern, by subtracting control RMS errors for each participant from those in the different masker conditions. The effect of masking on RMS errors depended in a complex way on various parameters manipulated in this study. First, the presence of maskers resulted in a larger increase in error in the more challenging reverberant condition, in particular at more lateral target locations. In the anechoic condition, however, the RMS errors tended to increase most at more central locations between -20° and 20°. This is in contrast to findings by Kopco et al. [3], where the RMS errors tended to increase most at target locations that corresponded to masker locations.

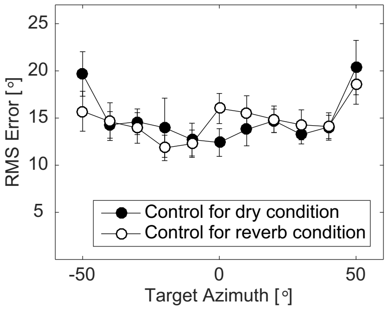

Figure 1: Localisation performance in the no-masker control condition. Across-participant averages (±1 SEM) of the RMS error are plotted as a function of the target location for both anechoic (dry) and reverberant (reverb) sessions.

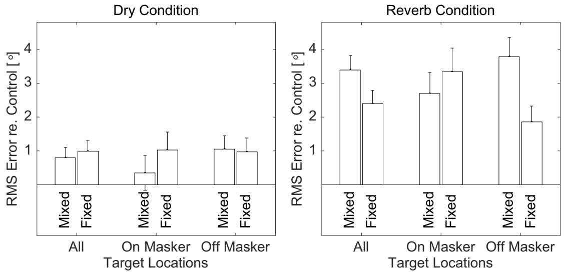

Figure 2: Effect of maskers on localisation accuracy shown as the increase in RMS error (*re.* the no-masker control condition).

The effect of maskers on RMS errors also depended on whether the masker locations were fixed or mixed within a run. This is better illustrated in Figure 2, which shows the increase in the RMS error averaged across all masker patterns and across either all target locations (all), across the target locations where the target was presented with a co-located masker (on-masker), or across the target locations where the target was not co-located with a masker (off-masker). The availability of *a priori* information about masker locations had a main effect in the reverberant condition. Averaged across all target locations, the RMS error reduction in the Fixed condition compared to the Mixed condition was 1.1° (or 31%). The effect of *a priori* knowledge was even larger when only the off-masker target locations were considered, reducing the RMS error by 2° (or 51%). On the other hand, the *a priori* information had a modest effect on the on-masker targets, increasing the RMS error by 0.6°. The effect is also modest under the anechoic condition. A repeated measures analysis of variance (ANOVA) confirms that the *a priori* information only has a significant main effect on localisation accuracy in the reverberant condition when all target locations are considered [$F = 5.56, p < .05$] or only the off-masker locations are considered [$F = 13, p < .005$]. No significant main effect was found for the on-masker data and in the anechoic condition.

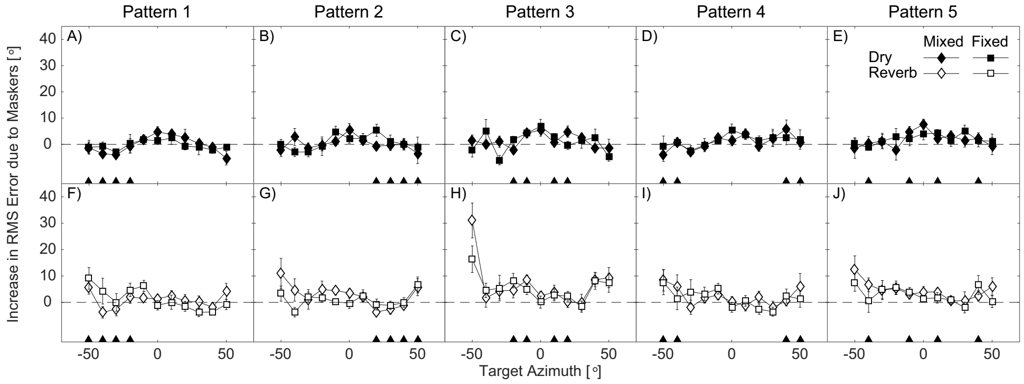

Figure 3: Across-participant average (± SEM) of the increases in RMS errors (*re.* the no-masker control condition) as a function of the target location. Masker locations are indicated by the filled triangles along the abscissa.

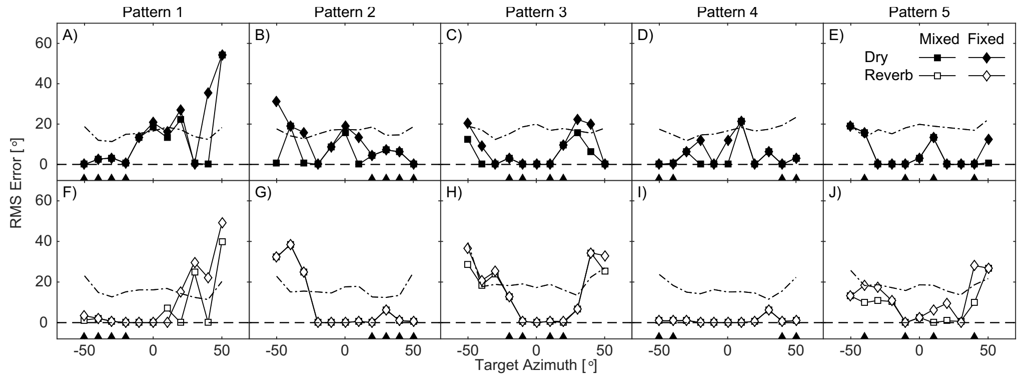

Figure 4: RMS errors of the proposed model as a function of the target location. Average listeners' RMS errors are plotted as dotted lines in each panel for comparison.

4.2. Results of model simulation

The proposed model achieved 100% target location accuracy in the no-masker control runs under both anechoic and reverberant conditions, compared to an average of 15° error by listeners.

Model RMS errors in masker conditions are plotted in Figure 4, which also shows the average listener data for comparison. The machine system outperformed listeners in many conditions. This can be largely attributed to the use of source models in the system (Section 3.3), without which performance was poor in this relatively challenging localisation task[1].

The system did not perform well when the target was not co-located with a masker, especially for masker patterns 1–3 in which the maskers were more clustered. Apparently, in such conditions the maskers disrupted the localisation cues for the target more than when maskers were distributed in space, and the DNN failed to indicate a high probability of a sound source occurring at the target location.

Figure 5 shows that the machine system also benefitted from prior knowledge about the masker locations. The average error reduction in the Fixed condition compared to the Mixed condition was 3.2° in the anechoic condition and 2° in the reverberant condition (both significant with $p < .005$). As in listeners' data, this error reduction became larger when only the off-masker locations were considered.

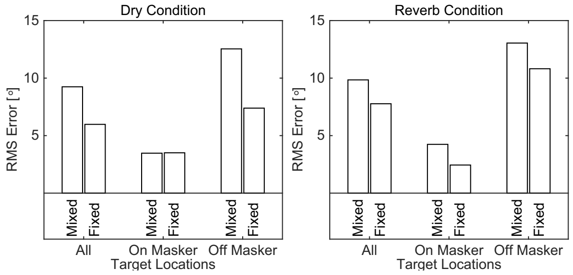

Figure 5: Effect of maskers on localisation accuracy of the proposed model shown as RMS error.

5. Conclusions

Listeners are able to exploit prior information about masker locations in Kopco et al.'s [3] task when listening over headphones, but only in reverberant conditions and when the target speech was not co-located with a masker. A computational model was able to match human data to some extent by exploiting the sources of knowledge available to listeners in this scenario, i.e. speaker characteristics and masker locations.

Future work will assess the benefit of individualised HRIRs in this task. The role of head movements will also be investigated, thus allowing listener performance to be compared with a DNN-based localisation system that uses head movements [9].

[1] Each masker was equal in level to the target, which is equivalent to about -12 dB target-to-masker-ratio with four maskers.

6. References

[1] C. Cherry, "Some experiments on the recognition of speech with one and two ears," *J. Acoust. Soc. Am.*, vol. 24, pp. 554–559, 1953.

[2] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[3] N. Kopco, V. Best, and S. Carlile, "Speech localization in a multitalker mixture," *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 1450–1457, 2010.

[4] N. Ma, G. J. Brown, and J. A. Gonzalez, "Exploiting top-down source models to improve binaural localisation of multiple sources in reverberant environments," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 160–164.

[5] G. J. Kidd, V. Best, and C. R. Mason, "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3793–3802, 2008.

[6] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *Proc. 130th Conv. Audio Eng. Soc.*, 2011.

[7] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, 2010.

[8] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley/IEEE Press, 2006.

[9] N. Ma, G. J. Brown, and T. May, "Robust localisation of multiple speakers exploiting deep neural networks and head movements," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3302–3306.

[10] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audiovisual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, pp. 2421–2424, 2006.

[11] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 845–848.

[12] G. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech. Lang.*, vol. 8, pp. 297–336, 1994.

[13] D. R. Begault, "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *Journal of the Audio Engineering Society*, vol. 40, no. 11, pp. 895–903, 1992.