Some notes on Audio-Visual Speech Perception from an AVSR perspective

(---- DRAFT ---)

Vitor M M C Pera

FEUP - Porto
December 2014 (?)

Abstract

This brief survey of the ...

Contents

Abbreviations

ASR	Automatic Speech Recognition
AVHSP	Audio Visual Human Speech Perception
AVSP	Audio Visual Speech Perception
AVSR	Audio Visual Speech Recognition
ERP	Event-Related Potential
fMRI	functional Magnetic Ressonance Imaging
HSP	Human Speech Perception
MEG	Magneto-Encephalographic (field recording)
PET	Positron Emission Tomography

...

1 Introduction

In the Audio Visual Speech Recognition (AVSR) community, at least, it is well known the importance of the visual clues to improve the performance of speech recognizers in many useful applications. Empirical results have systematically demonstrated that, such as expected, when the acoustics degrades and the visual stream conveys enough discriminative information then the AVSR approach generally becomes very effective. So, the interest in this R&D area has been increasing and many publications have been produced. By the other side, catching a glance over some of those publications, apparently the AVSR community has not been paying enough attention to related issues in Audio Visual (Human) Speech Perception (AVSP), an area that has been very active along more than half a century. Indeed, usually the papers on AVSR do not make any explicit reference to concepts or results from the AVSP field, or else just refer shortly the McGurck effect.

Is AVSR disregarding AVSP? If so, it shouldn't ...

Naturally, gaining some insight in AVSP can be important in the perspective of the AVSR endeavour. One of the reasons for this interest is that likely the knowledge acquired on AVSP allows to understand better and more deeply the achievements or the failures and limitations of some AVSR approaches designed for useful applications. For instance, if some particular AVSR system is performing worse than expected, eventually the AVSP knowledge may help to mislead a diagnostic based on the natural inter modal time misalignment, or else based on the excessive simplicity of the parametric model for the mouth region. By the way, some common explanations of relevant AVSR phenomena and mechanisms are too much simplistic. Moreover, many reported experiments in AVSP show quite clearly that substantial AVSR potential was not explored yet. Related to this last aspect there is another reason, quite obvious, that is based on the large amount of empirical results gathered along several decades of research in AVSP. A substantial part of those results can be very useful on developing AVSR systems, leading to the direct improvement of current techniques or else inspiring new ones. For instance, extensive studies on the analysis of linguistic and *paralinguistic* information associated to different facial features have been carried out, and potentially some of those results can lead to effective improvements in AVSR front-end modules. A third reason for the AVSP interest to AVSR is that the referred knowledge migration can also be advantageous in the case of some more theoretical work in the AVSP area. Even if some of those works are mostly speculative and lack an acceptable consensus, by the other side some of those discussions may become very inspiring. For instance, the discussion of essential aspects in AVSP concerning the early- *versus* late-integration of the acoustic and visual *stimuli* associated to speech can be quite interesting for AVSR, where that is a key topic too.

Why AVSR succeeds or fails and its potential.

Empirical knowledge migration.

Theoretically-based inspiration.

The author's main interests on speech technology are precisely in the AVSR sub-domain, having developed along approximately a decade a few small and medium vocabulary continuous speech recognizers based on the combination of acoustic and visual clues. Though, only recently and based on relatively casual readings the author realized how relevant work devel-

The casual origin of this text.

oped in AVSP can be for the AVSR field. So, he decided to follow that line based on selected readings focused on AVSP or on Human Speech Perception (HSP), which are fully listed in the Reference Section. Those readings allowed to gain an interesting insight of the AVSP *state of the art* and lead to this text, which can be seen as a brief survey on AVSP from an AVSR perspective.

First an AVSP overview, then focusing on some topics.

The structure of this document has two main parts. In the initial one (Section 2) is presented an overview of the AVSP *state of the art*. The introduction includes references to some of the discipline milestones, achieved along more than half a century. Basic knowledge on the speech-brain apparatus is then presented. Each one of the following subsections is dedicated to an AVSP topic. Some are mostly based on relevant empirical results, while other are more theoretical. In both cases the topics are addressed quite lightly. The second part (Section 3) focus on some of the mentioned topics, trying to get a deeper understanding. These topics were selected based essentially on their relevance, as much as it could be judged, for the AVSR technology. The empirical evidence acquired in AVSP concerning the relative utility of the facial features on *extended-lipreading*, or the more theoretical discussion between *early* and *late* audiovisual integration interpretations, are just two examples of those topics.

2 An overview of AVSP

text

2.1 Introduction

text

2.2 The *speech brain*

text

2.3 The AVSP in *early areas*

2.3.1 The acoustic processing

text

2.3.2 The visual processing

text

2.4 The audio visual stimulus integration

2.4.1 Basic theoretical notions

text

2.4.2 Empirical evidence

text

2.5 The audio visual speech realization

2.5.1 Relevant phenomena

text

2.5.2 Facial features *hierarchy*

text

2.6 OTHER

text

3 An AVSR perspective of some AVSP topics

3.1 Introduction

text

3.2 TOPIC-1

text

3.3 TOPIC-N

text

4 Conclusions

text

References

[1] Lynne E. Bernstein. Phonetic Processing by the Speech Perceiving Brain. In: David B. Pisoni, Robert E. Remez. The Handbook of Speech Perception. Blackwell; 2005. p 79-98.

[2] Dominic W. Massaro, Alexandra Jesse. Audiovisual speech perception and word recognition. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2nd ed. Elsevier; 2007. p 19-35.

[3] Lawrence D. Rosenblum. Primacy of Multimodal Speech Perception. In: David B. Pisoni, Robert E. Remez. The Handbook of Speech Perception. Blackwell; 2005. p 51-77.

[4] Carol A Fowler, Bruno Galantucci. The Relation of Speech Perception and Speech Production. In: David B. Pisoni, Robert E. Remez. The Handbook of Speech Perception. Blackwell; 2005. p 633-652.

[5] Roger K. Moore. Spoken language processing by machine. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2nd ed. Elsevier; 2007. p 723-738.

[6] Ruth Campbell. The processing of audio-visual speech: empirical and neural bases. Phil. Trans. R. Soc. B. 2008. (published online 7 September 2007) p 1001-1010.

[7] Friedmann Pulvermuller. Word processing in the brain as revealed by neurophysiological imaging. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2nd ed. Elsevier; 2007. p 119-139.

[8] Virginie van Wassenhove, Ken W. Grant, David Poeppel. Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45. 2007. p 598-607.

[9] Kaoru Sekiyama, Iwao Kanno, Shuichi Miura, Yoichi Sugita. Auditory-visual speech perception examined by fMRI and PET. Neuroscience Research 47. 2003. p 277-287.

[10] Lynne E. Bernstein, Edward T. Auer, Michael Wagner, Curtis Ponton. Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 39. 2008. p 423-435.

[11] Curtis W. Ponton, Lynne E. Bernstein, Edward T. Auer Jr. Mismatch Negativity with Visual-only and Audiovisual Speech. Springer Science+Business Media. 2009.

[12] Peter Indefrey. Brain-imaging studies of language production. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2nd ed. Elsevier; 2007. p 547-564.

[13] Lawrence J. Raphael. Acoustic cues to the Perception of Segmental Phonemes. In: David B. Pisoni, Robert E. Remez. The Handbook of Speech Perception. Blackwell; 2005. p 182-206.

[14] Roger P. G. van Gompel, Martin J. Pickering. Syntactic parsing. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2nd ed. Elsevier; 2007. p 289-307.

[15] Carol A. Fowler. Speech production. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2nd ed. Elsevier; 2007. p 489-501.

[16] Robert F. Port. The problem of speech patterns in time. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2nd ed. Elsevier; 2007. p 503-514.

[17] Helena M. Saldana, David B. Pisoni. Audio-visual speech perception without speech cues. -?-.

[18] Salvatore Campanella, Pascal Belin. Integrating face and voice in person perception. Trends in Cognitive Sciences. Vol.11 No.12 -?-

[19] Avril Treille, Camille Cordeboeuf, Coriandre Vilain, Marc Sato. The touch of your lips: haptic information speeds up auditory speech processing. -?-

[20] Gemma A. Calvert, Ruth Campbell. Reading Speech from Still and Moving Faces: The Neural Substrates of Visible Speech. -?-

[21] Dogu Erdener. Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages. The Language LEarning Journal. Vol.00, No.0. 2012.

[22] John P. J. Pinel. A conduo nervosa e a transmisso sinptica: como os neurnios enviam e recebem sinais. In: -?- p 105-128.

[23] John P. J. Pinel. Os mtodos de pesquisa em biopsicologia: compreendendo o que os biopsiclogos fazem. -?- p 130-142.

[24] Keith R. Kluender, Michael Kiefte. Speech Perception within a Biologically Realistic Information-Theoretic Framework. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2006. p 153-189.

[25] Jennifer S. Pardo, Robert E. Remez. The Perception of Speech. In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2006. p 201-239.

[26] Marta Kutas, Cyma K. Van Petten, Robert Kluender. Psycholinguistics Electrified II (1994-2005). In: Matthew J. Traxler, Morton A. Gernsbacher. Handbook of Psycholinguistics. 2006. p 659-663+.

[27] The Role of Speech Production System in Audiovisual Speech Perception Iiro P. Jskelinen Department of Biomedical Engineering and Computational Science, Aalto University, Espoo, Finland The Open Neuroimaging Journal, 2010, 4, 30-36

[28] Francisco Aboitiz, Ricardo Garca, Enzo Brunetti, Conrado Bosman. The Origin of Broca's Area and Its Connections from an Ancestral Memory Network. In: Broca's Region. Oxford Univ. Press. 2006. p 17-27.

[29] Michael Arbib. Broca's Area in System Perpective: Language in the Context of Action-Oriented Perception. In: Broca's Region. Oxford Univ. Press. 2006. p 153-167.

[30] Angela D. Friedirici. The Neural Basis of Sentence Processing: Inferior Frontal and Temporal Contributions. In: Broca's Region. Oxford Univ. Press. 2006. p 196-211.

[31] Audrey R. Nath, Michael S. Beauchamp. Dynamic Changes in Superior Temporal Sulcus Connectivity during Perception of Noisy Audiovisual Speech. The Journal of Neuroscience, 31(5). (Feb. 2) 2011. p 1704-1714.

[32] Argiro Vatakis, Asif A. Ghazanfar, Charles Spence. Facilitation of multisensory integration by the -*-unity effect-?- reveals that speech is special. Journal of Vision (2008) 8(9):14. 2008. p 1-11.

[33] Thomas Ethofer, Gilles Pourtois, Dirk Wildgruber. Investigating audiovisual integration of emotional signals in the human brain. In: Anders, Ende, Jungho, Ifer. Progress in Brain Research, Vol. 156 Kissler & Wildgruber Eds. Elsevier B.V. 2006.

[34] Ville Ojanen. Neurocognitive mechanisms of audiovisual speech perception. An academic dissertation for the degree of Doctor of Philosophy. Helsinki University of Technology. 2005.

[35] Jeremy I. Skipper, Howard C. Nusbaum, Steven L. Small. Lending a helping hand to hearing: another motor theory of speech perception. -?- 2005.

[36] Johanna Pekkola. Seeing and Hearing Speech, Sounds, and Signs: Functional Magnetic Resonance Imaging Studies on Fluent and Dyslexic Readers. Academic dissertation - Faculty of Medicine of the University of Helsinki. 2006.

[37] Julie J. Yoo. FMRI Studies of Effects of Hearing Status on Audio-Visual Speech Perception. Degree of Doctor of Philosophy in Speech and Hearing Bioscience and Technology at the MIT. 2007.

[38] Julien Besle, Catherine Fischer, Aurelie Bidet-Caulet, Francoise Lecaignard, Olivier Bertrand, and Marie-Helene Giard. Visual Activation and Audiovisual Interactions in the Auditory Cortex during

Speech Perception: Intracranial Recordings in Humans. The Journal of Neuroscience, Dec. 24 (2008); 28(52). 2008. p 1430-14310?

[39] Osamu Fujimura. Invariance and variability in speech production. AT&T Laboratories.

A Basics on neuro-transmission mechanisms

text

B Basics on neuro-imaging techniques

text

C Basics on event-related potentials

text

D Links related to AVSP or AVSR

Laboratories working on AVSR
text

Researchers working on AVSR
text

OTHER-1
text

OTHER-N
text