Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 540–542

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Introduction to the Special Issue on The listening talker: context-dependent speech production and perception

Speech is efficient and robust, and remains the method of choice for human communication. However, speech is often delivered by machines that are not aware of the context of the listener, potentially reducing the effectiveness of the communication process. The lack of context-awareness is particularly severe in one-way, output-oriented technologies such talking GPS and public address systems. Such machines lack an essential ingredient of face-to-face human interaction: feedback. *When people talk, they also listen. When machines talk, they do not listen*. As a result, there is no guarantee that the intended message is intelligible, appropriate or well-timed. The current generation of speech output technology is deaf, incapable of adapting to the listener's context, inefficient in use and lacks the naturalness that comes from rapid appreciation of the speaker–listener environment. Insufficient adaptation to the environment also has become a significant problem in human-to-human communications. The ability to communicate from anywhere to anywhere has resulted in a listening environment that is often challenging and rapidly-changing. The far-end party is generally only partly aware of the near-end environment, an issue that is exacerbated by the rapid improvement of noise suppression technology. The far-end party will therefore not adjust her speech to the environment faced by the near-end user.

This Special Issue on the *Listening Talker* brings together one review article and nine research reports that address the problem of the listening talker. The collection contains behavioural studies on the production and perceptual consequences of modified speech in both acoustic and visual domains, algorithms for post-processing speech with the aim of enhancing intelligibility, and text-to-speech synthesis systems designed to produce more robust speech output.

To start the Special Issue, *Cooke, King, Garnier and Aubanel* provide an extensive review of context-induced human speech modifications and algorithms designed to enhance intelligibility. Organised around four proposed goals of speech modification – improving audibility, increasing coherence, enhancing linguistic information and decreasing cognitive effort – the review concludes with a catalogue of speech modification candidates that might be explored in future behavioural and algorithmic studies.

Papers by both *Tweedy and Culling* and *Garnier and Henrich* examine active (i.e., non-automatic) processes in Lombard speech. While studies of the effect of noise on speech production began more than a century ago (Lombard, 1911), the article by Tweedy and Culling is the first to examine the influence of the perceived signal-to-noise ratio (SNR) experienced by the interlocutor on a talker's voice intensity. Faced with the problem of ensuring accurate control of interlocutor voice intensity, Tweedy and Culling introduce a novel approach in which pre-recorded conversational prompts are presented under experimenter control to simulate a live interaction. Their key finding is that the SNR of the interlocutor does not produce an effect on a speaker's voice intensity. Tweedy and Culling conclude that, by extension, for a fixed noise level, amplifying the interlocutor's voice is unlikely to prevent speakers shouting into mobile phones.

Garnier and Henrich consider the hypothesis that speakers modify their output in noise in ways that enhance the acoustic contrast between their speech and the background noise. They looked for evidence of three strategies: boosting (increasing speech level in parts of the spectrum dominated by noise), bypass (shifting energy to spectral regions not dominated by noise) or modulation of f_0 and intensity (enhancing the detectability of the target speech). They find no evidence for masker-specific boosting or bypass strategies over and above a general increase in vocal intensity, but

0885-2308/$ – see front matter
http://dx.doi.org/10.1016/j.csl.2013.11.002

do observe increases in f_0 and intensity modulations which they argue reflect an attempt on the part of the speaker to improve intelligibility.

Audible speech is often accompanied by visual information of the face of the speaker, which aids the listener. If noise is present, the importance of this visual information is increased. This increased reliance on visual information is compounded by changes in the facial movements when speech is produced in a noisy environment. As is discussed in both the papers of *Kim and Davis* and *Alexanderson and Beskow*, the visual articulation becomes more pronounced in a noisy environment. Alexanderson and Beskow show that this stronger visual articulation results in significantly increased intelligibility. Interestingly, their 3D avatar achieved a similar improvement in intelligibility as natural video for noisy environments, suggesting that avatars can be used for practical applications such as the learning of lipreading. They also found that if speech produced in a quiet environment is combined with the video channel (animated or natural) of speech produced in a noisy environment, its intelligibility is also increased. The mismatch must be limited, however: if speech in a quiet environment is combined with exaggerated visual channels such as that of whispered speech, the audio-visual signal is perceived as unnatural.

Kim and Davis focus on both the acoustic and visual consistency and distinctiveness of speech produced in noise. Their results suggest that the third formant may be more consistent when produced in noise, but this is not the case for the first two formants. The degree of dispersion of the formants for speech produced in quiet and in noise is the same. The main effect of the noisy environment that they observed is that the formant frequencies are increased when speech is produced in noise. For visual speech, Kim and Davis demonstrate larger jaw and mouth motion in noise compared to quiet conditions, but no increase in the consistency of speech production.

Two papers present human-inspired speech modifications that aim to increase speech intelligibility and operate as post-filters. The article by *Jokinen, Takanen, Vainio and Alku* is based on adaptation to noise characteristics as well as to the fundamental frequency of the talker, while *Godoy, Koutsogiannaki and Stylianou* introduce speech modifications that are both noise- and talker-independent. Jokinen et al. propose a low complexity post-filtering algorithm that imitates the Lombard effect. The algorithm consists of four stages: spectral tilt compensation, formant sharpening, fundamental frequency and noise-adaptive high pass filtering and gain control. Their approach is applied to narrow band speech for mobile communications. The authors provide evidence from listening tests that demonstrates that the proposed system outperforms other post-filtering approaches in terms of intelligibility as well as subjective quality.

Godoy et al. investigate the intelligibility gains from Lombard and ‘clear’ speaking styles. Increments of spectral energy around the second and third formants and vowel space expansion are confirmed by acoustic analysis for the Lombard and clear speaking styles respectively. Based on these findings, Godoy et al. suggest the incorporation of a frequency warping mechanism for vowel space expansion into an existing Lombard speech-inspired spectral shaping algorithm. Listening tests showed that although the vowel space of the unmodified speech was expanded following the patterns of clear speech, this did not result in additional intelligibility gains over those produced by the spectral shaping algorithm.

The three papers concerning speech synthesis all take their inspiration from human performance, with articles by *Raitio, Suni, Vainio and Alku* as well as that of *Picart, Drugman and Dutoit* attempting to mimic aspects of the behaviour of talkers, while *Valentini-Botinhao, Yamagishi, King and Maia* explore the generation of speech tuned to the capabilities of listeners. Picart et al. motivate their approach by Lindblom’s “H and H” theory (Lindblom, 1990) which proposes a continuum of degree of articulation from hypo (minimal articulation effort) to hyper (maximum clarity). An ability to synthesise speech from anywhere along this continuum allows a speaking style appropriate to the situation to be selected, such as hyper-articulated speech in the presence of noise. Picart et al. achieved this in a straightforward manner by adapting statistical parametric models using recordings of neutral, hypo and hyper speech, and provide experimental evidence that listeners perceive the degree of articulation correctly and that varying it has the appropriate effect on intelligibility in both noise and reverberation conditions. Raitio et al. also open their paper with a discussion of talker performance, this time in terms of vocal effort, which encompasses voice quality in addition to hypo and hyper articulation. As in Picart et al., statistical models are used to generate speech with low and high effort by adapting models trained on neutral speech. However, in addition to adapting the model parameters, Raitio et al. use a vocoder in which the vocal tract filter is excited by sampled glottal pulses, which can be selected from a library of pulses of the desired vocal effort level. A combination of intelligibility testing and continuous rating scales were used in the listening tests to evaluate the synthetic speech.

Whilst both Picart et al. and Raitio et al. rely on the availability of recorded speech in the desired style, whether that be hyper-articulated or Lombard speech, Valentini-Botinhao et al. propose a technique that does not require any such

recordings, although they also show benefits when it is combined with model adaptation techniques similar to those used by Picart et al. and by Raitio et al. Valentini-Botinhao's technique employs a computational model of audibility – the 'glimpse proportion' (Cooke, 2006). By modifying synthetic speech to maximise the glimpse proportion, audibility – and as a consequence intelligibility – is increased. Valentini-Botinhao et al. find that the type of modification performed needs to be highly constrained, since arbitrary modification of speech introduces excessive distortion. All of the methods proposed for synthetic speech rely on a parametric representation of speech – i.e., a vocoder – and they operate either by manipulating statistical models of this representation or, in Valentini-Botinhao et al's approach, by operating directly on the vocoder parameters. The methods of Picart et al. and Raitio et al. are limited to synthetic speech, since they involve model manipulation, but the method of Valentini-Botinhao et al. could in principle be applied to natural speech. They are all perhaps limited in quality by a requirement to vocode the speech. Picart et al. and Raitio et al. did measure the quality of the modified speech, whereas Valentini-Botinhao et al. only considered intelligibility.

We thank all the reviewers for their sterling work in assessing and improving the submitted articles for this issue.

References

Lombard, E., 1911. Le signe d'élévation de la voix [The sign of the elevation of the voice]. Annales des maladies de l'oreille et du larynx 37, 101–119.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H & H theory. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 403–439.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119, 1562–1573.

Martin Cooke [a,b,*]
[a] *Ikerbasque (Basque Science Foundation), Spain*
[b] *Language and Speech Laboratory, Facultad de Letras, Universidad del País Vasco, Vitoria, Spain*
Simon King
Centre for Speech Technology Research, University of Edinburgh, UK
Bastiaan Kleijn
Victoria University Wellington, New Zealand
Yannis Stylianou [a,b]
[a] *Toshiba, Cambridge, UK*
[b] *Department of Computer Science, University of Crete, Greece*

* Corresponding author.
E-mail addresses: m.cooke@ikerbasque.org (M. Cooke), simon.king@ed.ac.uk (S. King), bastiaan.kleijn@ecs.vuw.ac.nz (B. Kleijn), yannis.stylianou@crl.toshiba.co.uk (Y. Stylianou)