A French-Spanish Multimodal Speech Communication Corpus Incorporating Acoustic Data, Facial, Hands and Arms Gestures Information

Lucas D. Terissi[1] *Gonzalo Sad*[1] *Mauricio Cerda*[2] *Slim Ouni*[3]
Rodrigo Galvez[2] *Juan C. Gómez*[1] *Bernard Girau*[3] *Nancy Hitschfeld-Kahler*[4]

[1]CIFASIS-CONICET, Universidad Nacional de Rosario, Argentina
[2]SCIAN-Lab, Faculty of Medicine, Universidad de Chile, Santiago, Chile
[3]Université de Lorraine, LORIA, UMR 7503, Villers-lés-Nancy, France
[4]Computer Science Department, FCFyM, Universidad de Chile, Santiago, Chile
{terissi, sad, gomez}@cifasis-conicet.gov.ar, mauriciocerda@med.uchile.cl,
nancy@dcc.uchile.cl, {bernard.girau, slim.ouni}@loria.fr

Abstract

A Bilingual Multimodal Speech Communication Corpus incorporating acoustic data as well as visual data related to face, hands and arms gestures during speech, is presented in this paper. This corpus comprises different speaking modalities, including scripted text speech, natural conversation, and free speech. The corpus has been compiled in two different languages, *viz.*, French and Spanish. The different experimental setups for the recording of the corpus, the acquisition protocols, and the employed equipment are described. Statistics regarding the number and gender of the speakers, number of words, number of sentences, and duration of the recording sessions, are also provided. Preliminary results from the analysis of the correlation among speech, head and hand movements during spontaneous speech are also presented in this paper, showing that acoustic prosodic features are related with head and hand gestures.

Index Terms: multimodal speech, human-computer interaction, multimodal corpora

1. Introduction

In recent years, the study of human communication has benefited by the increasing number of multimodal corpora available to researchers in this field. Besides the acoustic signal during speech, the visual information related to facial expressions, hand gesture and body posture contributes significantly to the intelligibility of the message being transmitted, and to the perception of the actual meaning of the message. In addition, as pointed out in a recent survey about the interaction between gesture and speech [1], the parallel use of these modalities gives the listener access to complementary information not present in the acoustic signal by itself. For instance, when the speaker says "The dog is this tall" and simultaneously indicates with his hands the height of the dog, an extra information is being provided. By using his hands, the speaker avoids producing precise verbal description of a spatial quantity (the height of the dog). In addition to provide complementary information, gestures can be *redundant* relative to the accompanying speech. For instance when the speaker says "the dog is 30 cm tall", indicating this height with his hands.

In contrast to the abundance of audio-only data corpora, there is a limited number of multimodal corpora which, in addition to the audio signal, incorporate information about facial expressions, and possibly other modalities. A thorough overview of existing multimodal corpora and the challenges and limits involved in corpus building, can be found in [2] and [3]. As pointed out in [1], building a multimodal corpus requires to make decisions about several issues such as the number and gender of the participant, the modality of the recording (monologue from scripted text or free speech, dialogue), the number and characteristics of the recording devices (single camera, multicamera, microphones, motion capture systems, devices capable of capturing depth information, like Microsoft Kinect), the languages being used (single language, or multilingual), the signals to be captured (audio, facial expressions, hands and arms gestures, body posture), the words and sentences to be recorded in the case of scripted text monologues, etc. Most of these decisions are influenced by the particular application intended for the corpus.

One of the main applications of multimodal corpora is for the evaluation of Audio-Visual Speech Recognition Systems (AVSRS) [4][5][6][7][8]. For instance in [9] and [10], the authors proposed an AVSRS which incorporates facial depth information captured by Kinect. A 3D Audio Visual Corpus consisting of numbers, words and sentences by approximately 1000 speakers is described in [11]. The visual data collected by a stereo camera is converted to 3D meshes that can be used in 3D audio-visual applications. In [12], a large Audiovisual Polish speech corpus was presented aiming to evaluate the performance of a system built of acoustic and visual features and Dynamic Bayesian Network (DBN) models for AVSR. The database has 166 speakers, and one third of them are female. Both males and females contribute equally to the corpus in terms of total recordings duration for each gender. The corpus contains 117,450 words, where 13,784 words are unique and about half of them appear only once. The Modality database [3] was designed specifically to assist audio-visual speech recognition systems (AVSR) development. This database includes high-resolution, high framerate stereoscopic video streams from RGB cameras, depth imaging stream utilizing Time-of-Flight camera accompanied by audio recorded using both, a microphone array and a microphone built in a mobile computer. It is composed by the recording of 35 speakers uttering isolated words and sentences. For audio-visual speech recognition and lip-driven animation applications, the BL (Blue Lips) database was described in [13]. It consists of 238 French sentences uttered by 17 speakers, wearing blue lipstick. The recordings were performed in two sessions, using a single front-view camera and using 2 spatially aligned cameras and a depth camera. The audio was captured by 2 microphones. The database also includes time-aligned phonetic transcriptions of audio and

video data.

For multimodal identity verification applications, the BANCA database was presented in [14]. Video and speech data from 208 subjects were recorded, where each speaker uttered speech sequences composed of numbers, speakers name, address and date of birth. These recording were obtained in three different scenarios, *viz.*, high-quality video and low noise conditions, cheap webcam in noisy office environment, and high-quality camera, noisy environment. Another application for multimodal corpora is in the area of Audio-Visual Speech Synthesis. In [15], the authors describe a synchronous corpus of acoustic and 3D facial marker data from multiple speakers for adaptive audio-visual text-to-speech synthesis. The corpus is used by the authors as training data in a HMM-based speech synthesis system.

In recent years, research in the area of human-machine interaction has been focusing on the recognition and simulation of emotional states with the ultimate goal of mimicking the communication between humans, which relies on the capability of expressing and recognizing feelings. Fundamental for the research in this area is the availability of an audio-visual affective corpus. The authors in [16] introduce a 3D Audio-Visual Corpus for Affective Communication. The corpus incorporates the most important modalities used by humans to reflect their emotional states, *viz.*, the speech signal and facial expressions. This last modality is represented by dense dynamic 3D geometries. The characteristics of the 3D vision technology required to capture the multimodal corpus in [16] is presented in [17].

Most of the above mentioned multimodal speech communication corpora incorporate only the two main modalities in human communication: acoustic signal and face (2D or 3D) expression. This paper introduces a multimodal corpus which, in addition, contains hands and arms gestures information during speech. The corpus could be employed for different purposes, for instance, visual data can be used as ground truth to evaluate facial or arms tracking methods, and synchronized audio-visual data during speech can be employed for the development of new speech recognition algorithms based on acoustic and facial information. Beside these classical applications, this corpus aims to the study of the multimodal correlation among the acoustic signal and speaker head and hand movements during spontaneous speech, which has not been extensively explored yet.

The rest of the paper is organized as follows. In section 2, the main characteristics of the corpus are presented. The acquisition devices and setups employed to compiled the corpus are described in section 3. In section 4, data post-processing procedures for stream temporal and spatial alignment are described. Some preliminary multimodal correlation analysis is presented in section 5. Finally, some concluding remarks and perspectives for future work are included in section 6.

2. Corpus description

The bilingual corpus, recorded in Spanish and French, described in this paper is composed by data acquired in two different modalities, *viz.*, spontaneous and scripted (non-spontaneous) speech. In the scripted modality, the participants were asked to read a particular set of words and sentences, and to perform different facial expressions and sounds. In this case, audio-visual information during speech is represented by the acoustic signal and facial deformation. On the other hand, in the spontaneous modality, natural conversations about different topics, between the participant and an interviewer, have been recorded. In this case, in addition to the acoustic and facial information, the movements of the hands and arms during speech were also captured.

The corpus was recorded by 4 Spanish speakers (2 females and 2 males) and by 4 French speakers (1 female and 3 males). In the non-spontaneous modality, participants were asked to:

• pronounce a set of 20 words, composed by the digits and 10 actions such as open, save and stop. Each participant uttered these words three times, sorted in random order,

• perform 6 facial expressions: surprise, disgust, smile, kiss, close their eyes and show their teeth. Each expression was repeated twice,

• pronounce 27 consonant-vowel syllables, built by the combination of consonants $/b/$, $/d/$, $/f/$, $/k/$, $/m/$, $/p/$, $/s/$, $/t/$ and $/\int/$, and vowels $/a/$, $/i/$ and $/u/$,

• read 120 phonetically balanced short sentences.

In the spontaneous modality, three different situations were considered.

• Natural conversation between two people. The interviewer starts a natural conversation with the participant, asking questions and making comments about different topics (ex. hobbies, music, work). The interviewer guides the conversation. The interviewer provides around 20% of the conversation, while the participant with the remaining 80%.

• Description of a daily item/situation (an animal, a place, a work, etc.). The interviewer only interacts if the participant does not know what else to say.

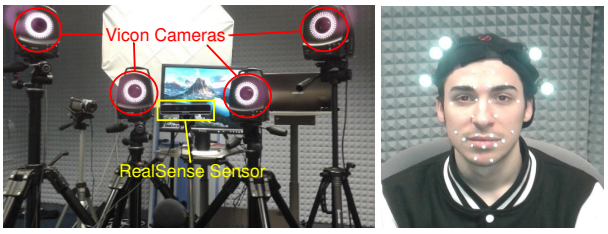
• Description of a situation or object associated with an old memory, where the participant has to make an effort to recall all the details. The interviewer only interacts if the participant does not know what else to say.

The above three situations were repeated twice, with the participant sitting down and standing up, respectively, resulting in 6 spontaneous speech recording sessions for each participant. The total duration of the spontaneous material is about 4 hours, approximately 30 minutes per participants, while the non-spontaneous recordings are around 6 hours length.

The complete corpus is freely available for research use at http://cimt.uchile.cl/mcc.

3. Data acquisition

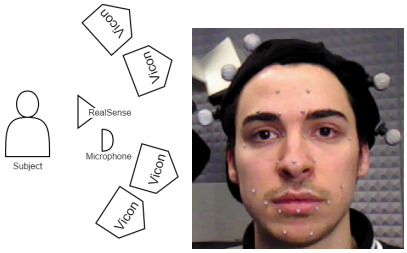
In order to capture audio-visual information simultaneously during speech three devices have been used. The acoustic signal was captured with a professional microphone and recorded at a sample rate of 44.1kHz, while for capturing visual information during speech, two commercial devices have been employed, a marker-based 3D motion capture system (Vicon Motion System Ltd.) and the marker-less capture system Intel RealSense™ F200. A Vicon system based on 4 infrared cameras (MX-3+ model) has been used to track reflective markers on the scene. The acquisition of the 3D position of the markers were provided by the Vicon Nexus software at a sampling rate of 100Hz. The Realsense system, based on a structured-light 3D sensor and a 2D camera (similar to Microsoft Kinect system), has been used to provide information about facial and hand movements at 60 Hz, based on deformable 3D generic models. In Fig. 1(a), the distribution of the acquisition devices in the scene is depicted. Examples of visual data captured by Vicon and Realsense systems are also depicted in Fig. 1(b) and Fig. 1(c), respectively. For acquiring audio-visual information three different setups were used, corresponding to the cases of non-spontaneous speech modality, and spontaneous modality in sitting down and standing up conditions, respectively.



(a) (b)

(c)

Figure 1: *Acquisition of visual information. (a) Vicon system and Realsense camera. (b) Example of reflective markers tracked by Vicon system. (c) and (d) markers captured by Realsense for the face and hands, respectively.*



(a) (b)

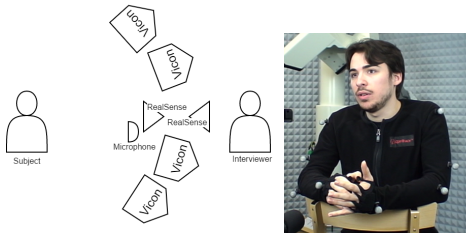
Figure 2: *Acquisition setup for non-spontaneous modality. (a) Devices and speaker location in the scene. (b) Markers layout used to capture facial movements with the Vicon system.*

3.1. Non-spontaneous speech setup

The distribution of the recording devices employed in this setup is depicted in Fig. 2(a). As mentioned before, in this acquisition modality facial movements are tracked by both Vicon and Realsense systems. The Realsense camera was placed close to the person's face, without interfering with the field of view of the Vicon cameras. In this setup, the participants were sat in front of the cameras, at a distance of 60cm from the Realsense sensor. In order to record facial movements during speech with the Vicon system, 23 reflective markers of 3mm diameter each were glued to the subject's face. These markers are distributed on the face as depicted in Fig. 2(b). Additionally, a hat with 5 markers (25mm diameter) was employed to capture global head motion, see Fig. 2(b). The text to be pronounced and the facial expressions to be performed by the participants, were prompted on a screen located behind the Realsense camera, at eye level.

3.2. Spontaneous speech setup

For the acquisition of audio-visual information in the spontaneous speech modality, in the case where the participants were sitting down, the capturing devices were located as depicted in Fig. 3. In this setup, the Vicon system was employed to track the movements of the arms. For that reason, the participant wore a vest with 8 markers (15mm) on it, located one on each shoulder, two around each elbow and one on each wrist, as shown in Fig. 3(b). In addition, a Realsense camera was used to capture participant facial and hand movements, and a second Realsense camera to capture the interviewer's movements. The participant and the interviewer were located at around 50cm from the Realsense cameras.



(a) (b)

Figure 3: *Acquisition setup for spontaneous speech modality for the case of participants sitting down.*

In the case where the participants were standing up, the distribution of the recording devices was similar to the case of sitting down condition. Nevertheless, in this setup, a single Realsense camera was not able to capture simultaneously facial and hand movements of a person, due to the fact that the face and the hands were not in the field of view of the camera all the time. For that reason, both Realsense cameras were used to capture participant visual information, one for the face and the other for the hands. The cameras were located at around 60cm from the participant's face, and at around 40cm from participant's hands. Vicon system was employed to capture arms movements.

4. Data processing

4.1. Temporal synchronization

In the recording setups described in the previous section, audio-visual data is recorded simultaneously by three (non-spontaneous modality) and four (spontaneous modality) capturing devices. This created the necessity to define a strategy for stream synchronization. In order to synchronize the information from three different devices (microphone, Vicon and Realsense cameras), a "beeper" was employed. This device simultaneously plays an audible tone (detected by the microphone), and turns on a red and an infrared LED (detected by Realsense and Vicon sensors, respectively). The beeper is activated twice at the beginning and twice at the ending of every recording. The signals emitted by the beeper are then detected on each stream, and used to manually synchronize them.

4.2. Spatial alignment

Each capturing device provides the visual information represented in its own 3D coordinate system. In order to have all the visual information in a common coordinate system, spatial alignment between the different streams was carried out.

In the case of the acquisition setup for the non-spontaneous speech modality, described in section 3.1, visual data was simultaneously captured by Vicon and Realsense systems. Some of the reflective markers glued to the person's face were located at the same position of some of the points in the face model used by Realsense system. Thus, there is a set of points that are represented in both data streams. These correspondences between both coordinates systems were used to compute the transfor-

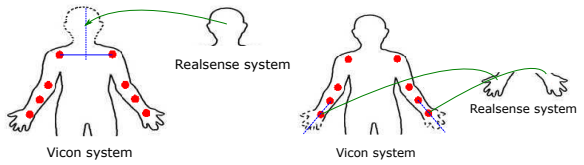


Figure 4: *(a) and (b) show the schematic representations of the spatial alignment procedures proposed for acquisition setups described in section 3.2.*

mation matrix from Realsense to Vicon coordinates using the method described in [18].

In the case of spontaneous speech modality in sitting condition, described in section 3.2, participant's facial and hand movements were captured by a Realsense camera, and arms movements by the Vicon system. Thus, in this setup, there is no data shared between the two systems. Nevertheless, taking into account that the head (captured by Realsense system) is located between the shoulders (captured by Vicon system), the spatial transformation from Realsens to Vicon coordinates can be approximated. To perform this alignment, a particular time position in the data streams is selected, where the person's is looking straight to the camera. The 3D location of the middle point between the shoulders, and the center of the head are computed in its corresponding coordinate reference systems. Then, the translation and rotation of the head are computed, where the horizontal and depth coordinates are determined by the middle point between the shoulders, and the vertical translation is manually adapted taking into account the participant's anatomy. This alignment procedure is schematically depicted in Fig. 4(a). Even though it is not a perfect transformation between both systems, it is good enough to perform the transformation from Realsense data to Vicon coordinates.

In the case of spontaneous speech modality in standing up condition, participant's movements have been captured by three devices, Vicon system for the arms, one Realsense for the face, and a second Realsense for the hands. The method used in the sitting down condition is employed to translate facial information from Realsense to Vicon coordinate system. Then, the mapping of the position of the hands from Realsense to Vicon coordinates has been performed, using a similar strategy, taking into account the position of the wrists. This transformation is schematically depicted in Fig. 4(b).

5. Multimodal analysis

In this section, preliminary results obtained from the analysis of the multimodal correlations among the fundamental frequency of the speech signal (F0), speaker head movements and hand movements during spontaneous speech, are presented. To perform this task, the Correlation Map Analysis (CMA) tool described in [19] is employed. CMA produces a 2D correlation map, in which correlation is characterized as a function of time and temporal offset, by computing instantaneous correlation for a range of time offsets between two signals. CMA is useful to identify correlated patterns that might be time-shifted. The results from these experiments agree with the analysis reported in [20], pointing out that there exists a strong relationship between head motion and acoustic prosodic features. The same experiment was performed on the non-spontaneous material of the corpus, where the speakers read out a set of sentences. However, in this case the CMA shows a low correlation, suggests

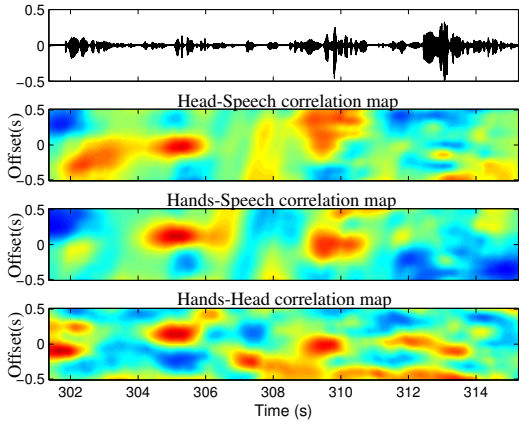


Figure 5: *CMA examples obtained by comparing head motion vs F0, hands motion vs F0, and head vs hand motion, respectively, during spontaneous speech.*

that the relationship between F0 and head motion is mainly associated with spontaneous speech. The experiments comparing F0 and hands motion also exhibit a high correlation between both signals. Analysing the associated video recordings, it was found that these high correlations occur, for example, when the speaker move their hands while he is describing an object or when he emphasises his speech.

Figure 5 shows three correlation maps obtained by comparing the speaker head motion vs F0, hands motion vs F0, and head vs hand motion, respectively. As it can be observed at 305 and 310 seconds, the three maps show high correlation. This simultaneousness in high values of correlation is frequently observed along the whole database. This could suggest that the three signals have a common event structure. The analysis of all the participants suggests that the correlation between F0 and hand movements depends on the expressivity of each speaker, showing different degrees of hand gesticulation. On the other hands, the level of correlation between F0 and head movements is similar for all the participants, independently of their expressiveness.

The previous comments hold for both Spanish and French speakers. Due to space limitations, video examples illustrating these analyses are included as additional material of this paper.

6. Conclusions

In this paper, a Bilingual (French-Spanish) Multimodal Speech Communication Corpus incorporating acoustic data as well as visual data related to face, hands and arms gestures during speech, was presented. The corpus comprises spontaneous and non-spontaneous speech. A preliminary analysis was presented showing the level of correlation between the different modalities (acoustic signal, hand gestures and head motion). The results are promising in the sense that they show the utility of the corpus for the study of multimodal communication. Future works regarding the corpus annotation are planned. The data will be annotated in terms of the speech, the facial movement, and the hands and arms gestures.

7. Acknowledgements

Research on this work has been partially funded by the STIC-AmSud Project: 15STIC-05 - Multimodal Communication Corpus (MCC).

8. References

[1] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[2] D. Knight, "The future of multimodal corpora," *RBLA*, vol. 11, no. 2, pp. 391–415, 2011.

[3] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski, "An audio-visual corpus for multimodal automatic speech recognition," *Journal of Intelligent Information Systems*, vol. 49, no. 2, pp. 167–192, Oct 2017.

[4] B. Lee, M. Hasegawa-johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "Avicar: Audio-visual speech corpus in a car environment," in *in International Conference on Spoken Language Processing, Jeju, Korea*, 2004, pp. 2489–2492.

[5] C. Sanderson and B. C. Lovell, *Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 199–208.

[6] T. J., M. Hrúz, P. Campr, and M. Železný, "Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition," 2008.

[7] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: using mobile phone data," in *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, Jul. 2012.

[8] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *In Proc. ICASSP*, 2002, pp. 2017–2020.

[9] G. Galatas, G. Potamianos, and F. Makedon, "Audio-visual speech recognition incorporating facial depth information captured by kinect," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO2012)*. Bucharest, Romania: EURASIP, August 2012, pp. 2714–2717.

[10] G. Galatas, G. Potamianos, D. Kosmopoulus, C. McMurrough, and F. Makedon, "Bilingual corpus for avasr using multiple sensors and depth information," in *Proceedings of the AVSP2011*. Bucharest, Romania: EURASIP, August 2011, pp. 103–106.

[11] C. Sui, S. Haque, R. Togneri, and M. Bennamoun, "A 3D audio-visual corpus for speech recognition," in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Sydney, Australia, December 2012, pp. 125–128.

[12] P. Żelasko, B. Ziółko, T. Jadczyk, and D. Skurzok, "AGH corpus of polish speech," *Language Resources and Evaluation*, vol. 50, no. 3, pp. 585–601, Sep 2016.

[13] Y. Benezeth, G. Bachman, G. Lejan, N. Souviraa-Labastie, and F. Bimbot, "BL-Database: A French audiovisual database for speech-driven lip animation systems," INRIA research report n RR-7711, Tech. Rep., 2011.

[14] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, *The BANCA Database and Evaluation Protocol*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 625–638.

[15] D. Schabus, M. Pucher, and G. Hofer, "Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC2012)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

[16] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-D audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, October 2010.

[17] ——, "3D vision technology for capturing multimodal corpora: Chances and challenges," in *Proceedings of LREC on Multimodal Corpora*, 2010, pp. 1–4.

[18] P. J. Besl and H. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[19] A. V. Barbosa, R.-M. Dchaine, E. Vatikiotis-Bateson, and H. C. Yehia, "Quantifying time-varying coordination of multimodal speech signals using correlation map analysis," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2162–2172, 2012.

[20] K. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004.