Introduction to the special issue on auditory-visual expressive speech and gesture in humans and machines

Jeesun Kim, Gérard Bailly, Chris Davis

Introduction to the Special Issue on Auditory-visual expressive speech and gesture in humans and machines

Jeesun Kim[1], Gérard Bailly[2], Chris Davis[1]

[1]The MARCS Institute, Western Sydney University, Australia; [2] GIPSA-Lab, Grenoble-Alpes University, France.

Abstract

We speak to express ourselves. Sometimes words can capture what we mean; sometimes we mean more than can be said. This is where our visible gestures - those dynamic oscillations of our gaze, face, head, hand, arms and bodies – help. Not only do these co-verbal visual signals help express our intentions, attitudes and emotion, they also help us engage with our conversational partners to get our message across. Understanding how and when a message is supplemented, shaped and changed by auditory and visual signals is crucial for a science ultimately interesting in the correct interpretation of transmitted meaning. This special issue highlights research articles that explore co-verbal and nonverbal signals, a key topic in speech communication since these are crucial ingredients in the interpretation of meaning. That is, the meaning of speech is calibrated, augmented and even changed by co-verbal/speech behaviours and gestures including the talker's facial expression, eye-contact, gaze-direction, arm movements, hand gestures, body motion and orientation, posture, proximity, physical contact, and so on. Understanding expressive signals is a vital step for developing machines that can properly decipher intention and engage as social agents. The special issue is divided into three parts: Auditory-visual speech perception; Characterization and perception of auditory-visual prosody; Computer-generated auditory-visual speech. Below, we introduce these papers with a brief review of relevant issues and previous studies, when needed.

1 Aim and scope

This special issue aims to provide a platform for consolidating research on human expression and interactive social robotics. Behavioural research seeks to identify the conditioning behaviors that support, structure and maintain social interactions along with auditory and visual cues that signal such things as attitude, emotion and importance. Research on interactive human-agent interaction (e.g., computer vision; machine learning, robot design, etc.) provides the methods, techniques to help quantify communicative signals and to test theories by implementing cues in interactive agents. In summary, the broad aim of the special issue is to connect human and machine research on expressive speech and gesture.

2 Processing auditory-visual expressive speech and gestures: What and why?

What is expressive speech and why is it interesting? A traditional way of classifying the information conveyed by speech is to consider three overlapping categories of information: linguistic, paralinguistic, and non-linguistic (e.g., Fujisaki, 2003). The topic of the special issue largely concerns the paralinguistic information conveyed by expressive speech; although in our view the division between linguistic, paralinguistic, and non-linguistic information is rather porous. In defining the scope of paralinguistic information, some authors (e.g., Crystal, 1974) have excluded visual information; clearly in the current special issue dedicated to auditory and visual expressive speech we do not impose such a limitation. Why are both auditory and visual information important? The key reason why both auditory and visual information are important is simple that the information from both is relevant to capturing the expressive aspects of speech. That is, if the aim is to

understand how intentions, beliefs, desires, and so forth, are signaled, then the relevant information must be in the input.

The articles in this special issue fall into three broad categories: Auditory-visual speech perception, the characterization and perception of auditory-visual prosody; computer generated auditory-visual speech.

2.1 Auditory-visual speech perception

Speech is produced by an elaborate and skilled play of gestures that shape the space in which the hybrid string-and-wind instruments of articulation sound. This process of articulation is visible, and the term "auditory-visual" speech captures the general idea that speech is more than simply sound. Research on auditory-visual (AV) speech perception has tended to cluster around extremes; like the remarkable McGurk effect (McGurk & MacDonald, 1976), or the everyday effect whereby seeing the talker assists speech perception in noise (Sumby & Pollack, 1954). While these two types of demonstration of the auditory and visual nature of speech are important, they tend to obscure other important phenomena and insights. The three papers in this section examine more nuanced aspects of AV speech production/perception.

Only relatively recently has it been demonstrated that seeing the exaggerated visible articulation of speech produced in noise (Lombard speech) enables better perception of auditory speech in noise, when compared to the visible articulation of quite speech (Kim, Sironic & Davis, 2011). The Alghamdi et al's paper (this issue) investigated the impact of artificially exaggerating the visual articulatory features (the mouth kinematics) on the benefit to speech recognition performance of seeing the talker. The study investigated two aspects of this visual exaggeration. First, it examined whether people can adapt to the conflict between the exaggerated visual motion and the unexaggerated acoustic signals. Then it investigated if the application of visual exaggeration could improve auditory and AV speech recognition when used in perception training. Understanding how and why visual speech can assist in learning to recognize speech is crucial for the development of automatic training procedures (e.g., Davis & Kim, 2001).

The Bear and Harvey paper (this issue) proposes that an obstacle to the progress of AV expressive speech research (and AV speech research more generally) is the lack of consensus on the recognition units of visual speech. Although, the idea of abstract recognition units is itself controversial, such units provide a principled way of understanding perceptual confusions and calculating error rates. The Bear and Harvey paper goes back to basics in reviewing different choices of visual recognition units and provides a full set of evaluations of the competing phoneme-to-viseme mappings.

The final paper in this section by Chong, Kim and Davis (this issue) reports on research that extends previous work on how the expression of emotion affects speech production (e.g., Drahota, Costall & Reddy, 2008; Jiang & Pell, 2017). Chong et al examined the emotion of disgust for two reasons. The first is that its expression appears to be firmly established. That is, its expression is tied to a clear selection pressure (pathogen avoidance) and so likely involves stereotyped patterns of action (something consistent with its uniform expression across cultures). The second reason is that its articulation specifically involves the mouth region, and as such should influence speech articulation and consequently resulting in acoustic change. The co-production of emotional and vocal expressions provides a concrete case of how the dynamics of speech production need comply with the articulation of other signals that are displayed on the face.

2.2 Characterization and perception of auditory-visual prosody

Information is exchanged in a conversation as part of a performance. The conversational actors juggle the symbols of information with practiced rhythm; grouping and highlighting ideas. Like an

acrobatic routine, interlocutors need to coordinate and to be aware of when each performance begins and ends. Research into such cues has traditionally focused on auditory properties, particularly those involved in prosody. However, since prosody is also conveyed visually (Cvejic, Kim & Davis, 2010, 2012), more recent work has considered both auditory and visual cues.

What is prosody and why is it important for communication? The term prosody generally refers to speech rhythm, and is typically operationalized by three properties: timing, amplitude and fundamental frequency. Whereas change in fundamental frequency only relates to auditory speech, timing and amplitude apply also to visual prosody. Prosody is important for communication because it helps decipher what the talker may mean or knows. For instance, prosody can provide the listener with clues about which speech elements the talker wants grouped together, and which elements may be more important. Prosody can also convey talker intention and affect.

The papers in this part of the special issue are mostly concerned with the linguistic and paralinguistic functions of prosody, rather than the affective function. The linguistic and paralinguistic aspects of prosody are voluntarily produced and part of an interlocutor's communicative goal. Understanding how speech conveys a talker's message by signaling his/her interest, intention and attitude has practical implications for human-machine communication. That is, although automatic speech recognition is reaching human-level performance at mapping speech onto linguistic units (word recognition, see Xiong et al, 2017), a coming challenge is for systems to go beyond the information specified by logical form of an utterance (the words) in order to get a better handle on what people intend and feel about what they are saying. In addition, effective auditory-visual speech synthesis will require knowledge of how to express intention and attitude.

The four papers investigated how the properties of auditory-visual prosody combine and interact (Ambrazaitis & House, this issue), how prosody is used to signal the end of a conversational turn (Bi & Swerts, this issue), and how attitudes can be expressed (Barbulescu, Ronfard & Bailly, this issue; Mixdorff et al, this issue).

We know that people are sensitive to the relationship between head/eyebrow motion and auditory prosody (Cvejic et al, 2010, 2012). Yet, how, when we make some part of speech prominent, do movements of the head and eyebrows actually relate to the properties of auditory speech? The answer to this question has implications for a range of areas, e.g., auditory-visual speech synthesis, automatic detection of prominence (Heckmann, 2018), and so on. The paper by Ambrazaitis and House (this issue) investigated how multimodal cues to prominence are coordinated. The interest was in examining the relationship between three verbal/visual prosodic markers of prominence: Focal pitch accent; head beats and eyebrow beats. The study examined whether these three markers combine in specific ways. Although this question has been examined in experimental settings (e.g., Kim, Cvejic & Davis, 2014), here, Ambrazaitis and House used something akin to a 'case study' approach. That is, they examined talker productions in a domain where the clear use of prosody is paramount, i.e., the productions of television news anchors.

The research by Bi and Swerts (this issue) builds on previous work that has identified pitch as an important auditory cue, and mutual gaze and eyebrow and head movements as important visual ones for turn taking. Much of this work has been conducted with non-tone languages. For example, Mixdorff et al (2015) looked at AV cues in English as perceived by German and English participants. The question that Bi and Swerts address is whether AV cues to prosodic boundaries differ in a tone versus a non-tone language (Chinese and English). There are several reasons why such a comparison is interesting in terms of how auditory and visual information is employed. First, in non-tone languages, pitch can be freely used to mark discourse information. However, in tone languages the use of pitch to indicate lexical information may constrain its use as a discourse marker. Second, due

to cultural differences, the way that eye gaze is employed in English to signal the end of an utterance may be different than in Chinese. Cross language/cultural contrasts are important, as such can identify properties that are unique to language/culture and properties that are common.

The two remaining papers in this section (i.e., Barbulescue et al and Mixdorff et al) examined how attitudes can be expressed by auditory and visual prosody. Before describing these topics in more detail, it is worth making a brief comment on the diversity of terminology used to describe the expression of attitudes and to touch on the broader context of this work.

Many different terms have been used to describe the expression of attitudes and intention. For example, the terms “social attitudes”, “prosodic attitudes” and “communicative intentions” as well as “dramatic attitudes”, “social affects”, have all been used (e.g., Barbulescu et al, this issue; Helbernd & Sammler, 2016; Kim & Davis, 2016; Mixdorff et al, this issue; Wichmann, 2000). The diversity of terms likely reflects different perspectives on the underlyingly nature of this type of expressive speech. To provide an insight into this, what follows is a brief description of how researchers have categorized different expressions of attitudes and intention.

Research on expressive speech generally tends to consider linguistic prosody “the organizational structure of speech" (Beckman 1996), separately from paralinguistic prosody that concerns the expression of attitudes and emotions. In addition, some researchers have proposed that the latter two categories (attitudes and emotions) should also be considered separately (Ohala, 1996; Moraes et al, 2010; 2011). For example, Moraes et al argued for a distinction between ‘social affects’ and emotional expression based on only the former being under voluntary control. Ohala made a similar argument, and added that emotional prosody is grounded in adaptive processes, where either the transmission of a signal has survival value or where a signal ‘leaks’ from a beneficial physiological state. The expression of attitudes, he argued, do not seem to confer any obvious survival benefit to the signaller and are probably acquired, i.e., learned (see also Wichmann, 2000, who argues that attitudes are best thought of as pragmatic inference that is reliant on context and situational knowledge).

This distinction based on putative differences in the underlying etiology of emotions and attitudes may be important for how well-established and characteristic the signaling of such can be. That is, Ohala (1996) suggested that emotional expressions are likely to be found cross-culturally, whereas the expression of attitudes are likely to vary from culture to culture. Moreover, he proposed that for attitudes to be appropriately communicated, they would need to be contextualized. In this regard, the paper by Mixdorff et al (this issue) examined how prosodic attitudes are perceived across culture, and the paper by Barbulescu et al (this issue) examined the ‘grain-size’ (frame-level, syllable-level, sentence-level) over which auditory-visual prosodic features are most effective in discriminating among expressive styles.

Moraes et al (2010; 2011) have proposed an additional division between propositional attitudes (those expressions involved in proposition content, e.g., irony, incredulity, etc.) and social attitudes (those expressions that references interpersonal relations, e.g., politeness, arrogance, etc.). Moraes et al argue that there is a difference in the degree that auditory and visual information is distinctive in these different types of attitudes. For the expression of propositional attitudes (more tied with linguistic meaning) they suggest that auditory expressions play a more important role; whereas for the social attitudes, they suggest that visual expression is important. This proposal points to the importance of determining variation (both across individuals and across prosodic types) in the production and perception and of these expressions, and to determine the extent to which context is required for unambiguous interpretation (see Hellbernd & Sammler, 2016; Kim & Davis, 2016).

Bearing the above in mind, Mixdorff et al (this issue) looked at a range of different types of expression and different ways of evaluating how well each can be decoded with use of a dimensional concept developed in research on affect to classify expressions of attitudes. Importantly these expressions were divided into propositional, social and mixed propositional/social types and auditory, visual and auditory-visual formats.

Barbulescu et al (this issue) emphasized the need for the collection of databases that consist of face/head motion and acoustic data. Collecting visual data is important for quantifying effects as well as for developing applications where the animation of virtual characters that can express complex mental states using expression that coordinate vocal prosody, head motion and facial expressions and head and gaze motion (see Barbulescu et al, 2017). The Barbulescu et al paper (this issue) makes an important contribution in identifying conditions illustrating talker dependent strategies and also attitude-specific ones.

2.3 Computer-generated auditory-visual speech

This part of the special issues highlights the development and challenges of synthesizing auditory-visual speech. We begin with a brief consideration of the need to go beyond the auditory only modality.

There has been a tremendous growth in voice-based AI interfaces, with the major technology companies making significant investments in this area. This is just the beginning. Smart voice assistants such as Amazon Echo or Google Home enable useful and successful spoken interaction with human users, particularly in a simple service-information-delivery role (home automation/entertainment). However, for tasks that work better with more engaging personalization, richer feedback and more natural conversation, embodied conversational agents (ECA, see Cassell, 2000) that provide co-verbal and non-verbal channels will be developed. As mentioned above, the movements of the body (face, head, hands, postures) of a virtual agent or a social robot provide additional channels of information that enrich the interaction with redundant and complementary information. That is, most speech gestures are clearly visible and are complemented with co-verbal gestures such as gaze, head and hand movements. This visual information not only helps the speaker to plan his/her discourse but it also helps the perceiver to interpret the discourse by providing non-audible cues about the speaker's emotional, psychological or physical state. As an example, the speaker's gaze direction and facial displays interact in the perception of emotional content and personal involvement (Sander et al, 2007).

While more and more research is dedicated to the generation of interactive behaviors for full body avatars including navigation of robots in human crowds, providing the face of embodied conversational agents with audiovisual expressivity is still an open challenge. A common problem addressed in the three papers of this part of the special issue is the domain/units over which audiovisual synchronization is performed.

One obvious solution will be to consider visual displays as add-ins to the prior generation of the speech signal. The literature is paved with direct speech-to-gesture mapping models trained on parallel audiovisual corpora (see as example Ding, Xie & Zhu, 2015 and Sakai et al, 2015 for head motion prediction or Karras et al, 2017, for speech-driven generation of facial movements). These works presuppose that the mapping models will unveil implicit latent multimodal representations that leave sufficient traces in the acoustic signal to influence co-verbal behaviors (see for example the speech to articulation mapping via GMM performed by Toda, Black & Tokuda, 2008 or Hueber et al, 2015). Interestingly, Sadoughi, Liu and Busso (this issue) demonstrate that such mappings (in their case speech-to-head mapping) strongly benefit from the explicit specification of the communication intent (in their case, the dialog act). Indeed, they propose to map synthetized neutral vs. emotional

speech computed by OpenMary with natural emotional head movements available in the motion capture database IEMOCAP (Busso et al, 2008). Given each utterance performed in the IEMOCAP together with its average arousal, valence and dominance, OpenMary generates a synthetic speech signal which is back-aligned with the original IEMOCAP acoustic signal. The parallel corpus of natural head motions and warped synthetic signals is then used to create predictive models of head motion driven by synthetic signals. They compare baseline Dynamic Bayesian Networks (DBN) mapping prosodic features (F0 & intensity) directly to head rotations and "constrained" DBN (CDBN) with nodes informed by the underlying 5-valued dialog acts. Ground-truth vs. generated head motions together with phonetic transcriptions drive the animation of a conversational agent. Subjective tests show that CDBN convincingly preserves the first-order statistics of the influence of head motions by dialog acts.

Another option is to coordinate the multimodal behavior via modal units. Here, units of the predominant modality are often chosen. Most interactive behavioral models (see the influential SAIBA model in Cafaro et al, 2014) use the structure of the acoustic signal as a multimodal skeleton: co-verbal trajectories or events are coordinated in relation to acoustic onsets of phonological units (phonemes, words, prosodic words, etc). Chuang and Bregler (2005) also generated facial expressions from speech via melodic units. Interestingly, Cakmak and Dutoit (this issue) show that multimodal generation may benefit from the organization of modal streams into cue-specific units. They exploit an original database AVLASYN (Cakmak et al, 2014) comprising 48 minutes of laughter performed by one male subject when watching funny movies. They propose a model for visual laughter generation composed of 3 parts (facial, head vs. eyelids) whose movements are segmented into specific units. Facial and head units are modelled as Hidden Markov Models (HMM). Eyelids movements are paced by blinks generated from the empirical probability density function of blinking frequency and blink duration. They show that one avatar driven by this model is subjectively rated as convincing as when driven by ground-truth movements. Of course, this impressive result should be interpreted in light of the fact that the synthetic movements are computed from ground-truth units (labels and timing) and that both ground-truth and synthetic movements are synchronized with natural laughter sound tracks. Note that this visual laughter renderer is driven by a few distinctive units: 3 for the facial part (with strong syntactic constraints) and 1 for the head oscillation. This would certainly have aided the development of the synchronization module to couple with their acoustic renderer (Urbain et al, 2014).

The benefit of freeing the generation model from common underlying representations is also illustrated by Filntisis et al (this issue). They compare the quality of several audiovisual text-to-speech mapping models (HMM, DNN, unit selection) trained from a unique set of 900 sentences uttered by one female Greek actress with 4 different emotions (neutral, anger, joy vs. sadness). The authors show that DNN with separate modelling of acoustic and visual features achieves significantly greater mean opinion scores in comparison with other architectures: forcing the audiovisual generator not only to share the input acoustic-contingent representation (e.g. positional features such as the current audiovisual frame with the acoustic states of the phone) but also common latent representations by imposing the prediction of joint audiovisual output features, i.e., does degrade generation performance.

Mapping models and multimodal generators should enable the simultaneous training of cue-specific dynamical systems monitoring the activity of each modality at its own pace, because the natural frequency, the damping parameters, the scope of the various communicative functions encoded by each cue are different. At the same time, these mapping models and multimodal generators should be able to couple these dynamical systems at key synchronization moments, e.g., coordination of head and eye gaze, speech and pointing gestures are tightly coupled in deixis for attracting attention and engaging into cognitive or physical action. Long short-term memory (LSTM) units have this

capacity to pace the activity of recurrent networks in order to reproduce a large variety of dynamical behaviors (see Nguyen, Bailly & Elisei, 2017). After their success in multimodal recognition, we expect such RNN models to shortly invade multimodal generation (see Rajagopalan et al, 2016).

3 Future directions

The current issue focused on facial movements related to speech and emotion. We are aware of course that the whole body "speaks": we expect future studies to address the processing of co-verbal gestures and body postures, i.e., multimodal characterization of expressive whole-body behaviors.

What also needs to be considered is the socio-communicative intentions of the speaker. This is because intentions have a strong impact on his/her multimodal behavior. These behaviors are also strongly influenced by the listener's overt behaviors and estimated socio-communicative intentions. We expect speech communication research will increasingly characterize and model multimodal behaviors in interaction, i.e., those involving the ambient environment including the interlocutors' actions and reactions to intended socio-communicative behaviors. Such phenomena as coadaptation, synchrony, and alignment mechanisms should be taken into account to consolidate research on human expression and interactive social agents.

In addition, the impact of time-varying cognitive, emotional and social settings should be studied and incorporated into multimodal interactive models that can adapt dynamically to human partners. This is because we behave differently as an interaction proceeds, i.e., models of interactive behaviors shaped by the initial situation and potential bias are updated and tuned to reflect an increasing awareness of the goals and social benefits of the interaction.

At a practical level, massive multimodal data collected in natural environments and ecological situations will be essential for gaining insights into how to enable communication between human and artificial agents seamless. As outlined in this special issue, modern machine learning should not be considered as a threat by the speech communication community but as a powerful statistical tool to utilize data and generate insights into human cognition.

Finally, more work is needed in understanding the fit between different types of information/interactivity and different use-case scenarios. What is needed is a twinning between the way we feel, think, speak and act; and the type/level of engagement with an automatic task-oriented system. For some tasks, systems that simply act as extended appendages are best (i.e., you do not need to have a conversation with your hand to turn the lights on – for such tasks automatic systems should be simple effectors); for some tasks, such as offering advice, engagement and companionship, a full range of interaction abilities is required; here both auditory and visual expressive speech is a must.

4 Acknowledgments

We thank the chief editor of Speech Communication, Bernd Möbius, for enabling this special issue to become a reality and for his strong commitment to this journal. Special thanks go to all authors that submitted articles to this special issue, and to the anonymous reviewers for their support in maintaining the high standard of this journal. The first & third authors acknowledge support from an ARC Discovery grant, DP150104600.

5 References

Alghamdi, N., Maddock, S, Barker, J., & Brown, G. J. (this issue). The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech, *Speech Communication*, 95, 127-136.

Ambrazaitis, G., & House, D. (this issue). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication*, 95, 100-113.

Barbulescu, A., Ronfard, R., & Bailly, G. (2017). A Generative Audio-Visual Prosodic Model for Virtual Actors. *IEEE computer graphics and applications*, 37(6), 40-51.

Barbulescu, A., Ronfard, R., & Bailly, G. (this issue). Which prosodic features contribute to the recognition of dramatic attitudes? *Speech Communication*, 95, 78-86.

Bear, H., & Harvey, R. (this issue). Phoneme-to-viseme mappings: the good, the bad, and the ugly, *Speech Communication*, 95, 40-67.

Beckman, M. E. (1996). The parsing of prosody. *Language and cognitive processes*, 11(1-2), 17-68.

Bi, R., & Swerts, M. (this issue). A perceptual study of how rapidly and accurately audiovisual cues to utterance-final boundaries can be interpreted in Chinese and English. *Speech Communication*, 95, 68-77.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., & Pelachaud, C. (2014, August). Representing communicative functions in SAIBA with a unified function markup language. In *International Conference on Intelligent Virtual Agents* (pp. 81-94). Springer, Cham.

Çakmak, H., & Dutoit, T. (this issue). HMM-based generation of laughter facial expression. *Speech Communication*.

Cakmak, H., Urbain, J., Dutoit, T., & Tilmanne, J. (2014). The AV-LASYN Database: A synchronous corpus of audio and 3D facial marker data for audio-visual laughter synthesis. In *LREC* (pp. 3398-3403).

Cassell, J. (Ed.). (2000). *Embodied conversational agents*. MIT press.

Chong, C., Kim, J., & Davis, C. (this issue). Disgust expressive speech: the acoustic consequences of facial expressions of emotions. *Speech Communication*.

Chuang, E., & Bregler, C. (2005). Mood swings: expressive speech animation. *ACM Transactions on Graphics (TOG)*, *24*(2), 331-347.

Crystal, D. (1974). Paralanguage. In *Current Trends in Linguistics*, *12, 265-295*.

Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, *52*(6), 555-564.

Cvejic, E., Kim, J., & Davis, C. (2012). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition*, 122(3), 442-453.

Davis, C., & Kim, J. (2001). Repeating and remembering foreign language words: Implications for language teaching systems. *Artificial Intelligence Review*, 16(1), 37-47.

Ding, C., Xie, L., & Zhu, P. (2015). Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, *74*(22), 9871-9888.

Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication*, 50(4), 278-287.

Fujisaki, H. (2003). Prosody, information and modelling with emphasis on tonal features of speech. In *Proc. Workshop SLP* (pp. 5-14).

Heckmann, M. (2018). Audio-visual word prominence detection from clean and noisy speech. *Computer Speech & Language*, 48, 15-30.

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88, 70-86.

Hueber, T., Girin, L., Alameda-Pineda, X., & Bailly, G. (2015). Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12), 2246-2259.

Karras, T., Aila, T., Laine, S., Herva, A., & Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics* (TOG), 36(4), 94.

Kim, J., Sironic, A., & Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception*, 40(7), 853-862.

Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, 57, 317-330.

Kim, J., & Davis, C. (2016). The Consistency and Stability of Acoustic and Visual Cues for Different Prosodic Attitudes. In *INTERSPEECH* (pp. 57-61).

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, 88, 106-126.

McGurk, H., & MacDonald J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-8.

Mixdorff, H., Hönemann, A., Kim, J., & Davis, C. (2015). Anticipation of turn-switching in auditory-visual dialogs. *Proceedings of FAAVSP 2015*.

Mixdorff, H., Hönemann, A., Rilliard, A., Lee, T., & Ma, M. K. (this issue). Audio-visual expressions of attitude: How many different attitudes can perceivers decode? *Speech Communication*, *95*, 114-126.

Moraes, J. A. D., Rilliard, A., Mota, B. A. D. O., & Shochi, T. (2010). Multimodal perception and production of attitudinal meaning in Brazilian Portuguese. In *Speech Prosody 2010-Fifth International Conference*.

Moraes, J. A., Rilliard, A., Erickson, D., & Shochi, T. (2011). Perception of attitudinal meaning in interrogative sentences of Brazilian Portuguese. In *Proc. of ICPhS*.

Filntisis, P. P., Katsamanis, A., Tsiakoulis, P., Maragos, P. (this issue). Video-realistic expressive audio-visual speech synthesis for the Greek language. *Speech Communication*, *95*, 137-152.

Nguyen, D. C., Bailly, G., & Elisei, F. (2017). Learning off-line vs. on-line models of interactive multimodal behaviors with recurrent neural networks. *Pattern Recognition Letters*, *100*, 29-36.

Sadoughi, N., Liu, Y., & Busso, C. (this issue). Meaningful head movements driven by emotional synthetic speech, Speech Communication, 95, 87-99.

Sakai, K., Ishi, C. T., Minato, T., & Ishiguro, H. (2015, August). Online speech-driven head motion generating system and evaluation on a tele-operated robot. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on* (pp. 529-534). IEEE.

Sander, D., Grandjean, D., Kaiser, S., Wehrle, T., & Scherer, K. R. (2007). Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. *European Journal of Cognitive Psychology*, *19*(3), 470-480.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.

Toda, T., Black, A. W., & Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, *50*(3), 215-227.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... & Zweig, G. (2017, March). The Microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP) 2017 IEEE International Conference on* (pp. 5255-5259). IEEE.