Active hearing, active speaking

Martin Cooke, Yan-Chen Lu, Youyi Lu, Radu Horaud

Active hearing, active speaking

Martin Cooke[1], Yan-Chen Lu[1], Youyi Lu[1] and Radu Horaud[2]

[1]*Speech and Hearing Research, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK*

[2]*INRIA Rhône-Alpes, 655, Ave. de l'Europe, 38330 Montbonnot, France*

ABSTRACT

A static view of the world permeates most research in speech and hearing. In this idealised situation, sources don't move and neither do listeners; the acoustic environment doesn't change; and speakers speak without any effect of auditory input from their own voice or other speakers. Corpora for speech research and most behavioural tasks have grown to reflect the static viewpoint. Yet it is clear that speech and hearing takes place in a world where none of the static assumptions hold, or at least not for long. The dynamic view complicates traditional signal processing approaches, and renders conventional evaluation processes unrepeatable since the observer's dynamics influence the signals received at the ears. However, the dynamic viewpoint also provides many opportunities for active processes to exploit. Some of these, such as the use of head movements to resolve front-back confusions, are well-known, while others exist solely as hypotheses. This paper reviews known and potential benefits of active processes in both hearing and speech production, and goes on to describe two recent studies which demonstrate the value of such processes. The first shows how dynamic cues can be used to estimate distance in an acoustic environment. The second demonstrates that the changes in speech production which take place when other speakers are active result in increased glimpsing opportunities at the ear of the interlocutor.

INTRODUCTION

The listening problem

The classical account of the issues faced by listeners has been illustrated by the *cocktail party problem (CPP)*: how do listeners manage to decipher speech in the presence of other sound sources, including competing talkers (Cherry, 1953)? The CPP has inspired both behavioural and computational studies which have focused on the use of cues such as fundamental frequency and interaural time differences, invoking principles such as old-plus-new and continuity to handle the introduction and tracking of new sources. The CPP has led to a focus on the existence of multiple sources and, in algorithmic terms, a welcome move away from the idea, prevalent in speech enhancement, that 'noise' is a quasi-stationary interference which can be suppressed. Several corpora based on the CPP now exist, and in a recent evaluation of computational techniques for identifying utterances in the presence of another talker, one approach achieved super-human performance in some conditions (Kristjansson *et al*., 2006).

But is the CPP a reasonable description of the true listening problem? By focusing mainly on the idea of attending to a single source amongst multiple non-stationary sources, the CPP account has downplayed a key aspect of auditory scenes, namely their *dynamics.* Consider instead the following scenario. You arrive at an airport. Your route takes you through the wide, high check-in hall and then through narrow tunnel-like corridors lined with glass and steel to the more comfortably furnished departure lounge, then down even narrower corridors

and on to the plane itself. All the way, you hear a succession of announcements on a variety of loudspeaker systems and pass knots of talkative passengers and music emanating from cafes and bars before listening to the captain's announcement to the accompaniment of engine roar. Yet, without realising it, you've been able to carry on a normal conversation with a colleague the whole time. How representative is the airport scenario? Consider cycling through any busy city or making use of a crowded public transport system. In these situations, the ability to answer questions about the immediate environment via hearing might be critically important.

What characterises this listening/communicative experience is change. The number of competing sources is never the same for very long; sources are often mobile; listeners make both fine and coarse movements so the head related transfer function is continually varying; sources gradually or suddenly enter the mix and exit similarly; transfer functions of the various transmission systems vary; reverberation characteristics vary with source-listener geometry; visual information may be unavailable at some times; attentional demands may vary as cognitive load changes (consider driving). The real listening problem is dynamic. By contrast, the CPP appears almost static.

Active processes

One response to the dynamic auditory scene interpretation problem may be to appeal to *active processes.* An active process might be defined as one which results in a purposeful response to the immediate environment. Figure 1 illustrates the "static/dynamic environment, active/passive response" distinctions by identifying some types of human or computational behaviour which may be appropriate or possible. For example, a passive response to a dynamic environment may be to treat the consequences of dynamics as an additional noise term. An active response in a static environment may be to use receiver motion to 'triangulate' the source location.

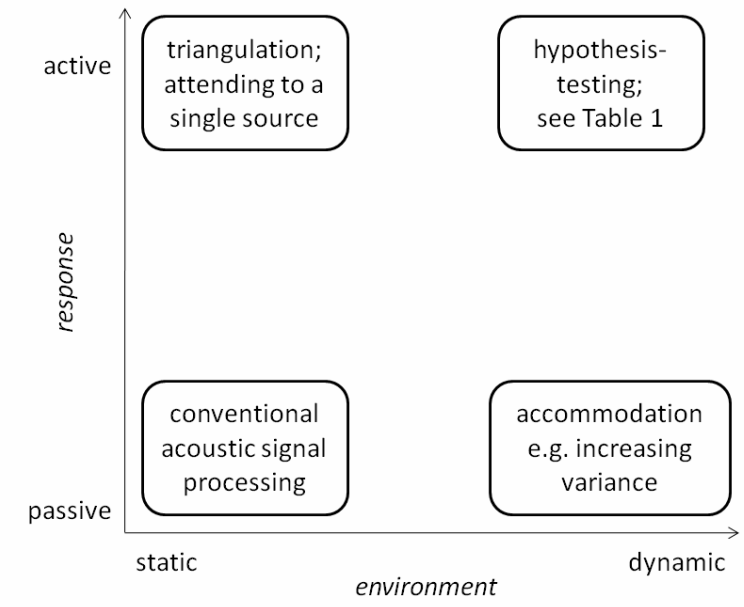

Fig 1: Static/dynamic environments, passive/active responses.

Active processes have great potential in hearing. A listener might use fine head movements to disambiguate possible source locations, or gross head/body movements to improve the level of the target source or reduce the contributions of interfering sources or reverberant components. Active attentional processes could select and track one source amongst many. A speaker might modify his speech productions to improve the assumed SNR at the ears of the listener. Table 1 identifies possible active processes together with evidence of their utility where it exists.

Table 1: Potential active processes in hearing and speaking.

Process	**Purpose**
Fine head movements (Wallach, 1940; Thurlow et al., 1967; Mackenson, 2004)	Disambiguate front-back confusions
Gross head movements (Loomis et al., 1990)	Improve SNR at best ear Use head shadow to reduce interferer Locate target in high resolution part of azimuthal plane
Body translation	Improve target signal (e.g. move closer) Improve line of sight for visual cues Reduce level of interference (e.g. move away) Reduce degree of reverberation (e.g. move away) Increase spatial separation between target and interferer
Motion (Speigle and Loomis, 1993; Ashmead et al., 1995)	Provide multiple samples for estimation of azimuth and distance via cues such as acoustic tau or motion parallax
Speech production modifications (Lombard, 1911; Lu and Cooke, submitted)	Compensate for energetic and informational masking at the listener's ears
Conversational processes (Local et al., 1986)	Turn-taking, signal agreement, self-repair, turn-completion, alignment

Parallels with active vision

The CPP mentioned above as well as its generalization to dynamic auditory analysis is very similar to the kind of tasks that vision has to solve. Indeed, the amount of visual information that falls onto our retinas is tremendous and it is barely static. As well as motion of objects of interest, it is necessary to take into account observer motion (egomotion) which consists of a combination of eye movements (3-4 saccades per second) as well as head and body movements. as well as most of the objects of interest. Moreover, we perceive a dynamic three-dimensional world through sequences of two-dimensional images. Therefore, from the very beginning, computer vision researchers have identified the analysis of motion as one of the fundamental problems to be solved.

The first problem that has been addressed, and solved, is the very simple situation of a moving observer looking at a static scene. This is known as the structure-from-motion problem and it has been one of the most investigated topics in vision (Maybank, 1993). The problem is twofold, i.e., find a one-to-one matching between pixels in one image and pixels in the next image that correspond to the same 3D scene point, and estimate the observer's motion. Once the matching and motion problems are solved, it is possible to recover depth at every matched point via triangulation. The combination of image processing and analysis (optical flow) with algebraic projective geometry and with robust statistics is at the heart of the modern approach to structure-from-motion.

The generalization to multiple moving objects is far from trivial and is still under investigation (Costeira and Kanade, 1998). One has to solve the structure-from-motion problem just mentioned for every single moving object. There are rigid objects, such as a car, articulated objects such as humans and animals, and more complex objects that move and deform in a completely unpredictable way, such as a flag waving in the wind. Although it is possible to have physical models for such situations, estimating the objects' parameters and tracking them over time seem to be a tremendously difficult task. Currently only simple situations have been addressed (Zelnik-Manor et al., 2006).

The segmentation of a scene into several objects can be tackled within a Bayesian probabilistic framework, and one of the most promising approaches is unsupervised clustering. Such an approach can accommodate the geometric and kinematic modelling of objects, of motion, and of the image formation process. It can also accommodate robust statistics and in particular with outlier rejection, which is crucial. The multi-body segmentation problem mentioned above is very relevant in the context of fusion of visual and auditory stimuli. Indeed, segmenting a scene into distinct objects and tracking them over time is one where each modality can help resolve ambiguities and constrain hypotheses in the other.

ACTIVE HEARING

General issues in active computational hearing

Attempts to incorporate active processes into hearing technology face a number of challenges. First, there is a paucity of behavioural or neurophysiological data with which to construct algorithms, especially for the complex, multiple sound source environments typical of everyday listening (Palmer et al., 2007). Second, there is usually a need to model attentional selection at some level in order to determine which auditory object to track and for how long. Third, features derived from interaural signals suffer from motion blur and there is also the possibility of actuator noise if the active hearing device is incorporated into a robotic platform (Okuno and Nakadai, 2000). A further issue is the need for real-time processing and indeed rapid 'gisting' of the auditory scene so that an appropriate active response can be initiated as early as possible (Harding et al., in press). Finally, evaluation is not trivial since the listener is part of the scenario, and any active movement changes the signals sensed in the environment, and also because ground-truth measurements of the ongoing listener-source geometry requires specialist equipment.

Case study: active localisation in azimuth and distance

To illustrate the potential of active hearing, the problem of dynamic source localisation is considered. Full source localisation involves the estimation of both azimuth and distance. While many computational models for the estimation of source azimuth relative to a receiver

exist, few consider the problem of distance, which is our focus here (see also Berglund and Sitte, 2005). A fuller account is available in Lu *et al.* (2007).

Potential cues to distance can be relative or absolute. The former include loudness and source spectrum, but prior information about the sound source is required to estimate absolute distance. For anechoic conditions, the loudness cue can be used to determine changes in the distance of a constant amplitude sound source according to the inverse square law. Differential absorption of frequencies along the propagation path is the major source of spectral cues. Familiarity, binaural information and reverberation deliver absolute cues. If the listener is sufficiently familiar with the sound source, relative cues can be used to judge absolute distance. Listener familiarity with both the source signals and the acoustic environment is clearly a key factor in any model of auditory distance perception. For near-field listening (distance < 1m), binaural cues based on interaural time and intensity differences provide not only directional but also distance information. A recent review of distance estimation is provided by Zahorik *et al.* (2005).

The cues described so far assume that the listener and source are stationary, but their (relative) motion can provide additional cues to auditory distance perception (figure 2). Listener motion creates a changing azimuth (A_{t-1} at time *t-1* to A_t at time *t* with respect to a stationary source) known as *motion parallax* which can be used to estimate source distance Δ_t via the listener translation distance, S. It has also been suggested that motion-induced rate of change of intensity (from I_{t-1} at time *t-1* to I_t at time *t* over listener movement *S)* can provide listeners with reliable distance information (Speigle and Loomis, 1993). This cue, known as *acoustic τ* (time-to-contact), may also be expressed as a ratio of distance to velocity when velocity is constant. Speigle and Loomis found that dynamic cues of motion parallax and acoustic τ influence an observer's judgment of source distance above and beyond static cues. However, their experiments involved relatively simple auditory scenes and it is an open question as to whether dynamic cues are more or less useful in realistic environments. The calculation of acoustic τ needs prior distance information from another cue such as motion parallax and hence must be exploited within a framework of multiple, coupled cues.

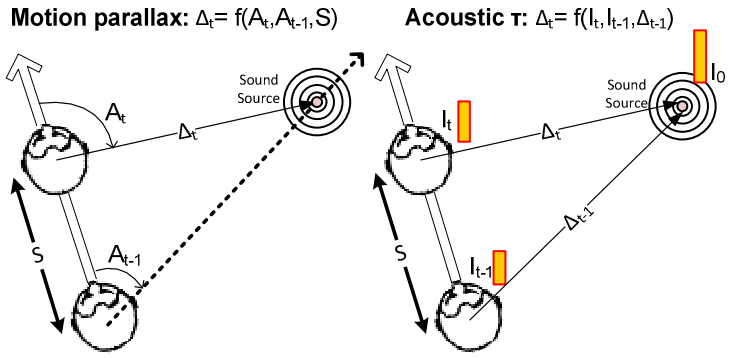

Fig 2: Dynamic auditory cues to distance: motion parallax (left) and acoustic τ (right).

A computational model of dynamic source localisation is illustrated in figure 3 (Lu et al., 2007). Dynamic cues are generated from successive measurements of cross-correlation and intensity derived from a model of peripheral auditory processing. Distance inference is based on triangulation for motion parallax and an adaptation of the inverse square law for intensity-based cues.

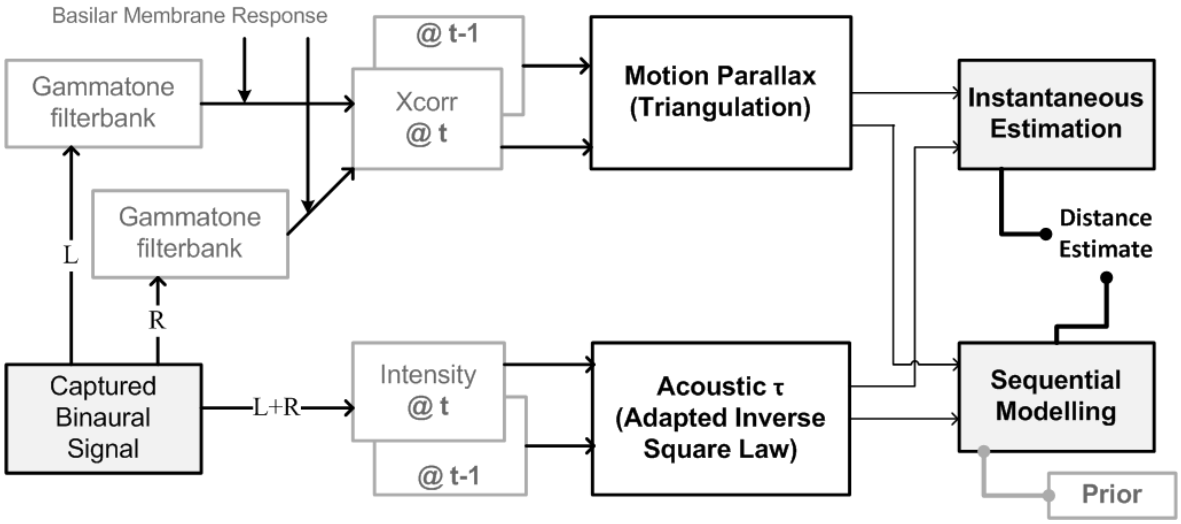

Fig 3: Computational model for distance and azimuth estimation (Lu et al., 2007).

Motion parallax and *acoustic* τ are combined and tracked through time using a technique known as *particle filtering*, a probabilistic sequential model for active hearing and vision applications. Particle filters have been used for robust acoustic source tracking in reverberant environments (Ward *et al*., 2003). In the particle filtering framework, each particle represents a hypothesis about the estimates of interest. Here, for example, each particle is a triple of random variables representing hypotheses for distance, azimuth and source intensity. Particles have weights which represent the belief that their associated hypotheses are correct.

Figure 4 shows the output of the model at a number of time steps for a static source and a moving receiver. The grey-level 'cloud' represents smoothed location estimates from the collection of particles, weighted by likelihood.

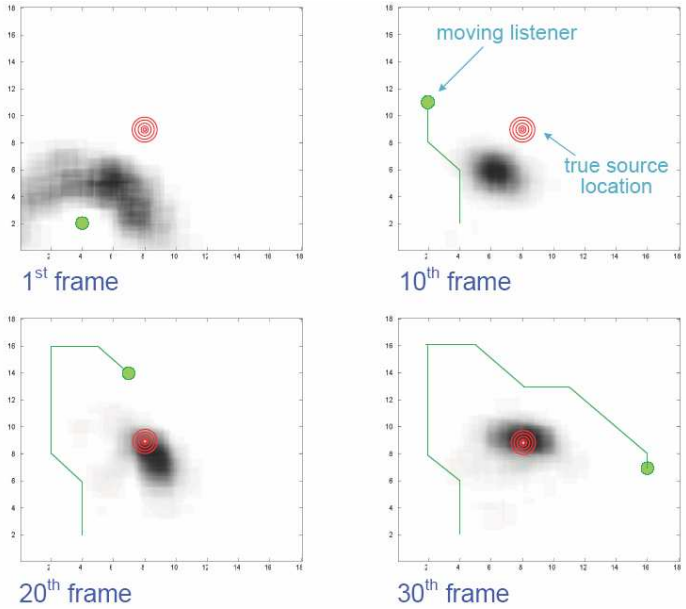

Fig 4: An illustration of active location estimation for a moving listener.

Formal evaluation of the algorithm employed a simulated acoustic environment (Campbell *et al.* 2005) of size 18 x 18 x 2.5 m. As well as an anechoic condition, two reverberant surfaces, "acoustic plaster" and "platform floor wooden", were used, with mean estimated T_{60} reverberation times of 0.34s and 0.51s respectively. The point sound source was a static pink noise source located as shown in figure 4. On each of 100 runs, the simulated listener moved for 50 time steps. Table 2 shows mean distance estimation errors for a number of algorithmic variants. The first two rows are 'instantaneous' estimates made without the use of the sequential particle filtering algorithm, for motion parallax alone and with acoustic tau. While the latter leads to a significant error reduction (less so for reverberant conditions), the error is still rather large. The remaining lines show the effect of adding particle filtering, which results in much better estimates. Although behavioural data on these stimuli are not available and are difficult to obtain given the moving listener scenario, it is notable that human distance estimation for this range of distance is quite poor. The power function approximation in Zahorik *et al*. (2005), while not strictly applicable since it is based on an average of many studies with different conditions, suggests a mean distance estimation error of around 3.5 m for listeners.

Table 2: Distance estimation errors in metres.

Model		*Anechoic*	*RT_{60}=0.3s*	*RT_{60}=0.5s*
No sequential model	Motion parallax	8.3	7.2	7.4
	Motion parallax + acoustic tau	5.6	5.2	6.2
With sequential model	Motion parallax	1.9	2.0	3.4
	Motion parallax + acoustic tau	1.4	1.8	4.4

The model of Lu et al. (2007) described above currently has a number of limitations. First, offline estimates of a pair of parameters which vary with room reverberation are required. Listeners appear to be able to learn reverberation "online" (Shinn-Cunningham, 2000). Second, egomotion is assumed to be known. However, in principle, egomotion can be estimated as part of the tracking process. Further, the model needs to be evaluated with non-simulated data in more complex environments where more than one source is active.

ACTIVE SPEAKING

Case study: Lombard glimpses

It has long been known that noise affects speech production and leads to a variety of acoustic consequences collectively known as the *Lombard effect* (Lombard, 1911). For example, increases in level, fundamental frequency, vowel duration and first formant frequency are usually observed, although there is some inter-speaker variation (Hanley and Steer, 1949; Summers et al., 1988; Junqua, 1993). A recent study by Lu and Cooke (submitted) examined the effect of N-talker noise on sentence production for a range of values of N varying from 1 (competing talker) to infinity (speech-shaped noise). The effect of noise on speech production increased with both the spectral density and level of the noise. Interestingly, increases in both spectral density and level result in an increase in the energetic masking effect of noise. As was found by Dreher and O'Neill (1957), noise-induced speech was always more intelligible than

speech produced in quiet when presented in a background of stationary noise. Lu and Cooke also demonstrated that the gain in intelligibility increased with spectral density and level, suggesting that talkers modify their productions in a proportionate fashion to ameliorate energetic masking at the ears of the listener.

The key question arising from studies of the Lombard effect is: why is speech produced in the presence of noise more intelligible than speech produced in quiet? Several aspects of Lombard speech might contribute to the intelligibility gain. The first 4 columns of table 3 demonstrate the extent of increases in a number of acoustic measurements as a function of the level of speech-shaped noise presented during speech production. The fifth column shows the increase in intelligibility for noise-induced speech over speech produced in quiet, in the presence of stationary noise added at an SNR of -9 dB. Increases in production level (column 1) will help to increase overall SNR. However, even when overall level differences are removed, as is the case for the constant SNR used here, Lombard speech is still significantly more intelligible. Increases in level alone appear to be insufficient to overcome the effect of background noise level: a level increase of 7.1 dB for a noise background of 82 dB SPL rises by only 2.4 dB for a noise background of 96 dB. Other factors which may contribute are spectral and temporal changes. Lombard speech is typically somewhat slower than speech produced in quiet (column 2) and energy is shifted to higher frequencies (column 3), partly as a result of increases in F0 (column 4). The auditory spectrograms in the left column of figure 5 illustrate some of the changes in spectro-temporal energy distribution.

Table 3: Effect of speech shaped noise on speech production and intelligibility

Noise level during production (dB SPL)	*Increase in speech production level (dB)*	*Increase in duration (%)*	*Increase in spectral centre of gravity (%)*	*Increase in F0 (semitones)*	*Increase in intelligibility (%)*	*Increase in glimpses (%)*	*Increase in glimpses per frame (%)*
82	7.1	3.3	32	1.7	58	25	21
89	8.3	5.3	34	2.0	-	-	-
96	9.5	7.6	38	2.5	68	36	27

One hypothesis for the increased intelligibility of Lombard speech is that speakers attempt to compensate for the masking effect of background noise at the listener's ears by modifying their articulations in such as way to increase the "glimpsing" opportunities for the listener. Lu and Cooke tested this idea using a glimpsing model which simulates the effect of energetic masking (Cooke, 2006). Columns 6 and 7 of table 3 show that both the overall proportion of glimpses and the duration-independent proportion of glimpses per unit time increase with background noise level. These figures do not incorporate the effect of level differences and show that both durational increases and changes in spectral energy distribution result in more opportunities to glimpse Lombard speech in noise. The right column of figure 5 depicts speech glimpses for the example utterances. Both the increased glimpsing opportunities and the overall shift to higher frequencies are apparent.

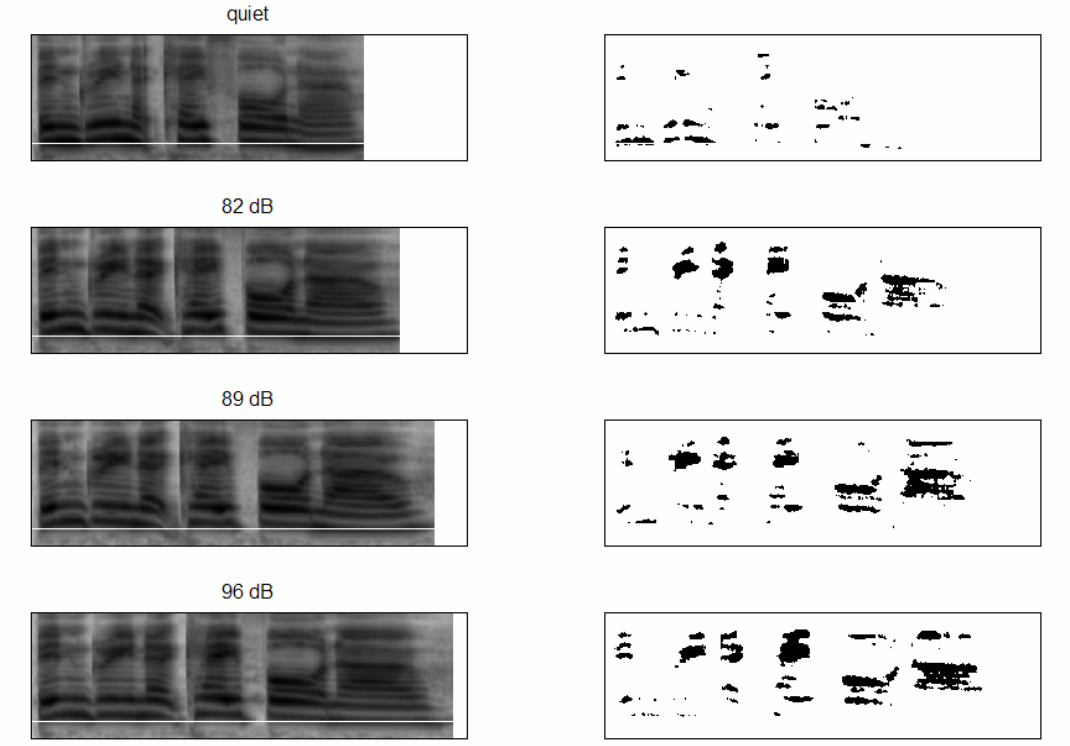

Fig 5: Auditory spectrograms and glimpses for the sentence "bin green at K 4 now" spoken by a female in quiet and in the presence of 3 levels of speech-shaped noise. Some effects of noise level on duration and spectral tilt are visible, as is an overall increase in F0 (horizontal lines indicate a frequency of 200 Hz).

The notion that "Lombard glimpses" can account for intelligibility gains is given additional support by a large-scale comparison of speech produced in a variety of noise backgrounds, including spectrally-sparse signals such as competing speech (figure 6).

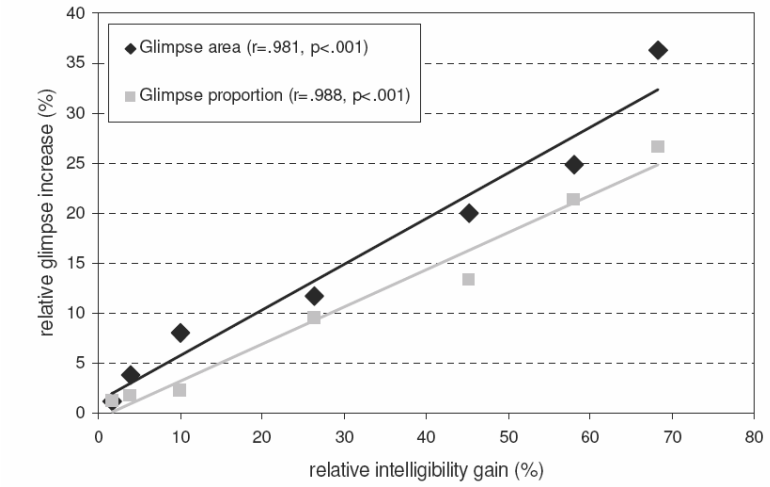

Fig 6: A comparison of intelligibility gains and increases in glimpsing opportunities for Lombard speech over speech produced in quiet, for a number of noise backgrounds. From Lu and Cooke (submitted).

It is, of course, possible that the intelligibility gains found by Lu and Cooke were fortuitous, since the upward shift in spectral centre of gravity is advantageous for the types of noise backgrounds used (since their long-term spectra were speech-shaped, decreasing with frequency). A follow-up study addressed the issue of whether speakers *actively* attempt to place spectral information in locations where it is less likely to be masked. That study compared the effect on speech production of low-pass and high-pass filtered noise. For low-pass noise, the increases in level, spectral centre of gravity, mean F0 and mean F1 were similar to those found in wideband speech-shaped noise when presented at the same overall level. However, for high-pass noise, significantly smaller increases in all 4 parameters were found. It is notable that speakers did not produce speech in which these parameters *decreased*, suggesting that background noise induces speech production changes with both a passive component (the original Lombard 'reflex') and an active component, which acts to resist the scale of the increases in level, centre of gravity etc.

Lu and Cooke found only weak evidence for the hypothesis that speakers modify their productions to decrease the informational masking (IM) effect for the listener. The number of duration of short pauses was higher for a single competing talker than for multitalker babble or stationary noise. However, speech produced in a matched competing talker background (i.e. the background used to induce the speech originally) was no more intelligible when presented in a competing talker background than speech produced in an unmatched competing talker background.

The task used by Lu and Cooke involved no communicative element. Different effects on speech production may surface when the communication of information is at stake. It is also possible that speakers are unable to execute strategies to minimise IM sufficiently rapidly. In this regard, studies of talk-in-interaction (e.g. Local et al., 1986) are relevant. In conversations, speakers (who are also listeners), are quite capable of 'aligning' conversational elements to signal agreement or complete each other's turns, with quite precise timing. Much of this may be due to a well-developed predictive ability linked to an internal model (see review in Moore, 2007). Of course, a critical difference in the masking situation is that the 'competing' speech message may not be the object of attention.

The relationship between the acoustic modifications produced by noise-induced speech and other forms of active speaking, such as clear speech (Chen, 1980; Picheny et al., 1986) is currently unclear. It is of interest to discover whether speakers use similar strategies to convey information in each of these cases.

APPLICATIONS AND FURTHER RESEARCH DIRECTIONS

Active hearing could find application in any situation where motion signals are available and could be incorporated into wearable audio devices such as hearing prostheses sensitive to listener movement. Active speaking promises to inform the next generation of speech synthesis technology by defining how synthesisers might dynamically modify their output as a function of prevailing noise conditions and with knowledge of a listener's hearing impairment.

While active hearing has great potential in dealing with dynamic auditory scenes and integration with active vision systems, it remains unclear how important active processes are. Further progress will depend on both the availability of 'ground truth' data as well as behavioural and neurophysiological studies in complex environments.

ACKNOWLEDGEMENT

This research was supported by the EU STREP “Perception on Purpose (POP)”. Thanks to Sue Harding for providing useful comments on the manuscript.

REFERENCES

Ashmead, D. H., Davis, D. L., and Northington, A. (**1995**). "Contribution of listeners' approaching motion to auditory distance perception," J. Exp. Psychol. Hum. Percept. Perform., **21**, 239-56.

Berglund, E. and Sitte, J. (**2005**). “Sound source localisation through active audition,” Proc. IEEE Int. Conf. Intelligent Robots and Systems, 509-514.

Campbell, D. R., Palomäki, K. J. and Brown, G. (**2005**). "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," Computing and Information Systems J., **9,** 48-51.

Chen, F. R. (**1980**). “Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level,” Unpublished master’s thesis, Massachusetts Institute of Technology, Cambridge.

Cherry, E. C., (**1953**). Some experiments on the recognition of speech with one and with two ears, J. Acoust. Soc. Am., **25**, 975-979.

Cooke, M. P. (**2006**). “A glimpsing model of speech perception in noise,” J. Acoust. Soc. Am. **119**, 1562-1573.

Costeira, J. and Kanade, T. (**1998**). “A multibody factorization method for independently moving objects,” Int. J. Computer Vision, **29**, 159-179.

Dreher, J. J. and O’Neill, J. (**1957**). “Effects of ambient noise on speaker intelligibility for words and phrases,” J. Acoust. Soc. Am., **29**, 1320-1323.

Hanley, T. D. and Steer, M. D. (**1949**). “Effect of distracting noise upon speaking rate, duration, and intensity,” J. Speech Hear. Disord. **14**, 363-368.

Harding, S., Cooke, M. P., and König, P. (in press). “Auditory gist perception: an alternative to attentional selection of auditory streams?” Lecture Notes in Artificial Intelligence.

Junqua, J. C. (**1993**). “The Lombard reflex and its role on human listeners and automatic speech recognizers,” J. Acoust. Soc. Am., **93**, 510-524.

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S. and Gopinath, R. (**2006**). “Super-human multi-talker speech recognition: the IBM 2006 Speech Separation Challenge system,” Proc. Interspeech, Pittsburgh, PA.

Local, J., Kelly, J., and Wells, W. (**1986**). “Towards a phonology of conversation: turn-taking in urban Tyneside speech,” J. Linguistics, **22**, 411-437.

Lombard, E. (**1911**). “Le signe de l’elevation de la voix. Annales des Maladies de l’Oreille,” du Larynx, du Nez et du Pharynx, **37**, 101-119.

Loomis, J.M., Hebert, C., and Cicinelli, J.G. (**1990**). “Active localization of virtual sounds,” J. Acoust. Soc. Am. **88**, 1757-1764.

Lu, Y., and Cooke, M. P. (submitted). “Speech production modifications produced by competing talkers, babble and stationary noise, submitted to J. Acoust. Soc. Am.

Lu, Y.-C., Cooke, M. P., and Christensen, H. (**2007**). "Active binaural distance estimation for dynamic sources," Interspeech, Antwerp, Belgium.

Mackenson, P. (**2004**). "Auditive localization. Head movements, an additional cue in localization," Ph. D. Thesis, Technical University of Berlin.

Maybank, S. J. (**1993**). "Theory of Reconstruction from Image Motion," in *Springer Series in Information Sciences, vol 28*. Springer-Verlag.

Moore, R. K. (**2007**). "Spoken language processing: piecing together the puzzle," Speech Communication, **49**, 418-435.

Okuno, H. G. and Nakadai, K. (**2003**). "Real-Time Sound Source Localization and Separation Based on Active Audio-Visual Integration," in *Computational Methods in Neural Modeling*, Lecture Notes in Computer Science **2686**. Springer.

Palmer A. R., Hall D. A., Sumner C. J., Barrett D. J. K., Jones S., Nakamoto K., Moore D. R. (**2007**). "Some investigations into non passive listening," Hearing Research 229, 148-157.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Speech Lang. Hear. Res. **29**, 434-446.

Shinn-Cunningham, B. G. (**2000**). "Learning reverberation: Considerations for spatial auditory displays," *Proc. International Conference on Auditory Display*, Atlanta, GA, 126-134.

Speigle, J. M. and Loomis, J. M. (**1993**). "Auditory distance perception by translating observers," *Proc. IEEE Symposium on research frontiers in virtual reality*, San Jose, CA, 92-99.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analysis," J. Acoust. Soc. Am., **84**, 917-928.

Thurlow, W. R., Mangels, J. W., and Runge, P. S. (**1967**). "Head movements during sound localization," J Acoust Soc Am, **42**, 489-493.

Wallach, H. (**1940**). "The role of head movements and vestibular and visual cues on sound localization," J . Exp. Psychol., **27**, 339-368.

Ward, D. B., Lehmann, E. A., and Williamson, R. C. (**2003**). "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Processing*, **11**, 826-836.

Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (**2005**). "Auditory Distance Perception in Humans: A Summary of Past and Present Research," *Acta Acustica united with Acustica*, **91**, 409-420.

Zelnik-Manor, L., Machline, M., and Irani, M. (**2006**). "Multi-body Factorization With Uncertainty: Revisiting Motion Consistency," Int. J. Computer Vision, **68**, 27-41.