AUDIOVISUAL SPEECH: ANALYSIS, SYNTHESIS, PERCEPTION, AND RECOGNITION

Sascha Fagel

Berlin University of Technology
sascha.fagel@tu-berlin.de

ABSTRACT

In many cases research in the fields of audiovisual speech analysis, synthesis, perception and (automatic) recognition is carried out separately with only limited account for the neighboring areas. But the author claims that these neighboring areas yield huge, currently idle potential to improve and better understand the field under investigation and that human speech as a phenomenon should be looked at from a more holistic point of view. This paper briefly looks into the fields of audiovisual speech research and tries to identify existing links between them as well as future collaboration for promising prospective mutual benefit.

Keywords: speech production, speech perception, speech synthesis, automatic speech recognition.

1. INTRODUCTION

Speech is often seen as audio communication between humans my means of words. But it is not only what is said that is important, but also how it is said, i.e. information in speech is not only transmitted by the linguistic content but also by additional prosodic cues. A close look reveals speech production as a physiological process that becomes audible and visible where the auditory and visual sensory channel provide complementary information: some physiological properties (e.g. lip rounding) can be heard and seen, others are only audible (e.g. hoarseness), and others still are only visible (e.g. frown) – mostly a mixture of different auditory and visual information occurs. The senses yield different views to the same phenomenon. Hence, a holistic view includes both.

Furthermore, in a communication situation the speaker usually acts also as a perceiver of his or her own speech. In the latter case vision is rarely informative but audition, proprioception, the sense of touch from the articulators and the articulatory motor control commands are in principle available for perceiving one's own speech. So a broader look at the phenomenon of speech involves the speech production process at the speaker, the speech perception process at the listener, and – nonetheless – the feedback of one's own speech especially during speech acquisition (which goes beyond the main scope of this paper).

Although much is known about the speech production process, audio speech synthesis is commonly not carried out by implementing this knowledge. Analogously current automatic speech recognizers are not built by "reverse engineering" the human perceptual system. To some extent this is also the case in the field of visual speech synthesis and recognition. Knowledge of the speech production and perception processes is already being used but still hold potentials for improvement of automatic systems.

The investigation of speech production is mainly regarded in this paper as analysis of observations of the speech production process by means of acoustical, optical and other (e.g. articulatory) measurements. Current and possible links between speech production/analysis, perception, synthesis, and recognition are discussed in this paper. These are not exhaustive as each individual section is capable of filling a book.

2. HUMAN MACHINE SPEECH COMMUNICATION

Figure 1 shows the communication chain from the communication intention to the linguistic and paralinguistic information (excluding the internal feedback path where the speaker acts simultaneously as perceiver). The speech generation process is either natural or synthetic and analogously the decoding is done either by human perception or by automatic recognition. The interface between speech generation and registration instances is the transmission of the acoustical, optical and physiological (in case of articulation measurements) manifestations.

3. PRODUCTION AND PERCEPTION

Speech is produced to be perceived. The human speech organs and the sensory system have evolved jointly and hence can be assumed to be well adjusted to one another. Furthermore a human has access to his or her own motor control "data" and his or her own sensory system while speaking and thus have a broad feedback of what is produced – even though a) the acoustic feedback has different characteristics

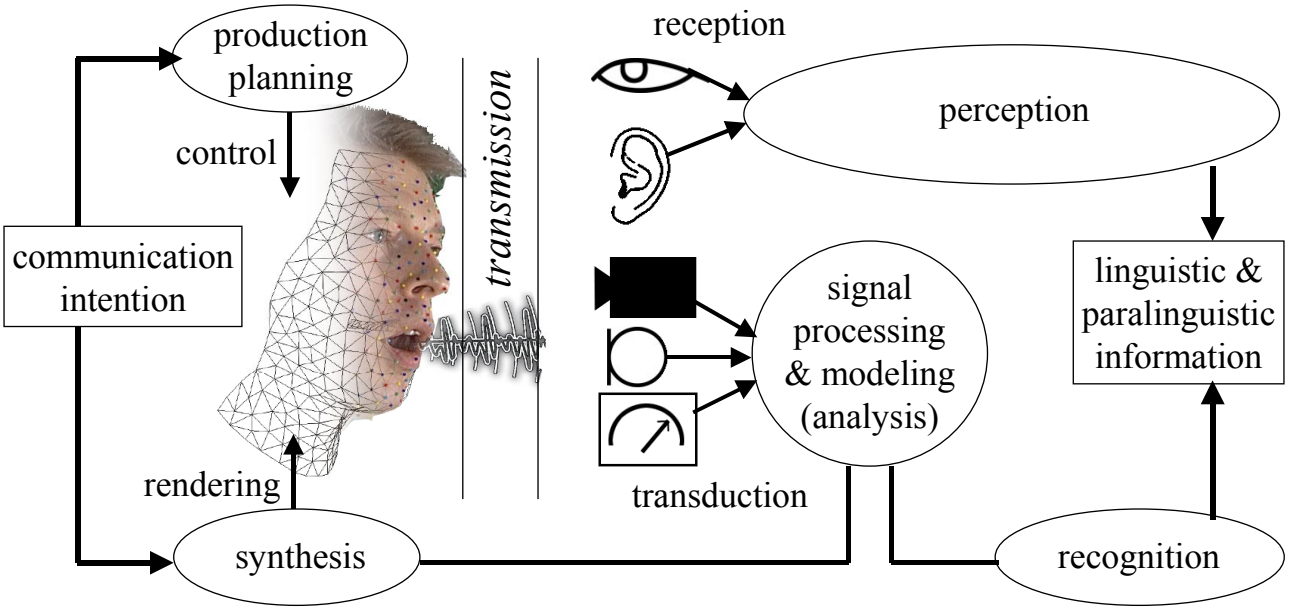

Figure 2: *Speech communication chain.*

than the signal that reaches the ear when another person is speaking and b) the optical feedback is only marginal. (The strongest link between speech production and perception is stated in the Motor Theory [17]; see a brief discussion in section 6.) Insights into both the production and the perception processes can be yielded by the view of speech production and perception as a partly self-organizing system. [5] see the potential benefits of understanding such an architecture mainly in terms of speech perception such as coarticulation, analysis-by-synthesis, motor theory, categorical perception, invariant speech perception, word superiority, and phonemic restoration. An interesting initial implementation of this idea in the form of an artificial speech acquisition and synthesis system can be found in [12].

The problem of pronunciation adaptation (section 4) is also present in human speech recognition – e.g. when the speaker is using a language that is not native to him or her [29]. But human listeners are highly capable of adapting to the speaker's specific pronunciation in case of accented or dialect speech. It is unknown whether this is carried out by switching between different models, as in multiple baseform ASR systems [31], or by an adjustment of a generic recognition scheme. But, while the perceiver adapts to the speaker specifics, these characteristics and linguistic information of the utterance are available and used by the listener at the same time.

4. PRODUCTION AND AUTOMATIC RECOGNITION

Solving the problem of variation in human speech production is one major issue of ASR (automatic speech recognition) systems as the same utterance is never spoken in exactly the same way more than once. The pronunciation adaptation implemented in current ASR systems on the one hand enhances the system performance [28], on the other hand it can be used to measure the similarity between a speakers pronunciation and that in the training data [8]. There are still many characteristics of human speech production that are not taken into account in most of the current ASR systems, e.g. that the movements of the articulators are shifted in phase to one another [17], that the speech sounds and their properties do not occur sequentially [10], and that articulation changes with background noise [20] and speaking rate [24].

5. PERCEPTION AND AUTOMATIC RECOGNITION

Many properties of the perceptual system are still not implemented in current ASR systems. The time window of most current ASR systems is quite short compared to that of human perception. Whereas knowledge of the speech perception process is intuitively valuable for ASR development (e.g. the non-uniform frequency resolution of the auditory system), some ASR researchers argue that results from ASR research also yield insights into human speech perception. For example the promising outcome of temporal pattern based ASR compared to conventional spectral envelop-based ASR [11].

The visual modality of speech is already included in many experimental ASR systems (see e.g. [21]). The main goal is to reduce recognition errors although it is also known that in human speech

perception the visual modality speeds up the neural processing of auditory speech [32]. Another known effect implemented in recent ASR systems [18] is that humans can better detect the presence of speech in noise if the according lip movements are visible [14]. New valuable insights into the problem of speaker variation have been published [25]: Speaker characteristics are represented in the human perception system in a non-unimodal way. It was shown that the auditory intelligibility increases when the listener is presented with silent lip movements of the same speaker before listening to his or her audible utterance. Hence, humans can estimate the characteristics of the voice by seeing the lip movements. Such a human speaker adaptation yields potential for bimodal speaker adaptation in automatic systems.

6. ANALYSIS, SYNTHESIS AND PERCEPTION

The theories that have been developed at the intersection of speech production and speech perception – such as Acoustic Invariance Theory, Adaptive Variability Theory, Motor Theory and Direct-Realist Theory (find a comparative description in [22]) – are functional rather than computational in nature and hence cannot yet feasibly be applied directly to speech synthesis and automatic speech recognition. Models that can be used to generate synthetic speech deal with the physical manifestation of speech; therefore they do not actually incorporate knowledge of speech production, but rather describe the acoustic or visual representation in data-driven audio (e.g. [4]) or video (e.g. [2]) synthesis systems or the physical properties that are then rendered in articulatory synthesizers (e.g. [1]) or "Talking Heads" (e.g. [7]). Rather than define the accuracy of the model in terms of variance explanation or error measures it is common to evaluate the derived systems subjectively using human listeners. The perceptual relevance of variance explanation or error measures in re-synthesized speech is to date far less sophisticated compared to technical speech quality assessment (PESQ [13]).

7. SYNTHESIS AND RECOGNITION

Techniques for speech synthesis commonly have little overlap with those applied in automatic speech recognition. Furthermore, both are concerned with one machine-sided aspect of human-computer-interaction (human perception of synthesized speech or automatic recognition of human utterances, respectively). But there are examples where both research areas can benefit from one another. One application is to use speech recognition features to re-synthesize speech. [15] showed that speech reconstruction from only MFCC features results in reasonably intelligible speech, where the pitch and the phase additionally have to be reconstructed to achieve good speech quality in terms of naturalness. The use of speech synthesis for training and testing ASR systems can be extended to automate and hence facilitate assessment of both the speech synthesis and the ASR system that are involved.

8. AUDIO-VIDEO CORRELATION

As visible and audible speech derives from a single physiological process, correlations between the acoustic and optical representations can be found [33] (and [9] who found phoneme-specific correlations between linear combinations of visual features and linear combinations of audio features). Nevertheless the optical and the acoustical manifestation yield somewhat complementary information. Both sources of information in conjunction lead to better intelligibility [30] and to higher recognition rates [23] than the use of only one of the channels. This points to partial mutual exclusive information in the two channels and hence to a complementary nature of audible and visible speech with respect to linguistic features. To extract information from the sensory channels, audition and vision are processed jointly in human perception. Speech information present in these channels is known to be integrated at a very early stage of perception [27] and the auditory and visual cortices are not purely sensory-specific [3]. The joint usage of audition and vision in human perception leads to a synergy effect where information of one channel becomes accessible by the use of the other channel. This was evidenced on the one hand by [25] who could show that a highly manipulated (sinewave) speech signal – that was non-informative when presented alone – was able to increase visual intelligibility, and on the other hand by [27] who showed the opposite case that a non-informative speech video (when played alone) was able to increase auditory intelligibility.

The complementary nature of audiovisual speech does not apply only to linguistic information. Some paralinguistic information, like frowning, is available only in the visual domain, whereas, e.g. hoarseness of the voice is only audible. Recent research has proven that many prosodic features like word prominence [16] or the level of speaker confidence [6] are spread over the audio and the video channel.

9. SUMMARY

This paper attempts to show recent joint work and possible mutual benefit of the different scientific areas of speech research and argues that speech communication science is interdisciplinary in nature. Engineers in the field of speech technology often have profited from results originating from humanities and vice versa. But still there are many things to learn from one another.

10. REFERENCES

[1] Birkholz, P., Jackèl, D. 2003. A Three-Dimensional Model of the Vocal Tract for Speech Synthesis. *Proceedings of the 15th ICPhS*, Barcelona, 2597-2600,

[2] Brand, M. 1999. Voice Puppetry. *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, Los Angeles, 21-28.

[3] Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S. 1997. Activation of Auditory Cortex During Silent Lipreading. *Science* 276, 593-596.

[4] Carlson, R. 2002. Data-Driven Formant Synthesis. *Proceedings of Fonetik* 44(1), 121-124.

[5] Cohen, M.A., Grossberg, S., Stork, D.G. 1988. Speech Perception and Production by a Self-Organizing Neural Network. Y.C. Lee (Ed.): *Evolution, Learning, Cognition, and Advanced Architectures*, Hong Kong, 217-231.

[6] Dijkstra, C., Krahmer, E., Swerts, M. 2006. Manipulating Uncertainty: The Contribution of Different Audiovisual Prosodic Cues to the Perception of Confidence. *Proceedings of the Speech Prosody Conference*, Dresden.

[7] Fagel, S., Bailly, G., Elisei, F. 2007 (in print). Speaker Cloning in 3D: Intelligibility of Natural and Synthetic Visual German Speech. *Proceedings of the Workshop on Auditory-Visual Speech Processing*, Hilvarenbeek.

[8] Franco, H., Neumeyer, L., Digalakis, V., Ronen, O. 2000. Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication* 30, 121-130.

[9] Goecke, R., Millar, J.B. 2003. Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. *Proceedings of the Workshop on Auditory-Visual Speech Processing,* St. Jorioz, 133-138.

[10] Goldsmith, J. 1990. *Autosegmental and Metrical Phonology*. Blackwell.

[11] Hermansky, H. 2001. Human Speech Perception: Some Lessons from Automatic Speech Recognition. *Proceedings of the International Conference on Text, Speech and Dialogue*, Zelezna Ruda 187-196.

[12] Howard, I.S., Huckvale, M.A. 2005. Learning to Control an Articulator Synthesizer by Imitating Natural Speech. *ZAS Papers in Linguistics*, 63-78.

[13] ITU 2001. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. *ITU-Recommendation P.862*, http://www.itu.int/rec/T-REC-P.862/en.

[14] Kim, J., Davis, C. 2003. Testing the Cuing Hypothesis for the AV Speech Detection Advantage, *Proceedings of the Workshop on Auditory-Visual Speech Processing,* 9-12.

[15] Kleinwächter, T. 2006. Re-Synthesis of Speech from ASR Features. *Masters Thesis at the Royal Institute of Technology*, Stockholm.

[16] Krahmer, E., Swerts, M. 2006. Hearing and Seeing Beats: The Influence of Visual Beats on the Production and Perception of Prominence. *Proceedings of the Speech Prosody Conference*, Dresden.

[17] Liberman, A.M., Mattingly I.G. 1985. The Motor Theory of Speech Perception Revised. *Cognition* 21, 1-36.

[18] Liu Peng, Wang Zuoying 2006. Audio-Visual Voice Activity Detection. *Frontiers of Electrical and Electronic Engineering in China* 1(4), 425-430.

[19] Löfqvist, A. 2004. Speech Motor Control – Laryngeal Function in Speech, *The Japan Journal of Logopedics and Phoniatrics* 45(4), 290-291.

[20] 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- 199.

[21] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J. 2000. Audio-Visual Speech Recognition. *CSLP* W*orkshop Report, The Johns Hopkins University*, Baltimore.

[22] Perrier, P. 2005. Control and Representations in Speech Production. *ZAS Papers in Linguistics*, 109-132

[23] Petajan, E.D. 1984. Automatic Lipreading to Enhance Speech Recognition. *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, 26-29.

[24] Pfitzinger, H.R. 2001. Phonetische Analyse der Sprechgeschwindigkeit. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)* 38, 117–264.

[25] Rosenblum, L.D., Miller, R.M., Sanchez, K. 2007. Lip-Read Me Now, Hear Me Better Later: Cross-Modal Transfer of Talker-Familiarity Effects. *Psychological Science* 18(5).

[26] Saldaña, H., Pisoni, D. 1996. Audio-Visual Speech Perception Without Speech Cues. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, 2187-2190.

[27] Schwartz, J.-L., Berthommier, F., Savariaux, C. 2002. Audio-Visual Scene Analysis: Evidence for a "Very-Early" Integration Process in Audio-Visual Speech Perception. *Proceedings of the International Conference on Spoken Language Processing*, Denver, 1937-1940.

[28] Strik, H. 2001. Pronunciation Adaptation at the Lexical Level. *Proceedings of the ITRW on Adaptation Methods For Speech Recognition*, Sophia-Antipolis, 123-131.

[29] Strik, H. 2003. Speech is Like a Box of Chocolates... *Proceedings of 15th ICPhS*, Barcelona, 227-230.

[30] Sumby, W., Pollack, I. 1954. Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America* 26, 212-215.

[31] van Compernolle, D., Smolders, J., Jaspers, P., Hellemans, T. 1991, Speaker Clustering for Dialectic Robustness in Speaker Independent Recognition. *Proceedings of EUROSPEECH*, Genova, 723-726.

[32] van Wassenhove, V., Grant, K.W., Poeppel, D. 2005. Visual Speech Speeds up the Neural Processing of Auditory Speech. *Proceedings of the National Academy of Sciences (PNAS)* 102(4), 1181-1186.

[33] Yehia, H., Rubin, P., Vatikiotis-Bateson, E. 1998. Quantitative Association of Vocal-Tract and Facial Behavior. *Speech Communication* 26(1-2), 23-43.