Audiovisual speech analysis with deep learning

Pedersen, Nicolai Fernández

Publication date:
2021

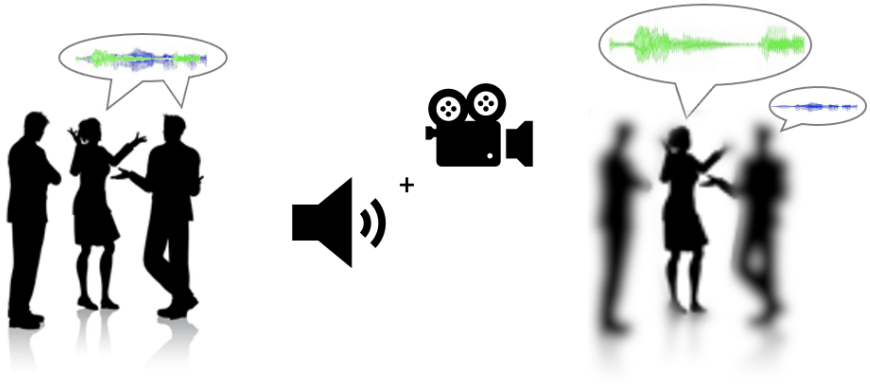
Document Version
Publisher's PDF, also known as Version of record

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 50

Nicolai Fernández Pedersen

Audiovisual speech analysis with deep learning



DTU Health Tech
Department of Health Technology

Audiovisual speech analysis with deep learning

PhD thesis by
Nicolai Fernández Pedersen

Technical University of Denmark

2021

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Health Technology, Technical University of Denmark.

The project was partly financed by COCOHA/CHESS (2/3) and by the Technical University of Denmark (1/3).

Supervisors

Senior Researcher Jens Hjortkjær
Hearing Systems Group
Department of Health Technology
Technical University of Denmark
Kgs. Lyngby, Denmark

Prof. Torsten Dau
Hearing Systems Group
Department of Health Technology
Technical University of Denmark
Kgs. Lyngby, Denmark

Prof. Lars Kai Hansen
Cognitive Systems
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

During speech production, the movement of speech articulators creates visual signals that are temporally aligned with the acoustic speech signal. These visual speaker cues have been found to facilitate speech perception in humans, especially in noisy auditory environments such as "cocktail-party" scenarios. Besides facilitating human speech perception, it is also well established that machines can learn to utilize visual speaker cues to inform auditory representations of speech. Visual speaker cues from target speakers have thus been shown to improve the performance of both automatic speech recognition systems and speech separation systems in contrast to audio-only systems. However, while many studies have investigated the temporal correspondences between auditory and visual signals, there is still a lack of knowledge about the nature of these audiovisual (AV) cues and how the two modalities are related.

This thesis aimed to contribute to a better understanding of the relationship between auditory and visual cues created during speech production. By utilizing recent advances in computer vision and data-driven approaches, natural AV speech was investigated across thousands of speakers. First, using a linear canonical correlation analysis (CCA), two primary temporal ranges of envelope fluctuations related to facial motion across speakers were identified. Amplitude envelope modulations distributed around 3-4 Hz were related to mouth openings, whereas 1-2 Hz modulations were related to more global face and head motion. Next, nonlinear neural networks were trained through a self-supervised learning scheme to learn correlated AV embeddings from natural AV speech videos. Highly correlated AV features primarily located around the mouth and jaw were identified. Based on these insights, it was examined whether the different AV features could assist a speech separation model in extracting the acoustic speech stream of a target talker from multi-talker audio mixtures. More correlated AV feature embeddings translated to better speech separation performance. Notably, the speech separation models achieved a performance comparable to more computational complex systems while showing promise for real-time implementation.

Overall, this thesis provided new insights into how auditory and visual speech cues are related and showed their usefulness in audiovisual speech separation.

Resumé

Under taleproduktion danner bevægelsen af taleartikulatorer visuelle signaler som tidsmæssigt matcher variationer i det akustiske talesignal. Disse visuelle cues har vist sig at fremme taleforståelsen hos mennesker, specielt i lydmæssigt støjende miljøer såsom "cocktail-party"scenarier. Udover at fremme menneskelig taleforståelse, så er det også veletableret at maskiner kan bruge visuelle cues til at informere auditive repræsentationer af tale. Visuel information om taleren har således vist at forbedre både automatiske talegenkendelsessystemer og taleseparationssystemer i modsætning til traditionelle systemer der kun processerer lyden. Mens mange studier har undersøgt de tidsmæssige sammenhænge mellem audio- og video-signaler, er der stadig mangel på viden om karakteren af disse audiovisuelle (AV) cues, og hvordan de to modaliteter hænger sammen.

Denne afhandling har til formål at bidrage til en bedre forståelse af sammenhængen mellem auditive og visuelle signaler skabt under taleproduktion. Ved at bruge de seneste fremskridt inden for computervision og datadrevne metoder kunne naturlig AV-tale undersøges på tværs af tusindvis af talere. Først præsenteres en lineær kanonisk korrelationsanalyse, der identificerer to primære frekvensområder for envelope-fluktuationer, der korrelerer ansigtsbevægelser på tværs af talere. Amplitude envelope-modulationer fordelt omkring 3-4 Hz var temporalt associeret med mundåbninger, hvorimod 1-2 Hz modulationer korrelerede mere med globale ansigts- og hovedbevægelser. Dernæst blev nonlineære neurale netværk trænet gennem et self-supervised læringsskema til at lære korrelerede AV-embeddings fra naturlig AV-tale video. Højt korrelerede AV-features blev primært identificeret omkring mund og kæbe. På baggrund af disse fund blev det undersøgt om forskellige AV-features kunne hjælpe en taleseparationsmodel til at udtrække den akustiske talestrøm fra en bestemt taler fra multi-taler lydblandinger. Højere korrelerede AV-feature-embeddings medførte til en bedre taleadskillelsesperformance. Navnlig opnåede taleseparationsmodellerne en performance der kan sammenlignes med mere beregningsmæssige komplekse systemer og viste samtidigt lovende resultater for realtidsimplementeringer.

Samlet set gav denne afhandling ny indsigt i hvordan auditive og visuelle cues er relaterede og viste deres anvendelighed i audiovisuel taleseparation.

Acknowledgments

First of all, I would like to express my gratitude toward my supervisors. It has been extremely encouraging to work with such intelligent, positive, and open-minded supervisors. Jens Hjortkjær, thank you for all your support, feedback, guidance and for being so passionate about the project. Thanks to Torsten Dau for the exciting discussions, constructive feedback, and always finding time to help me or just discuss German football. Thanks also to Lars Kai Hansen for your supervision during my B.Eng., my M.Sc., and my Ph.D., and for seeing a scientist in me.

Thanks to all my colleagues and friends at the Hearing Systems Group and the Acoustic Technology Group. Thank you for being supportive and for creating a great working atmosphere. Thank you to my fellow PhD-office-friends for good times at the office. Also, a big thank you to Caroline van Oosterhout for the good talks and for helping me with administrative tasks.

A special thanks to my friends and family. To my friends from football for always being there throughout the ups and downs of the past three years. To Lasse for your help and for the many hours we spent studying together in the past three years. I would also like to give a big thanks to my parents and sister for taking care of me. Last but not least, thanks to my wife Isabel for your unconditional support and love throughout the good and bad times in the past three years; you made this possible.

Related publications

Journal papers

- Tuckute, G., Hansen, S. T., Pedersen, N., Steenstrup, D., and Hansen, L. K. (**2019**). "Single-Trial Decoding of Scalp EEG under Natural Conditions," Computational Intelligence and Neuroscience, 2019, 1687-5265

- Pedersen, N. F., Dau, T., Hansen, L. K., and Hjortkjær, J. (**In prep**). "Modulation Transfer Functions for Audiovisual Speech"

- Pedersen, N. F., Dau, T., and Hjortkjær, J. (**In prep**). "Self-Supervised Learning of Correlated Audiovisual Features"

- Pedersen, N. F., Dau, T., Wen, C., Ceolini, E., and Hjortkjær, J. (**In prep**). "Audiovisual Speech Separation with Multisensory Features"

Contents

1 General introduction

1.1 Audiovisual cues during speech communication

It is well known that seeing a talker's face can improve the comprehension of auditory speech compared to listening without visual inputs (Sumby and Pollack, 1954). Especially in noisy settings, speech comprehension is improved if the talker's articulatory gestures are visible to the perceiver (Erber, 1975; Helfer, 1997; MacLeod and Summerfield, 1987; Ross et al., 2007; Schwartz et al., 2004) and improve comprehension when communicating semantically complex messages (Arnold and Hill, 2001; Reisberg et al., 1987). Seeing the face of a speaker has also been shown to ease speech recognition in hearing-impaired as well as cochlear implant listeners (Grant et al., 1998; Rouger et al., 2007; Schorr et al., 2005). While seeing the face of the talker in many cases benefits the listener, situations can arise where the visual integration can cause illusory effects (McGurk and MacDonald, 1976). The "McGurk illusion" represents a famous illusion of this, which was accidentally discovered when McGurk and MacDonald (1976) observed that an auditory 'ba' sound presented together with a visual presentation of a 'ga' caused the listeners to hear an illusory 'da' sound.

Access to visual cues does not only help human auditory perception but can also inform machine hearing. In particular, video information can greatly enhance the performance of automatic speech recognition (ASR) or speech separation (SS) systems (Girin et al., 2001; Ochiai et al., 2019; Potamianos et al., 2003). As in human perception, access to visual information allows artificial systems to extract representations of a target speech source in a noisy acoustic background (Afouras et al., 2020). In recent years, AV speech separation systems based on deep neural networks have outperformed traditional statistical approaches (Michelsanti et al., 2021). Deep neural network (DNN)-based speech separation or speech enhancement systems benefit from the abundance of readily available AV data, which has made it possible to train better performing and more complex deep learning models (Afouras et al., 2018a,b, 2020; Ephrat

et al., 2018).

Nevertheless, while both humans and machines can utilize visual cues to inform auditory representations of speech, there is still a lack of knowledge about what those cues are. It is generally assumed that temporal correspondences between auditory and visual signals are pivotal. During speech production, the movement of speech articulators creates visual signals that are temporally aligned with the speech signals. Deciphering the underlying relationship between the facial movements and the resulting speech has therefore attracted attention. Using infrared emitting diodes on talking faces, Munhall and Vatikiotis-Bateson (1998) tracked the face's movements and found that vertical facial motions tend to be below 10 Hz. Müller and MacLeod (1982) found that facial movements during speech tended to be dominated by slow quasiperiodic motion below 10 Hz. Ohala (1975) studied the jaw motion during oral reading and concluded that the primary spectral peak of the movements was around 4 Hz. Although variation across languages and speakers exists, these results have been found to correspond well with the average rate at which syllables are produced in natural speech (Goswami and Leong, 2013; Jacewicz et al., 2009; Pellegrino et al., 2011; Varnet et al., 2017). While the motions related to speech production usually tend to move at rates around 4 Hz, larger and slower head movements, such as head nodding and eyebrow movements, have been found to be related to prosodic speech events (Guaïtella et al., 2009; Hadar et al., 1984; Hadar et al., 1983; Kim et al., 2014; McClave, 1998). Extraoral facial movements also provide valuable cues for speaker identification, especially if oral cues are absent (Thomas and Jordan, 2004).

Generally, the motion of orofacial articulators during speech production is correlated with amplitude envelope fluctuations in the acoustic speech in the 1-8 Hz range (Chandrasekaran et al., 2009; Ding et al., 2017; Munhall and Vatikiotis-Bateson, 1998). Using video data of talking speakers, Chandrasekaran et al. (2009) estimated the correspondence between wideband envelopes and the mouth area and found a strong correlation in the range 2-7 Hz. However, it is not yet well understood at what rates speech modulations correlate to visible movement in different parts of the talker's face or head. More generally, the range of temporal AV correspondences that humans and machines can exploit is still not well explored.

A more solid knowledge of the AV statistics that underlie human and machine perception could also lead to better and more lightweight AV feature

representations. Such feature representations might be vital in implementing real-time AV speech separation systems in low-resource applications, such as hearing assistive devices and smart devices.

1.2 Audiovisual speech separation

Both humans and machines experience difficulties in extracting a specific sound source in noisy environments (Michelsanti et al., 2021; Shinn-Cunningham and Best, 2008). Thus, potential applications for systems capable of extracting target sounds from noisy audio mixtures are countless, making it an attractive research topic.

Audio-only speech separation systems often perform relatively well when the background noise is stationary, making it easily distinguishable from the speech (Wang and Chen, 2018). However, audio-only speaker-independent speech separation systems suffer from the "source permutation problem" that arises when the separated speech signals are inconsistently assigned to the sources (Michelsanti et al., 2021). By including the visual information related to the talker, source separation becomes less of a problem, as the system can utilize the visual information to extract the target speech. Furthermore, the visual information typically is unaffected by noisy actions in the acoustic scene, making it a reliable supporting signal. Similar to what has been observed in humans, including visual information also makes AV speech separation systems perform better than audio-only systems (Michelsanti et al., 2021). While early-generation AV speech separation systems were based on traditional statistical approaches, the recent amount of readily available AV data have made it easier to train better performing and more complex deep learning models (Afouras et al., 2018a,b, 2020; Ephrat et al., 2018). These systems, however, often only work in offline settings and are too computational heavy to be implemented in low-resource devices, which greatly reduces the number of potential applications for such systems.

A general approach to train AV speech separation models is to use carefully designed low-level AV features, such as Mel Frequency Cepstral Coefficients (MFCCs) and facial landmark-based features. These low-level AV features are distilled versions of the original audio and video data that have properties thought to be beneficial for speech separation models' ability to perform speech separation. It is, however, not guaranteed that the low-level AV features are

optimal for speech separation. Contrary to training on low-level AV features, recently proposed AV speech separation models based on DNNs have been trained directly on raw visual and acoustic signals (Ideli et al., 2019; Wu et al., 2019). When training directly on raw input data, the neural networks are forced to distill the audio and video information into AV features that enable speech separation. Training directly on raw input data is usually a computationally demanding task that requires much training data, but it can potentially lead to specifically optimized AV features. Analyzing the AV features learned by the neural networks might, therefore, assist in understanding how the two modalities are related (Aldeneh et al., 2021; Ravanelli and Bengio, 2018a).

1.3 Overview of the thesis

This thesis aims to contribute to a better understanding of the relationship between audio and visual cues created during speech production. AV speech is investigated using data-driven approaches that enable the analysis across thousands of speakers. It is examined whether visual speech cues that are correlated with auditory speech can be exploited for better speech separation. A speech separation system is presented that uses visual speaker cues to extract the corresponding speech from single-channel audio mixtures.

In *Chapter 2*, a large-scale analysis of natural AV speech is presented. Canonical correlation analysis (CCA) is used to decompose the facial movements constituting speech production and to identify the modulation transfer function of each of these decomposed facial movements. CCA is also used to decompose the audio signal in the envelope domain to investigate which envelope rates are related to facial movement in different parts of the face. The analysis is based on recent technological advances in deep learning to estimate three-dimensional (3D) facial landmarks on two large-scale AV speech datasets, which enables the analysis of many hours of natural AV speech video. Compared to studies based on, e.g., manual motion tracking, this approach allows the study of natural AV statistics on a larger scale and the investigation of how AV correlations generalize across many speakers.

Chapter 3 extends the linear CCA approach in chapter 2 by using non-linear neural networks to learn AV correspondences. A self-supervised approach is presented to train interpretable AV-based convolutional neural networks (CNNs) directly on raw video and audio inputs. Two AV fusion models are trained and

compared. Both models are trained to maximize the correlation between AV feature embeddings when the segments are temporally aligned while trained to minimize the correlation when the segments are misaligned. In one approach, standard one-dimensional (1D) convolutions are employed. In the second approach, sinc-based convolutions are employed to ease the interpretability of the learned audio filters. Besides analyzing the AV feature representations, the proposed network architecture also allows the visualization of which aspects of the raw input data the AV network focuses on.

In *Chapter 4*, it is investigated how the findings from *chapter 2* and *chapter 3* can be utilized for automatic speech separation. The AV features learned by CCA (in *chapter 2*) and the deep sincnet (from *chapter 3*) are compared, along with a novel AV fusion strategy. It is hypothesized that higher AV correlation in the feature embeddings should provide for a strong input in the downstream task of speech separation. To test this, speech separation models are trained to use visual feature embeddings from target speakers as guiding signals to help the models extract the target speech from single-channel audio mixtures. The performance of the AV speech separation system is evaluated in different acoustic conditions, and perspectives for real-time implementations are discussed.

Finally, *Chapter 5* summarizes the main results, discusses their implications, and explores future directions for data-driven analyses of AV speech.

2

Modulation Transfer Functions for Audiovisual Speech[a]

Abstract

Temporal synchrony between facial motion and acoustic modulations is a hallmark feature of audiovisual speech. The moving face and mouth during natural speech are known to be correlated with low-frequency acoustic envelope fluctuations (below 10 Hz), but the precise rates at which envelope information is synchronized with motion in different parts of the face are less clear. Here, we used regularized canonical correlation analysis (rCCA) to learn speech envelope filters whose outputs correlate with motion in different parts of the speaker's face. We leveraged recent advances in video-based 3D facial landmark estimation, allowing us to examine statistical envelope-face correlations across a large number of speakers (∼4000). Specifically, rCCA solutions were regularized to learn modulation transfer functions (MTFs) for the speech envelope that significantly predict correlation with facial motion across speakers. The AV analysis revealed bandpass speech envelope filters at distinct temporal scales. A first set of MTFs showed a peak around 3-4 Hz and were correlated with mouth movements. A second set of MTFs captured envelope fluctuations in the 1-2 Hz range correlated with more global face and head motion. The two distinctive timescales emerged only as a property of natural AV speech statistics across many speakers. A similar analysis of fewer speakers performing a controlled speech task highlighted only the well-known temporal modulations around 4 Hz correlated with orofacial mo-

[a] This chapter is based on Pedersen, N. F., Dau, T., Hansen, L. K., and Hjortkjær, J. "*Modulation Transfer Functions for Audiovisual Speech*" (in prep).

tion. The different bandpass ranges of AV correlation align notably with the average rates at which syllables (3-4 Hz) and phrases (1-2 Hz) are produced in natural speech. Whereas periodicities at the syllable rate are evident in the envelope spectrum of the speech signal itself, slower 1-2 Hz regularities thus become prominent when considering AV signal statistics. This may indicate a motor origin of temporal regularities at the timescales of syllables and phrases in natural speech.

2.1 Author summary

Natural speech signals are dominated by slow fluctuations (<10 Hz) in the acoustic speech envelope. A peak in modulation energy around 3-4 Hz corresponds to the average rate at which syllables are produced in natural speech, but speech carries temporal information at multiple timescales. Here, we show that audiovisual speech statistics derived from natural speech across many speakers reveal different and distinct timescales of envelope fluctuations correlated with different kinematic components of the speaker's face. Using regularized canonical correlation analysis, we analyzed a comprehensive natural speech video dataset to derive modulation transfer functions for the speech envelope conditioned on correlations with facial motion. Distinct timescales of audiovisual correlation emerged: (i) speech envelope fluctuations around 3-4 Hz correlated with mouth openings, as expected, and (ii) slower 1-2 Hz envelope fluctuations correlated with more global face movements. These different envelope frequency regions align notably with the timescales of syllables and phrases in natural speech and may point to a motor origin of temporal regularities in speech at these privileged rates.

2.2 Introduction

Seeing a person's face is known to influence auditory speech perception (McGurk and MacDonald, 1976) and can improve speech intelligibility in noisy environments (Sumby and Pollack, 1954). Visual cues can also inform automatic speech recognition (Potamianos et al., 2003) or speech separation systems (Ephrat et al., 2018). Audiovisual speech perception is thought to hinge on temporal correspondences between the auditory and visual signals received by the perceiver.

Both amplitude envelope fluctuations in the acoustic speech signal and the motion of orofacial articulators during speech production are dominated by slow 'rhythms' predominant in the 1-8 Hz range (Chandrasekaran et al., 2009; Ding et al., 2017; Munhall and Vatikiotis-Bateson, 1998). However, the details of how speech modulations at different rates correlate with visible movement in different parts of the talker's face or head are still not fully understood.

Orofacial movements during speech production display relatively slow quasi-regular kinematics. Studies measuring jaw, lip, or tongue movements during speech have reported regular motion patterns predominantly below 8 Hz (Bennett et al., 2007; Lindblad et al., 1991; Matsuo and Palmer, 2010; Munhall and Vatikiotis-Bateson, 1998; Walsh and Smith, 2002). Ohala (1975), for example, reported histograms of intervals between jaw openings measured during running speech, showing a peak frequency in the 3-6 Hz range. This corresponds to the average rate at which syllables are produced in natural speech, although variation exists across languages and speakers (Jacewicz et al., 2009; Pellegrino et al., 2011; Varnet et al., 2017). The natural syllable production rate has also been argued to determine the shape of the modulation spectrum of natural speech signals (Greenberg et al., 2003), consistently showing a peak frequency around 4 Hz across different languages and speech corpora (Ding et al., 2017; Singh and Theunissen, 2003; Varnet et al., 2017).

However, the co-existence of slow periodicities in face movements and in the produced speech signal does not by itself specify the details of how they are related. It also does not reveal which dynamic visual cues are available in audiovisual speech perception or decodable from video inputs of a speaker's face. Some periodic movements occurring during speech may not be related to the production of sound or necessarily correlated with any acoustic events (e.g., blinking). Conversely, natural speech sounds contain amplitude modulations that may not be directly related to any visible movement available to the perceiver (such as speech modulations produced predominantly by phonatory activity). Although the two domains share a temporal axis, the temporal characteristics of the relation between visible motion and speech acoustics remain to be specified.

Several previous studies have examined correlations between orofacial movements and different features of the acoustic speech signal (Chandrasekaran et al., 2009; Kuratate et al., 1999; Munhall and Vatikiotis-Bateson, 1998; Yehia et al., 2002). Most work has considered temporal envelope representations ex-

tracted by low-pass filtering the speech audio waveform. Chandrasekaran et al. (2009) reported a correlation between speech envelopes and the area of mouth openings extracted from speech videos. To extract the envelope, the speech signal was first filtered in the audio frequency domain, Hilbert transformed, and down-sampled to 25 Hz, but the envelope was not decomposed further. To examine the relation between the mouth area and the speech envelope as a function of temporal modulation frequency, the spectral coherence between the audio and video signal features was examined. This suggested that mouth openings and speech envelopes both contain temporal modulations in the 2-6 Hz range. Alexandrou et al. (2016) reported a similar range of spectral coherence between speech envelopes and electromyographic lip and tongue recordings. Coherence analyses of this type demonstrate that auditory and visual signals display some degree of periodicity in the same spectral range. However, spectral coherence does not extract potential different sources of covariance in the spectral range where coherence is observed. This requires a decomposition of the covariance structure in the envelope domain.

The majority of studies have focused on the correlation between speech acoustics and movements of the mouth. However, other parts of the face or body move as well during natural speech (Wagner et al., 2014). Some of these may be coupled with orofacial articulators in speech motor control. Other gestures performed during naturalistic speech may not be directly involved in sound production but may nonetheless be consistently correlated with sound features. Rhythmic head nodding or eyebrow movements during speech, for instance, have been associated with speech prosody (Guaïtella et al., 2009; Hadar et al., 1984; Hadar et al., 1983; Kim et al., 2014; McClave, 1998). Head or body movements may thus also correlate with variations in acoustic features (Munhall et al., 2004; Pouw et al., 2020b; Yehia et al., 2002) but presumably at slower rates given the kinematics of head or body motion (Grimme et al., 2011). More generally, it remains unclear how different parts of the talking face and head may be correlated with different rates of acoustic variation in the speech signal during natural speech.

This question is complicated by the fact that different moving parts of the face are themselves mutually correlated during natural speech. Individual articulators do not move independently but are synergistically coordinated via common neuromuscular control (Vatikiotis-Bateson et al., 1996) or biomechanical coupling (Pouw et al., 2020a). For example, movements of the hyoid, jaw,

and tongue display a unique and rate-specific degree of coupling during speech, and the coupling is distinct from other behaviors such as chewing (Ghazanfar et al., 2012; Hiiemae et al., 2002; Matsuo and Palmer, 2010; Moore et al., 1988). Since different parts of the speech motor system are coordinated, it is necessary to consider how different parts of the face form groups with common kinematics. Data-driven dimensionality reduction techniques have been used to analyze facial motion data recorded during speech production in order to identify spatial components that follow shared motion patterns (Kuratate et al., 2005; Lucero et al., 2005; Lucero and Munhall, 2008; Ramsay et al., 1996). Lucero & Munhall (2008) used QR factorization to identify groups of linearly dependent facial markers, revealing a set of *kinematic eigenregions* in the speaking face (Lucero and Munhall, 2008). Consistently across two talkers, such eigenregions were identified for the lower and upper parts of the mouth and each of the mouth corners. Regions in other non-oral parts of the face were also identified, such as the left and right eyebrows and the two eyes (Lucero and Munhall, 2008). Such data-driven analyses of facial markers may capture the spatial *degrees of freedom* or dimensionality of facial kinematics during speech production but may also identify spatial components that are not necessarily related to the acoustic speech signal.

In the current study, we present an AV analysis approach based on *canonical correlation analysis* (CCA) that linearly transforms *both* visual and audio signals to capture the correlational structure between them. This approach simultaneously segments facial landmarks (as in previous work) while filtering the speech audio signal in the envelope domain. We adapt an idea originally proposed for the analysis of electrophysiological responses to speech (Cheveigné et al., 2018) that uses CCA to learn modulation transfer functions (MTFs) in the audio envelope domain. De Cheveigné et al. (2018) applied a multichannel FIR filter bank to speech envelopes as input to the CCA (the second input being EEG brain signals) (Cheveigné et al., 2018). Each component of the CCA then linearly recombines the envelope subbands to find a filtered audio envelope that maximizes the correlation with the second input. With an appropriate choice of filters, the filter bank constitutes a *filter basis*, and CCA learns optimal coefficients on that basis (Cheveigné et al., 2018). Here, we adapt this idea to learn envelope filters that correlate with visual motion in different regions of the speaker's face. Specifically, CCA simultaneously learns a set of envelope filters and a corresponding set of eigenregions of the face. The MTFs of the envelope

filters learned by CCA can then be used to characterize the range of temporal modulation frequencies that correlate with different kinematic regions of the face.

MTFs have traditionally been used to characterize how an acoustic transmission channel, such as a room, attenuates or enhances certain modulation frequencies in the input sound signal (Houtgast and Steeneken, 1973). MTFs have also been used to characterize the sensitivity to amplitude modulations in auditory perception (Dau et al., 1996; Elliott and Theunissen, 2009; Viemeister, 1979) or physiology (Delgutte et al., 1998; Edwards and Chang, 2013). In the context of AV speech analysis, we adapt the MTF concept to characterize the range of envelope frequencies in the speech signal that are correlated with visual motion. Similar to MTFs in auditory physiology or perception, we speculated that the relation between the acoustic speech envelope and the visual face might have a band-pass character, i.e., that narrower ranges of speech modulation frequencies might be related to visible motion in different parts of the face. In contrast to its application in room acoustics or perception, the MTFs of AV speech do not map the acoustic speech signal directly to the visual signal but instead transform both signals to a latent representation learned by CCA. This is motivated by the fact that the visual signal is not directly caused by the acoustic signal or vice versa. Instead, the audio and video signals are both related to the underlying speech production system (Scholes et al., 2020) and its neuromuscular control (Fuchs and Perrier, 2005; Vatikiotis-Bateson et al., 1996).

Here, we analyzed an extensive video dataset of natural speech using CCA. Our primary analysis was based on the LRS3 (lip-reading sentences) dataset consisting of single-talker video recordings collected *in the wild* (videos from TED and TEDx talks, (Afouras et al., 2018a)). We exploited novel deep learning techniques to estimate 3D facial landmarks directly from 2D videos of the speakers. In contrast to previous work based on motion tracking, the estimation of face points from video enabled us to model the statistics of facial kinematics and their relation to speech envelope variations across a large number of speakers (>4000). Specifically, we used regularized CCA (rCCA) with regularization parameters optimized to generalize across speakers. This used this approach to examine patterns of head and face movement that are consistently correlated with the speech modulations across speakers. We derived audiovisual MTFs that were predictive across a large number of speakers in the LRS3 dataset. We

also compared the results to more well-controlled speech recordings (the GRID dataset, (Cooke et al., 2006)) used in a number of previous AV speech studies.

2.3 Materials and methods

2.3.1 Data

LRS3 dataset

The main analysis was conducted on the LRS3 dataset (Afouras et al., 2018a), containing *in the wild* videos of single speakers extracted from TED and TEDx talks in English. The predefined `trainval` training dataset consisting of 32,000 videos or approximately 30 hours of video data was used. The dataset is composed of video clips from 4,004 different speakers. The videos were recorded with a frame rate of 25 fps, an audio sample rate at 16 kHz, and the clips vary from one to six seconds in duration. Videos were excluded if the face landmarks could not be estimated, leaving a total of 30,934 videos corresponding to approximately 29.5 hours of video data.

GRID dataset

For comparison, the analysis was also performed on the GRID dataset (Cooke et al., 2006), used in a number of previous AV speech studies (e.g., Chandrasekaran et al. (2009)). In contrast to LRS3, the GRID dataset consists of data from fewer speakers performing a controlled speech task. The data consists of audio and video recordings of 34 native English speakers, each reading 1,000 predefined matrix sentences. Each sentence consists of six monosyllabic words: command, color, preposition, letter, digit, and adverb, e.g., *"place green by D 2 now"* out of a total vocabulary of 51 words. The speaker is situated in front of a neutral background and facing the camera. All videos have a duration of 3 seconds and are recorded with a video frame rate of 25 frames per second (fps) and an audio sample rate of 50 kHz (Cooke et al., 2006). Videos for one of the speakers (speaker 21) were not available. From the 33,000 available videos, a total of 32,731 videos were included in the analysis, corresponding to approximately 27 hours of video data.

2.3.2 Feature Extraction

Audio Envelope Extraction

We estimated an envelope representation of the speech audio signals (see Fig 2.1). First, the audio files were resampled to 16.000 Hz and converted to mono. The speech waveform signals were passed through a gammatone filterbank (Patterson et al., 1987) consisting of 31 filters spaced from 80 to 8000 Hz. The envelope was then computed in each gammatone subband via the Hilbert transform. Next, the envelopes in each subband were passed through a modulation filterbank comprising a set of 25 equally spaced first-order Butterworth bandpass filters with a bandwidth of 0.75 Hz and a spacing of 0.5 Hz. Each envelope subband was then averaged across the gammatone filters and resampled to 25 Hz to match the video framerate.

Visual Feature Extraction

3D-facial landmarks were extracted from the videos on a frame-by-frame basis. The landmarks were extracted using the deep learning-based face alignment network presented in Bulat and Tzimiropoulos (2017). The network first performs face identification in a given frame and then estimates the 3D position of 68 facial landmarks (see Fig 2.1). Each landmark is composed of an x, y, and z coordinate, where the x and y coordinates correspond to the location of a given landmark in the image frame, and the z coordinate is the estimated depth location of the landmark.

The landmark time series were first low-pass filtered at 8 Hz to remove jitter in the frame-to-frame estimation. Energy above this range is unlikely to stem from speaker motion that can be detected at the 25 Hz sampling rate of the video (Yehia et al., 2002). The landmarks were finally normalized to have zero mean and unit variance for each individual video.

2.3.3 Canonical Correlation Analysis

Given two multidimensional datasets, CCA learns linear transforms that maximize correlation in the shared projected space. Given two zero-mean datasets $\mathbf{X_A} \in \mathbb{R}^{T \times J_A}$ and $\mathbf{X_V} \in \mathbb{R}^{T \times J_V}$, where T denotes time, and J_A and J_V are the number of features in the different dataset. CCA estimates two weight matrices $\mathbf{W_A} \in \mathbb{R}^{J_A \times J_0}$ and $\mathbf{W_V} \in \mathbb{R}^{J_V \times J_0}$, where $J_0 \leq \min\{J_A, J_V\}$, such that linear projec-

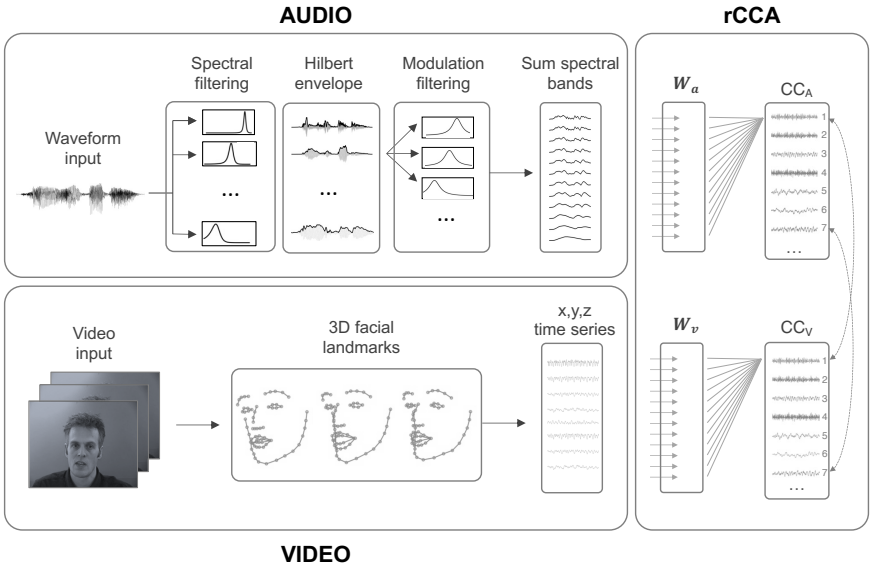


Figure 2.1: **Analysis procedure.** Regularized CCA (rCCA) combines speech envelope filter outputs and 3D landmarks of the speaker's face. Resulting pairs of canonical components (CCs) are linear combinations of envelope filter outputs for audio (CC_A) and facial landmarks for video (CC_V).

tions of each dataset $\mathbf{X_A W_A}$ and $\mathbf{X_V W_V}$ are maximally correlated. Pairs of columns of $\mathbf{X_A W_A}$ and $\mathbf{X_V W_V}$ are denoted the canonical variates or *canonical components* (CCs). The first CC pair is the linear transformation of the datasets yielding the highest correlation. The next J pairs of canonical components have the highest correlation each orthogonal to the preceding component. The components are thus ordered with respect to the size of correlation.

The objective function maximized in CCA can be formulated using the sample cross-covariance $\mathbf{\Sigma_{AV}} = \mathbf{X_A}^\intercal \mathbf{X_V}$ and the covariance matrices $\mathbf{\Sigma_A} = \mathbf{X_A}^\intercal \mathbf{X_A}$ and $\mathbf{\Sigma_V} = \mathbf{X_V}^\intercal \mathbf{X_V}$:

$$\begin{aligned} \rho &= \max \frac{(\mathbf{X_A} \cdot \mathbf{W_A})^\intercal \cdot (\mathbf{X_V} \cdot \mathbf{W_V})}{||\mathbf{X_A} \cdot \mathbf{W_A}||||\mathbf{X_V} \cdot \mathbf{W_V}||} \\ &= \max \frac{\mathbf{W_A}^\intercal \mathbf{\Sigma_{AV}} \mathbf{W_V}}{\sqrt{||\mathbf{W_A}^\intercal \mathbf{\Sigma_A} \mathbf{W_A}||||\mathbf{W_V}^\intercal \mathbf{\Sigma_V} \mathbf{W_V}||}}. \end{aligned} \tag{2.1a}$$

Since scaling of $\mathbf{W_A}$ and $\mathbf{W_V}$ does not change the correlations, we can add the constraints that $\mathbf{W_A}^\intercal \mathbf{\Sigma_A} \mathbf{W_A} = 1$ and $\mathbf{W_V}^\intercal \mathbf{\Sigma_V} \mathbf{W_V} = 1$, and hence reformulate as a Lagrangian that can be solved as a generalized eigenvalue problem.

CCA can be regularized to avoid overfitting. An L2 regularization term can

be incorporated into the objective function in Eq. (2.1a) as follows:

$$\rho = \max \frac{\mathbf{W_A}^\intercal \mathbf{\Sigma_{AV}} \mathbf{W_V}}{\sqrt{(\mathbf{W_A}^\intercal \mathbf{\Sigma_A} \mathbf{W_A} + \lambda_A \|\mathbf{W_A}\|^2) \cdot (\mathbf{W_V}^\intercal \mathbf{\Sigma_V} \mathbf{W_V} + \lambda_V \|\mathbf{W_V}\|^2)}} \tag{2.2}$$

Note that by adding regularization we effectively relax the orthogonality constraint of the canonical components.

2.3.4 AV modulation transfer functions

Here we use CCA to simultaneously learn a set of temporal modulation filters and spatial decompositions of the facial landmarks. The CCA analysis pipeline is illustrated in Fig 2.1. $\mathbf{X_A}$ is the data matrix of J_A (25) filtered subband audio envelopes, and $\mathbf{X_V}$ is the data matrix of visual features of size $T \times J_V$, where J_V is the total number of facial landmarks (3*68). We assume that linear combinations of audio and video features are correlated by virtue of both being generated by the same speech production source. Specifically, let $\mathbf{X_A} = \mathbf{A_A}\mathbf{S} + \epsilon$ and $\mathbf{X_V} = \mathbf{A_V}\mathbf{S} + \epsilon$ be a generative forward model, where the same set of speech production sources $\mathbf{S} \in \mathbb{R}^{T \times J_0}$ generate both envelope fluctuations in the audio signal $\mathbf{X_A}$ and spatial motion in the face points $\mathbf{X_V}$. $\mathbf{A}$ are filters that map between the speech source and the observed audio envelopes and video landmarks, i.e., a spectral filter in $\mathbf{A_A}$ and spatial filter in $\mathbf{A_V}$. Given the audio and video features, CCA produces two transform matrices, $\mathbf{W_A}$ and $\mathbf{W_V}$, that instead map 'backward' from the observed features to estimate the latent sources (the CCs). However, the CCA weights cannot be directly interpreted as the filter parameters $\mathbf{A}$ in the corresponding forward model (Haufe et al., 2014). The size of the CCA weights reflect both a weighting of those features that are correlated (particular combinations of envelope subbands and spatial landmarks), but also a suppression of 'noise', i.e., envelope fluctuations or visual motion that are not related to the shared speech source. However, the parameters of the corresponding generative model can be estimated as $\mathbf{A} = \mathbf{\Sigma W}$ (Haufe et al., 2014), also referred to as the canonical loadings. Unlike the CCA weights, the columns of the $\mathbf{\Sigma W}$ matrix indicate the correlation between CCs and the input features, i.e., the strength of the latent speech source in each of the observed features.

For the audio envelope features, each CC learned by CCA represents a weighted sum of the envelope subbands from outputs of the modulation filterbank. Due to the distributivity of convolution, a signal summed at the output

of an N-channel parallel filterbank with impulse responses $h_1, h_2, .., h_N$ is equivalent to a filter given by the sum of impulse responses $h_1 + h_2 +, .. + h_N$. The effective modulation transfer function learned by CCA is therefore given by the weighted sum of the impulse responses of the modulation filterbank. If $\mathbf{H} \in \mathbb{R}^{F \times J_A}$ is the set of transfer functions for the modulation filterbank with J_A channels and F frequencies, the effective MTFs learned by CCA is thus given by $\mathbf{H\Sigma_A W_A}$.

The MTFs can also be visualized by inspecting the CCs, i.e., the output of the learned filters $\mathbf{X_A \Sigma_A W_A}$. In the results below, we plot the average spectrum of $\mathbf{X_A \Sigma_A W_A}$ computed per video for each CC (Figs 2.2 and 2.4). This takes the effective average modulation energy across speakers in the dataset into account, i.e., it shows the effective outputs of the filtering process learned by CCA.

On the visual side, CCA decomposes the facial landmarks into correlated groups. The landmarks corresponding to each CC can similarly be visualized in the face by the canonical loadings, i.e., the CCA weights for each landmark scaled by the sample covariance: $\mathbf{\Sigma_V W_V}$ (also shown in Figs 2.2 and 2.4).

2.3.5 Optimization scheme

To identify statistically significant AV correlations that generalize across speakers, we trained the rCCA model using a cross-validation scheme. The dataset was first split into a test set and a training set consisting of 10% and 90% of the data, respectively. Cross-validation was then performed on the training set by further splitting the training data into five folds. Importantly, no speakers appeared in more than one data split, both for the test and training sets and for the individual cross-validation folds. This implies that the model was optimized to predict AV correlations across speakers. The rCCA was trained using a match-mismatch scheme (Cheveigné et al., 2021). During cross-validation, rCCA models were trained on correctly matching video and audio data on four of the five folds, and correlations for each rCCA component were computed on the held-out validation fold. Correlations for each component were then computed on 1000 mismatching segments of audio and video to generate an empirical null-distribution. The difference between the median correlation obtained from the mismatching data and the correlation for the matching data defined the objective function that was used to optimize the two regularization parameters. Only matching components exceeding the 95th percentile of the null-distribution were considered.

For optimization, Bayesian Optimization via Gaussian Processes was used. The optimization scheme was implemented using `scikit-opimize.gp_minimize` (Head et al., 2018). The search space for both regularization parameters, RegA and RegV, were chosen to be between $[10^{-5}, 10^{0}]$. The `scikit-opimize.gp_minimize` algorithm was initialized with a random search for the two regularization parameters, which were drawn from a log-uniform distribution with upper and lower bounds defined by the search space. After evaluating the five random searches, the algorithm approximated the next five regularization parameters with a Gaussian process estimator using a Matern kernel. The `gp_hedge` acquisition function was used, which chooses probabilistically among the three acquisition functions: lower confidence bound, negative expected improvement, and the negative probability of improvement, at each iteration. This process was repeated for each of the five validation folds, and the regularization parameters yielding the highest difference in correlations across the five-folds were used to train a final rCCA model on the entire training set. The significant rCCA components of this final rCCA model were determined on the independent test set. Significant components were defined as those exceeding the 95th percentile of the null-distribution obtained with mismatching audio and video.

2.4 Results

We used CCA to relate speech envelope information and facial motion across a large number of speakers (∼4000). Specifically, CCA learns envelope filterings that correlate with visual motion in groups of facial landmarks (see Fig 2.1). Fig 2.2 shows statistically significant canonical components (CCs) with a correlation above 1% for the main analysis on the LRS dataset. Importantly, the significance of the CCs was determined by whether they generalize across talkers. The left panels show the envelope filters learned by CCA for each CC. The right panels show the corresponding contribution of facial landmarks visualized by the 2D projection of the landmark CCA loadings. The color bars indicate the relative contribution of the x, y, z-directions.

A dynamic visualization of the facial CCs for an example speaker can be seen on GitHub [a]. This example is not a facial animation but a dynamic plot of

[a] https://github.com/NicolaiP/cca_facial_animations

the visual CCs back-projected to the input landmark space to aid interpretation of how CCA decomposes face and head movements during speech.

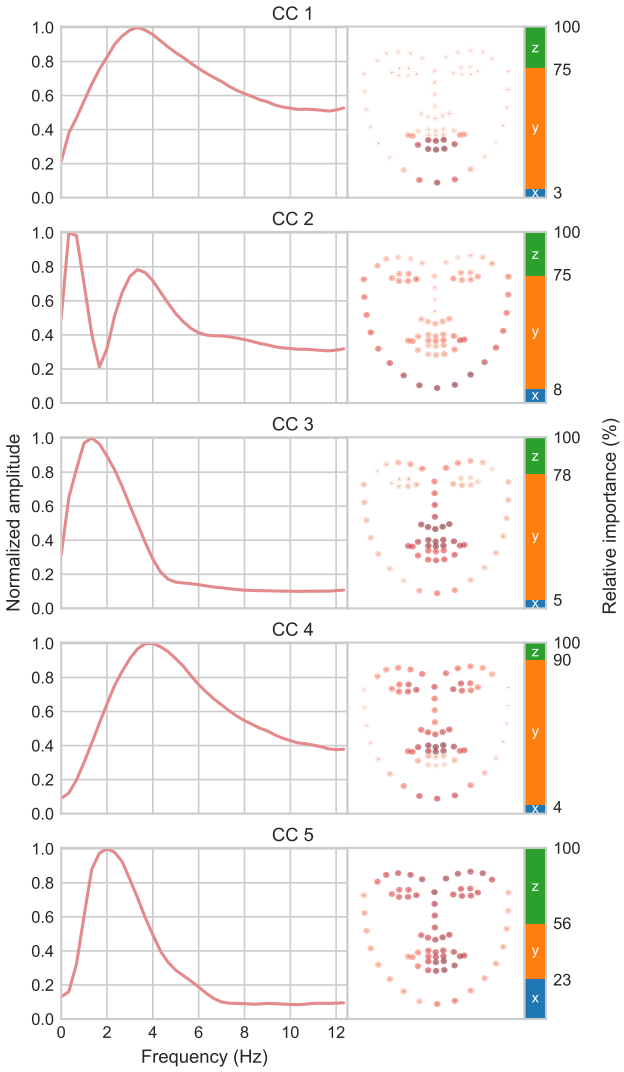


Figure 2.2: **CCA results for the LRS3 dataset.** *Left*: CCA-derived temporal modulation filters for the first 5 significant canonical components (CCs). *Right*: corresponding facial landmark loadings. Darker red indicates higher weights. The 3D landmarks are shown in 2D projection, and the colorbar indicates the relative contribution of the x (blue), y (orange), and z (green) directions.

The first canonical component, CC1 represents the largest correlation between the AV features. As can be seen in Fig 2.2, CC1 extracts motion of the lower lip and jaw, mainly in the vertical direction, which is correlated with speech modulations at rates peaking around 3-4 Hz. CC4 complements CC1 by

extracting envelope information in a similar envelope frequency range with a peak around 4 Hz but correlated with vertical movement of the upper lip and upper parts of the head. Together, CC1 and CC4 represent a modulation transfer function for the envelope that aligns with the average modulation spectrum for natural speech, with a peak around 3-4 Hz (Ding et al., 2017; Varnet et al., 2017). Our analysis indicates that this 4 Hz peak is statistically correlated with two main sources of visual face motion centered at the lower and the upper parts of the mouth. The first (CC1) relates to mandibular motion that can be performed relatively independently of other head movements. The second (CC4) relates to maxillary movements that are naturally coupled with pitch axis rotations of the head relative to the mandible. These two components thus appear to capture two main kinematic dimensions of mouth open-close cycles during speech production.

The envelope filters associated with mouth openings (CCs 1 and 4) are relatively broadly distributed around 4 Hz. This may partly reflect differences in speaking rate across talkers (Jacewicz et al., 2009; Pellegrino et al., 2011). To investigate this, we computed the spectral peaks of the envelope CCs separately for each video (and thus for each speaker) in the dataset (see S1 Fig). The distribution indeed matches the tuning width of the filters learned by CCA, indicating an influence of individual differences in speaking rate.

Whereas CC1 and CC4 capture mouth openings correlated with envelope modulations distributed around 3-4 Hz, CCs 2, 3, and 5 capture slower modulations around 1-2 Hz correlated with more global head and face movements. CC3 specifically extracts pitch axis rotations of the head, whereas CC5 relates to rigid head movements in all spatial directions. The spatial decomposition learned by CCA thus isolates rigid 3D head rotations by a single component (CC5) while removing x and z rotations from the remaining components. While CCs 3 and 5 capture head rotations, loadings on oral landmarks are also high, in particular for CC3. This indicates that head and mouth movements are mutually correlated and together correlated with slower speech envelope information. This occurs, for instance, when head nods are synchronized with mouth openings to produce accents on important words, thereby yielding envelope fluctuations at a slower rate. CC2 combines envelope information at the two rates in one component.

Together, the visual face and head appear to carry speech envelope information at two distinct timescales during natural speech. Envelope fluctuations

peaking around 4 Hz are specifically associated with mouth openings (CCs 1 and 4), while slower 1-2 Hz modulations are correlated with coordinated motion across the face and head (CCs 2, 3, 5). For illustration, CC1 and CC3, for an example talker are plotted in Fig 2.3. As can be seen, modulations around 4 Hz captured by CC1 track speech at the level corresponding to syllable onsets, while the slower 1-2 Hz modulations of CC3 capture variations at the level of phrases. Local time shifts between face and envelope CCs can occur, as can be seen when inspecting CC loadings for individual speakers. For instance, in the example shown by CC3 in Fig 2.3, a vertical head rotation used to emphasize the final statement ('*do offer*') precedes the acoustic modulation associated with the produced stress.

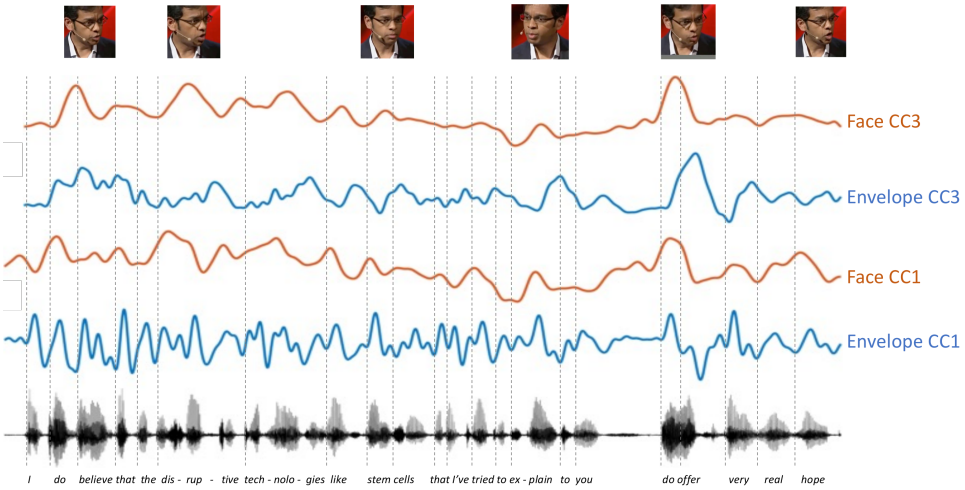


Figure 2.3: **CC1 and CC3 for an example speaker.** CC projections for the speech envelope are shown in blue. CC projections of the facial landmarks are shown in red. Vertical lines indicate word onsets. CC1 represents speech envelope fluctuations corresponding to the onset of individual syllables, while CC3 tracks slower variations corresponding to words or phrases.

Because of the data-driven nature of the analysis, it is important to determine the consistency of the learned AV components. To investigate reliability, we split the dataset into two equal halves and performed the same analysis separately on each split. None of the speakers overlapped between the two halves. The results of the split-half analysis are shown in S2 Fig. As can be seen, the CCA-derived envelope filters and corresponding face loadings are highly similar in the two separate analyses. This indicates that the observed temporal regularities are stable when considering AV speech statistics across many speakers. S3 Fig also illustrates this point by showing MTFs for CCA solutions

computed with a varying number of speakers. With increasing amounts of data, the bandpass filter shapes become increasingly stable, in particular for the most prominent components.

Analysis of the GRID dataset

As a supplemental analysis, we performed the same rCCA analysis on the GRID speech database. Unlike the LRS videos of natural speech in the wild, the GRID corpus consists of videos of a smaller number of speakers (34) instructed to perform simple and syntactically identical monosyllabic sentences (such as 'put red at G9 now') (Cooke et al., 2006). Movements beyond those involved in sound production are thus minimized in this data. The GRID data is comprised of numerous videos from each speaker, whereby the total amount of data included in the GRID analysis was similar to the LRS analysis.

The components learned for the GRID data are shown in Fig 2.4. Again, components with a correlation above 1% that generalize significantly across speakers are shown. As can be seen, CCA again learns envelope filters distributed around 4 Hz. CCs 1, 2, and 5 again capture mouth openings and associated movements of the lower (CC1, CC2) and upper (CC5) parts of the face, highly similar to CCs 1 and 4 found for the LRS data. Unlike the LRS data, however, all components for the GRID data have envelope filter peaks in the 3-5 Hz range and relate more closely to orofacial motion. In addition to the upper and lower part of the mouth, regions around the two lip corners emerge as separate CCs (CC3, CC4, CC7). Slower envelope rates in the 1-2 Hz range related to head motion do not emerge when talkers are not gesturing freely, as in the LRS dataset. Instead, the GRID data highlights several details of the oral motion.

2.5 Discussion

In the current study, we present a CCA technique to learn speech envelope filterings that are correlated with visual face motion. Our analysis relates different rates of acoustic envelope variation to visual motion in different parts of the talking face. The main results for the LRS natural speech dataset indicated two primary temporal ranges of envelope fluctuations related to facial motion across speakers. The first is distributed around 4 Hz and relates to mouth openings. The second range of modulations peaks around 1-2 Hz and relates to more global face and head motion. Envelope information at both rates were

correlated with landmarks distributed across the face, reflecting the fact that natural speech involves highly coordinated motor activity. This also implies that envelope cues are not only available from mouth movements but can be retrieved from non-oral parts of the face and head. Importantly, the derived AV correlations were predictive across speakers implying that these temporal cues are consistent in natural AV speech statistics.

Bandpass envelope MTFs

Our analysis revealed modulation transfer functions with a bandpass character. A number of previous studies have investigated the relation between speech envelopes and facial movement, e.g., by correlating motion data with the low-passed Hilbert envelope of the audio waveform (Chandrasekaran et al., 2009; Kuratate et al., 1999; Munhall and Vatikiotis-Bateson, 1998; Yehia et al., 2002). However, our analysis indicated that envelope information is correlated with visual face motion at specific temporal scales. This echoes the sensitivity of the auditory system to envelope information at different timescales (Poeppel and Assaneo, 2020). In the auditory domain, bandpass-like modulation sensitivity has been modeled as a modulation filterbank, with filters acting as AM detectors at different rates (Dau et al., 1996; Nelson and Carney, 2004). For instance, accurate prediction of speech intelligibility in fluctuating noise maskers has been argued to rely on the signal-to-noise ratio in the envelope domain, e.g., after modulation-frequency selective filtering (Jørgensen and Dau, 2011). While sensitivity to higher modulation frequencies may be unique to audition, slower temporal cues may be processed in a multisensory fashion (Chandrasekaran et al., 2009; Rosenblum, 2008). Our analyses indicate that AV envelope cues are available at two distinct timescales below 10 Hz. These are not simply different low-passed versions of the broadband envelope but bandpass modulation filters in the 1-2 Hz and 3-7 Hz ranges, respectively. Envelope modulations in these two distinct ranges were mutually uncorrelated (as a property of CCA) and thus appear to capture unique sources of AV correlation.

Two rates of AV regularity

These two rates of speech modulations correspond well to the rates at which syllables (3-4 Hz) and phrases or prosodic features (1-2 Hz) are produced in natural spoken language (Goswami and Leong, 2013; Inbar et al., 2020). The

onsets of individual syllables are pronounced energy transitions in a speech signal, as reflected by the fact that the average modulation spectrum is dominated by energy around 4 Hz (Greenberg et al., 2003). Acoustic cues for segmenting a continuous speech signal into phrases are less prominent in the envelope spectrum, where energy falls off below 3 Hz. However, when considering speech as an audiovisual signal (rather than a purely acoustic one), slower envelope rhythms in the 1-2 Hz range emerge as a distinct range of temporal regularity. AV correspondences at these two different timescales may thus provide cues for segmenting the continuous speech signals at the level of syllables and phrases.

This might also indicate a motor origin of temporal regularities at these two distinct timescales. Rhythmic head or limb movements performed during speech are typically slower than mouth movements involved in syllable production. Head nodding or hand gestures during speech have been reported to be synchronized with envelope or pitch variations below 2 Hz (Krahmer and Swerts, 2007; Munhall et al., 2004; Pouw et al., 2020a,b), consistent with our analysis. Mouth open-close cycles during speech, on the other hand, matches the natural syllable production rate around 4 Hz (Chandrasekaran et al., 2009; Ohala, 1975). Different temporal regularities imposed by these oral and non-oral motor components may emerge in facial communication before language and persist in speech. It has been proposed that the use of faster mouth movements to produce acoustic modulations at the syllable rate may be a unique adaptation in humans (MacNeilage, 1998). MacNeilage (1998) proposed that the motor capacity for rhythmic orofacial control in speech may have evolved via slower ingestion-related mandibular cycles. Macaque monkeys can produce rhythmic vocalizations in the 3-4 Hz range (i.e., vocalizations with modulation spectra similar to speech) accompanied by a single facial movement trajectory, rather than by synchronized open-close cycles of the mouth (Ghazanfar and Takahashi, 2014a). Faster cyclic movements of the jaw, lips, and tongue in the 3-7 Hz range are used in non-vocal visuofacial communication (lip smacking, teeth chattering) in non-human primates (Ghazanfar et al., 2012), and may have been adapted for vocal behavior in humans (Brown et al., 2021; Ghazanfar and Takahashi, 2014a; Risueno-Segovia and Hage, 2020). A parallel transition between two rates of vocal production can be observed in human speech development. In the first year of life, infants begin to produce rhythmic babblings (repeated consonant-vowel-like sequences like 'bababa') synchronized with mouth open-close cycles that are below 3 Hz (Dolata et al., 2008) and coordinated with

rhythmic limb movements (Ejiri and Masataka, 1999, 2001; Esteve-Gibert and Prieto, 2014; Iverson and Fagan, 2004; Iverson and Thelen, 1999). From slower and more variable vocal rhythms in infancy, faster and more regular envelope-mouth synchronization above 4 Hz as in adult speech emerge gradually during development (Smith and Zelaznik, 2004; Walsh and Smith, 2002).

Thus, slower vocalizations coordinated with limb movement can be viewed as a precursor to faster vocalizations synchronized with mouth openings at the syllable rate (Ghazanfar and Takahashi, 2014b; Iverson and Thelen, 1999). However, speech modulations at the syllable rate do not necessarily replace slower modulations but may be superimposed on them. Our analysis points to the co-existence of two unique sources of AV correlation, e.g., slower (1-2 Hz) rates of speech modulations synchronized with head and face movement co-exist with faster mouth-envelope synchronization.

The two distinct rates of AV correlation only emerged when considering natural speech across many speakers. Analysis of the GRID data highlighted the well-known synchronized mouth-envelope modulations in the 4 Hz range (Chandrasekaran et al., 2009). Only the analysis across many speakers in the LRS dataset revealed the slower timescale to be a consistent source of AV correlation in natural speech. These differences between datasets suggest an interesting predisposition in AV speech studies. Controlled speech production, as in the GRID matrix sentences, strips away important gestural features that are prominent in natural speech. Speech can be produced with minimal gestural movement (Butterworth and Hadar, 1989), but gestures consistently accompany natural speech (McNeill, 1992). Gestures occur even in conversations between blind people (Iverson and Goldin-Meadow, 1998), suggesting a nonincidental association. Analysis of the GRID data confirmed the prominence of speech modulations distributed around 4 Hz (Ding et al., 2017; Singh and Theunissen, 2003) correlated with mouth open-close cycles (Chandrasekaran et al., 2009), but the analysis does not fully reflect the prominence of envelope information below the syllable rate. It also does not fully capture the degree to which envelope information is consistently correlated with motion in many different parts of the face. Different data splits within each dataset yielded highly consistent CCA components (S2 Fig), indicating that differences between the two datasets stem from differences in the nature of the data. Different speech materials based on different speech tasks thus appear to implicitly zoom in on particular features of AV speech.

AV decomposition of the speaking face

While slower modulations were not found in the analysis of the GRID data, the GRID data revealed a number of more detailed orofacial components. Decomposition of the face during speech has been pursued in previous work using PCA (Kuratate et al., 1999; Ramsay et al., 1996), ICA (Müller et al., 2005) or other matrix factorization algorithms (Lucero et al., 2005; Lucero and Munhall, 2008). Lucero et al. (2008) identified independent kinematic components for the upper and lower parts of the mouth and the two mouth corners that were also identified in our CCA analysis of the GRID data (Lucero and Munhall, 2008). In contrast to previous work, our CCA performs a joint dimensionality reduction in the visual and auditory domain to identify facial regions that are correlated with envelope information. The GRID analysis indicated that the different local kinematic regions of the mouth (upper lip, lower lip, left and right corners), also found in visual-only face decompositions (Lucero and Munhall, 2008), correlate with envelope information in the 3-7 Hz range. The independent kinematics of lip corners could potentially relate to grimacing unrelated to acoustic information (e.g., smiling), but this does not appear to be the case. Other spatially local components, such as the eyes or eyebrows that appear as independent components in visual-only decompositions of the face (Lucero and Munhall, 2008), were not identified as isolated components in our AV analysis. However, a number of components showed high loadings on landmarks around the eyes and upper parts of the face in combination with oral ones. This suggests that e.g., raising of eyebrows at prosodic events (Graf et al., 2002) is consistently coupled with movement in other parts of the face. While non-oral facial parts, such as the eyebrows, may display independent kinematics (Lucero and Munhall, 2008), only movements that are coordinated across the face are consistently correlated with envelope information in our analysis. This high redundancy also implies that similar envelope information is available from many parts of the face.

Neural sensitivity

We note that the two distinct modulation frequency regions emerging from our AV analysis align noticeably with the modulation sensitivity of auditory cortex. Human auditory cortical activity is known to track envelope fluctuations at distinct rates below 10 Hz in speech or other natural stimuli (Ding and Simon,

2014). Speech envelope tracking occurs specifically in the theta (4-8 Hz) and delta (1-3 Hz) frequency bands of the human electroencephalogram (Ding et al., 2016; Doelling et al., 2014; Keitel et al., 2018; Rimmele et al., 2021), and synchronization of cortical activity in these bands have been proposed as a neural mechanism for parsing speech at the level of syllables and phrases (Giraud and Poeppel, 2012). Yet, the fact that these same modulation frequency ranges emerge from AV signal statistics could suggest that temporal modulation tuning in the auditory cortex is adapted to the statistics of natural AV stimuli. The auditory cortex is known to integrate correlated visual signals (Luo et al., 2010; Schroeder and Foxe, 2005), and AV correlations at different timescales may have shaped band-pass modulation selectivity in the auditory cortex, persisting with auditory-only inputs. Rather than a language-specific mechanism for tracking syllables and phrases, cortical envelope tracking specifically in the delta and theta ranges may thus reflect a cortical envelope tuning adapted to temporal regularities that are ultimately determined by auditory-motor constraints.

Perceptual relevance

Our analyses suggest the availability of temporal cues at distinct rates from different parts of the face, but not how these are used in perception. It is well known that viewing a talker's mouth aids auditory speech perception (Bernstein et al., 2004; Sumby and Pollack, 1954). Degrading visual temporal cues, e.g., by reducing the frame rate in videos of the speaker's face, reduces the AV perception benefit (Vitkovitch and Barber, 1996). Non-oral facial movements also contribute to AV perception, as seen by the fact that AV perception benefits occur when the mouth is visually occluded (Thomas and Jordan, 2004). Seeing head motion can improve speech intelligibility (Munhall et al., 2004) and has been argued to provide prosodic speech cues (Guaïtella et al., 2009; Hadar et al., 1984; Hadar et al., 1983; Kim et al., 2014; McClave, 1998; McNeill, 1992). This is consistent with our analyses indicating an association between slower envelope information and head movement. While envelope information distributed around 4 Hz was closely related to mouth openings, these components were also correlated with non-oral facial landmarks. This also implies that envelope information at both timescales is available when only seeing parts of the face. Temporal modulations at these rates are particularly important for speech intelligibility (Elliott and Theunissen, 2009), making coordinated movements across the face a useful perceptual cue. Being distributed across the

face, temporal modulations are likely not perceived via the motion of individual speech articulators but as motion patterns of coordinated facial components. Johnston et al. (2021) recently reported that subjects were highly sensitive to the degree of synchronicity between mouth and eyebrow motion, suggesting that coordinated motion across the face facilitates perceptual binding.

Modelling AV speech across speakers

In contrast to much earlier work, our analysis takes a between-speaker approach to AV speech. The CCA regularization scheme was designed to extract AV statistics that are predictive across many speakers. Much finer details of face-speech correlation can be observed at the individual level, but speaker-specific analyses do not reveal which AV patterns generalize across talkers. Ginosar et al. (2019) recently proposed a deep neural network model that predicts hand gestures of an individual speaker from speech audio of that speaker. Models were trained on large amounts of data from a few speakers in order to synthesize the gestural styles of the individual speakers convincingly. In contrast, we focused our analysis on little data from a large number of speakers in order to identify AV speech-face correlations that generalize across speakers. The person-specific approach of Ginosar et al. (2019) and others was motivated by the argument that speech gesture is essentially idiosyncratic (McNeill, 1992), and that different speakers use 'different styles of motion' (Ginosar et al., 2019). While speaker-specific models may indeed capture most variance in speech gesture data, our between-speaker approach demonstrates that some aspects of AV gesture are also predictive across talkers. It is perhaps unsurprising that mouth movements directly associated with speech production generalize across talkers, but also AV components related to more global gestural head movements appear to generalize. Although gestures like hand or head movement may have acoustic consequences (Pouw et al., 2020b), speech can be produced with limited gestural movement (Butterworth and Hadar, 1989; McClave, 1998), and their consistency across speakers must be established empirically.

Applications

Previous work has used CCA for audiovisual applications, such as speech separation (Sigg et al., 2007), audiovisual synchronization (Sargin et al., 2007; Slaney and Covell, 2001), or facial animation (Mariooryad and Busso, 2012). In such

applications, feature extraction is typically performed to optimize the performance of the particular application. Here, we focused on learning generalizable features that are informative about automatic AV speech, but relevant applications can also be highlighted. Our approach regularizes the CCA across speakers to identify features that are consistently correlated across talkers, making the approach attractive for AV speaker identification. For instance, our approach can be used to identify which of N separated audio sources (e.g., from an acoustic source separation system) belongs to which talking face in multi-talker video data (see S4 Fig). CCA is a linear technique, and the feature transforms are fast to compute, making them appealing for real-time applications.

Limitations

Some limitations in the current approach must also be highlighted. First, our analysis does not account explicitly for time lags between the audio and video. The degree to which audio might lag visual speech is debated (Chandrasekaran et al., 2009; Schwartz and Savariaux, 2014). Speech gestures such as head nods do not have to occur simultaneously with the speech (Butterworth and Hadar, 1989), and time lags may vary between speakers (Kim et al., 2014). This individual variation is explicitly ignored in our between-speaker approach. CCA can readily be extended to account for time-lags (Cheveigné et al., 2018), but will likely require more finely sampled video signals. However, a narrowly spaced envelope filterbank covering low modulation frequencies is likely to be able to absorb time shifts between the signals (Cheveigné et al., 2018), at least within the temporal range normally considered to be relevant for AV integration (Stevenson and Wallace, 2013).

Speech datasets like the LRS3 enable large-scale studies of AV statistics across speakers, but the nature of the data also limits such investigations. The differences between our analysis of natural speech in the LRS dataset and the GRID dataset illustrate the fact that differences in the data influence the results. While the recordings of TED talks in the LRS dataset can be considered natural speech, most natural speech occurs in the form of dialogues or conversations involving turn-taking. Speech rhythms during turn-taking may be adapted to the temporal structure of turn-taking behavior (Hadar et al., 1984; Roberts et al., 2015; Trujillo et al., 2021; Zhang and Ghazanfar, 2020), which may not be captured when analyzing video of monologues. Unfortunately, large video speech datasets involving natural communication are currently missing.

Importantly, CCA is a linear technique, and our approach only considers linear relations between visual and acoustic features. The relation between visible articulators and the produced speech signal is non-linear in important aspects (Scholes et al., 2020; Yehia et al., 2002), and a linear model is therefore principally limited in capturing these. Yehia et al. (2002) found that a non-linear neural network outperformed a linear model in predicting head motion from acoustic features (Kuratate et al., 1999; Yehia et al., 2002). Nonlinearities may, in principle, be accounted for by appropriately transforming the acoustic and visual features. However, here, the main goal was to learn these feature transformations from the AV speech data. The availability of extensive speech datasets and improved techniques for facial landmark estimation may enable data-hungry non-linear models to learn feature transformations from more simple input features. However, this arguably involves a trade-off between model accuracy and interpretability. In our approach, CCA learns a linear combination of linear envelope filters, which is itself an envelope filter. This implies that the components can be investigated directly in the envelope domain, i.e., we can directly investigate which envelope frequencies relate to motion in different parts of the face. The fact that results can be linearly related back to the input space arguably facilitates interpretation.

2.6 Supporting information

S1 Fig. Spectral peak distribution. Distribution of spectral peaks on envelope CCs for the individual speakers in the LRS dataset. The distribution aligns with the width of the CCA-derived modulation filter functions (Fig 2.1), suggesting an influence of individual differences in speaking rate.

S2 Fig. Split-half reliability. The same CCA analysis was performed on two independent halves of the LRS3 dataset (∼1950 speakers in each split). Envelope filters (left panels) and spatial decompositions of the visual face (right panels) learned via CCA were highly similar between the two data splits.

S3 Fig. MTFs for varying number of speakers. MTFs were computed for different amounts of speakers (n=10, 100, 1000, 2000) by subsampling the data. For a given number of speakers, 9 CCA solutions were computed. As can be seen, a higher number of speakers lead to more convergent solutions. Regularisation

parameters were still optimized to predict AV correlation across speakers on the full dataset.

S4 Fig. Speaker identification. The AV CCA model enables fast speaker identification. Here, the CCA model is used to identify which of 2 (solid lines) or 3 (dashed lines) different audio segments correspond to 2 or 3 video segments. The AV pair with the highest correlation on CC1 is chosen as the matching pair. Only videos not used for training the CCA model were used for speaker identification. Identification performance is shown as a function of AV segment duration for the LRS3 (blue) and GRID (orange) data. Shaded regions show ± SEM.

Acknowledgments

JH was supported by the Novo Nordisk Foundation synergy Grant NNF17OC0027872 (UHeal).

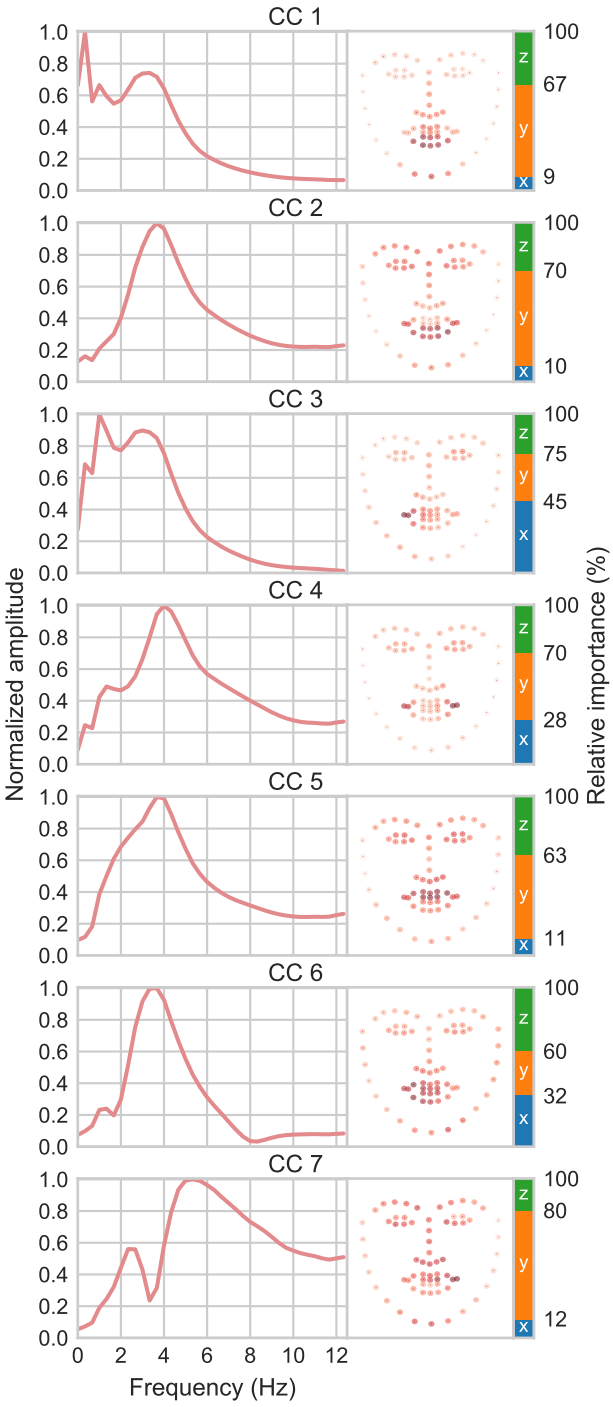


Figure 2.4: **CCA results for the GRID dataset.** CCA-derived envelope filters (*left*) and corresponding face loadings (*right*) for the GRID dataset. Unlike *in the wild* recordings of natural speech such as the LRS3, the GRID corpus is composed of simple, syntactically identical six-word sentences.

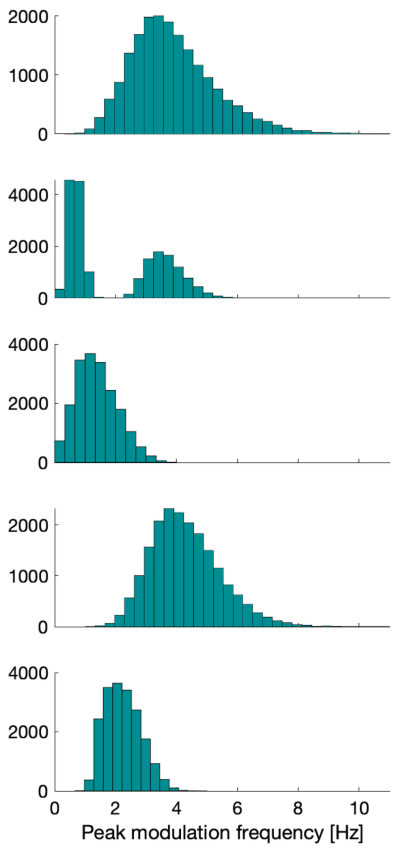


Figure 2.5: **S1 Fig. Spectral peak distribution.** Distribution of spectral peaks on envelope CCs for the individual speakers in the LRS dataset. The distribution matches the width of the CCA-derived modulation filter functions (Fig 2.1), suggesting an influence of individual differences in speaking rate.

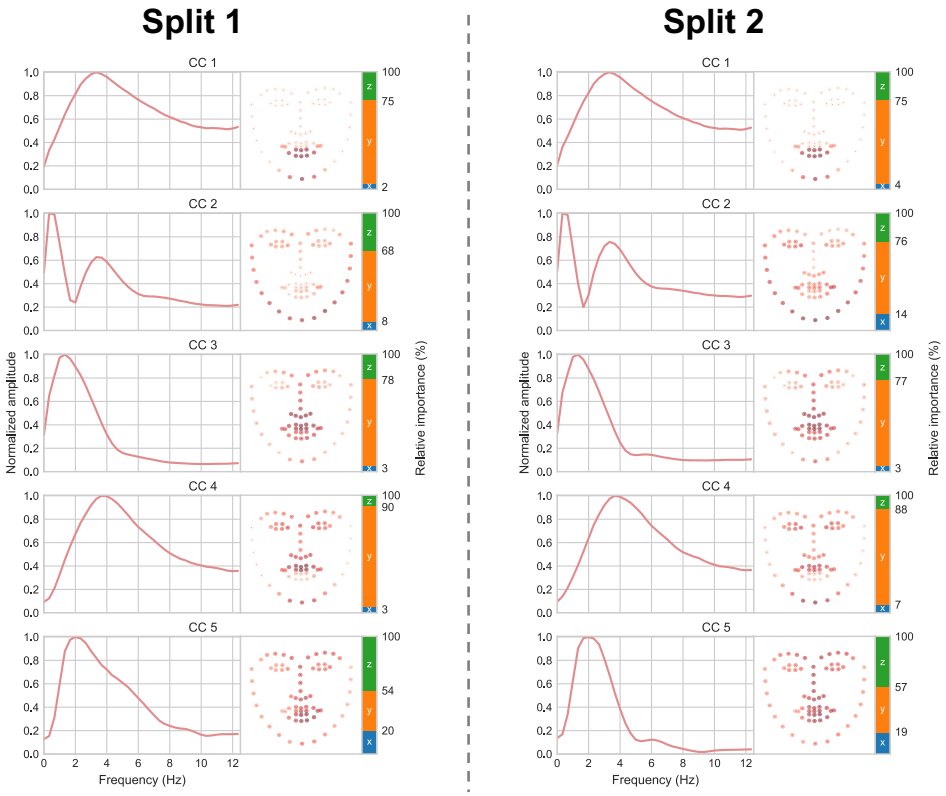


Figure 2.6: **S2 Fig. Split-half reliability.** The same CCA analysis was performed on two independent halves of the LRS3 dataset (∼1950 speakers in each split). Envelope filters (left panels) and spatial decompositions of the visual face (right panels) learned via CCA were highly similar between the two data splits.

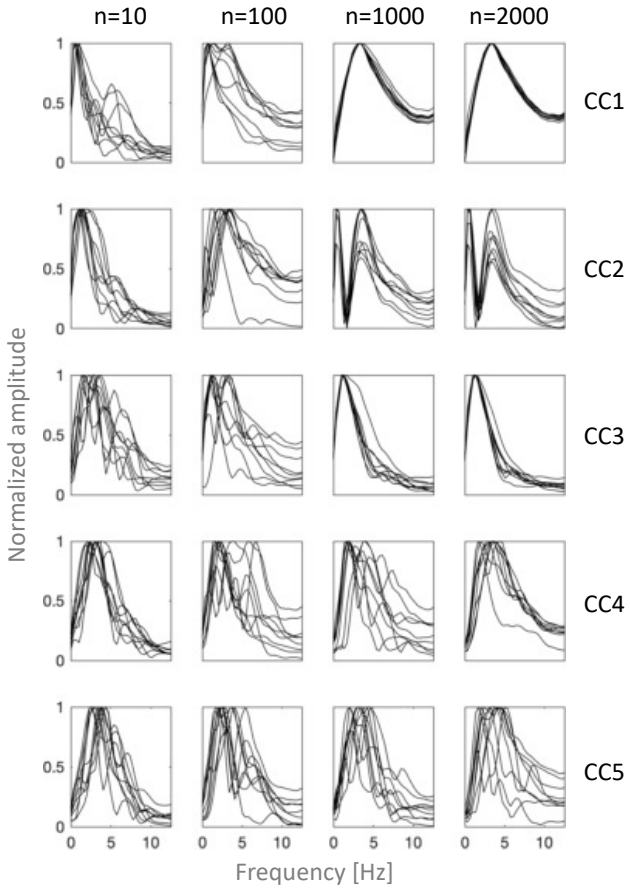


Figure 2.7: **S3 Fig. MTFs for varying number of speakers.** MTFs were computed for different amounts of speakers (n=10, 100, 1000, 2000) by subsampling the data. For a given number of speakers, 9 CCA solutions were computed. As can be seen, a higher number of speakers lead to more convergent solutions. Regularisation parameters were still optimized to predict AV correlation across speakers on the full dataset.

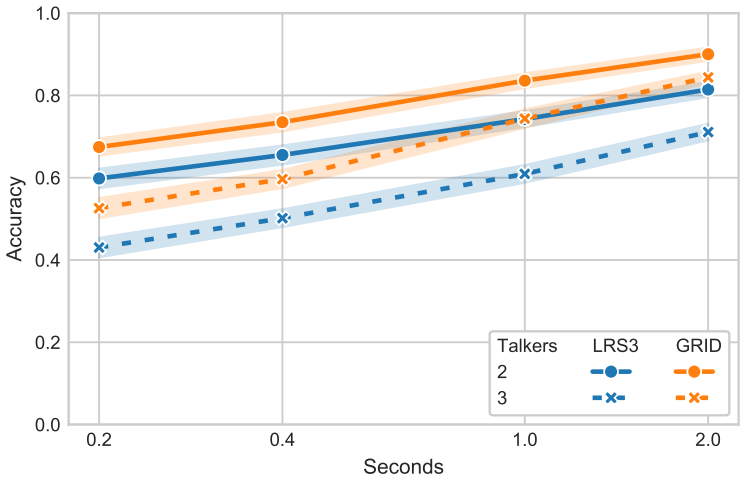


Figure 2.8: **S4 Fig. Speaker identification.** The AV CCA model enables fast speaker identification. Here, the CCA model is used to identify which of 2 (solid lines) or 3 (dashed lines) different audio segments correspond to 2 or 3 video segments. The AV pair with the highest correlation on CC1 is chosen as the matching pair. Only videos not used for training the CCA model were used for speaker identification. Identification performance is shown as a function of AV segment duration for the LRS3 (blue) and GRID (orange) data. Shaded regions show ± SEM.

3

Self-Supervised Learning of Correlated Audiovisual Features[a]

Abstract

When producing speech, correlated audiovisual (AV) signals are generated. Generally, studies concerned with AV speech rely on AV features to investigate AV correspondences or build applications such as AV speech separation systems. While the AV features tend to reflect prior knowledge about the individual modalities, they are not guaranteed to capture the shared information between the two modalities. This study proposes a self-supervised learning approach to train interpretable AV-based convolutional neural networks (CNNs) directly on raw audio and video inputs. Using a novel correlation scheme, CNNs are trained on matching and mismatching AV segments to learn AV features that are correlated when the AV segments match. We compared AV features and first- and second-layer audio filters learned by two CNNs trained on natural AV speech video. One of the CNNs relied on standard one-dimensional (1D) convolutions, whereas sinc-based convolutions, specifically designed to learn bandpass filters, were used to ease interpretation of audio-filters in the other network. Evaluated on a test dataset, both models achieved almost 100 % accuracy in a three-speaker identification task, while the average correlation between the learned AV features was found to be 70 % for matching AV segments and approximately 0 % for mismatching AV segments. Moreover, we demonstrated how the AV features could be backtracked to the input space revealing the attentional focus of the

[a] This chapter is based on Pedersen, N. F., Dau, T., and Hjortkjær, J. "*Self-Supervised Learning of Correlated Audiovisual Features*" (in prep).

CNNs in both the auditory and visual domains. In the visual domain, both models primarily learned to focus on mouth movements during speech, while they focused on extraoral face movements during periods of silence in the speech. In the audio domain, the audio features seemed to capture fluctuations in the audio envelope. The presented method has multiple compelling properties that would make it useful for both analyses of AV speech and for extraction of AV features that can be used in downstream tasks such as AV-based speech separation models or speech recognition models.

3.1 Introduction

Speech perception is fundamentally multisensory. Producing speech generates temporally aligned visual and auditory signals, and their co-occurrence is an essential cue for binding them together in perception (Johnston et al., 2021). Temporal synchronicity of audio and video is also a useful feature for learning AV feature representations. During speech, it is well known that mouth movements are linearly correlated with slow amplitude fluctuations in the speech signal (<10 Hz). Yet, the statistics of AV speech are likely much richer and more complex than what linear correlation statistics capture (Scholes et al., 2020; Yehia et al., 2002).

In recent years, deep neural networks have been successful in learning latent AV representations from co-occurrence statistics (Afouras et al., 2018b; Owens and Efros, 2018). The temporal correspondence between audio and video can be used efficiently as an objective function to guide learning. Specifically, networks can be trained in a self-supervised manner to detect temporal misalignment between audio and video (Afouras et al., 2018b; Owens and Efros, 2018). Discriminating aligned and misaligned (synthetically shifted) audio and video does not require labeled data, and network training can harness the abundance of video data. Networks that learn robust AV representations in a self-supervised manner can, in turn, be used for downstream tasks such as speech separation, speech enhancement, and speaker identification (Afouras et al., 2018b; Ephrat et al., 2018; Nagrani et al., 2020; Owens and Efros, 2018).

In this work, we present a self-supervised neural network framework to learn AV representations directly from raw video pixel and audio waveform inputs. Raw inputs allow the network to learn any type of correspondence between

the audio and video signals without prior assumptions about their relation. The trained network can then, in turn, be inspected to investigate the signal transformations that lead to shared AV representations. However, when both inputs are given in their raw form, the networks also face a tremendously challenging task. They need to account for the enormous variability of the input data, e.g., the large variability in pose and illumination in the visual domain and the variability in speaking rate, background noise, pronunciation, etc. in the audio domain. The network must learn latent invariant representations given this huge amount of variability at the input level. This is in many ways similar to the self-supervised learning process during human speech learning, where infants at an early stage must learn to combine temporally correlated signals across modalities (Dupoux, 2018). To form common representations, humans and neural networks are thought to exploit the compositional hierarchies of natural signals by extracting multiple levels of representations with increasing complexity (LeCun et al., 2015). Hence, it is intriguing to analyze these representations learned by neural networks and compare those with prior knowledge reflecting the properties of the human auditory and visual system.

Advances in AV deep learning have primarily been driven by their usefulness in applications, e.g., for speech separation. Less energy has been devoted to in-depth analyses of the interrelationship between the visual and the audio signals and the feature extraction learned by the networks. Increasing the interpretability of these networks is of great importance if we want to better understand how the networks function and how they might generate signal transformations that are informative. In recent years, an expanding body of work, aiming at deciphering the DNNs and yielding higher interpretability, has been pursued (Montavon et al., 2018). These analyses have shown that neural networks learn to exploit the compositional hierarchies of natural signals, in which higher-level features are obtained by composing lower-level ones. Analysis of audio networks has, for example, shown that features like phones and phonemes are extracted in early layers, whereas later stage layers capture features of, higher abstraction level, like words and sentences (LeCun et al., 2015). Other approaches, such as Class Activation Mapping (CAM), have been used to visualize what visual neural networks learn to attend (Zhou et al., 2016). As a means to increase the interpretability of the individual filters, Ravanelli and Bengio (2018a,b) presented SincNet. SincNet filters are bandpass filters parametrized by only two parameters. SincNets have been shown to yield interpretable frequency trans-

formations in the first layer of audio-only networks. Here, we extend this to AV networks with two SincNet layers - an architecture similar to cascaded bandpass filters (separated by a nonlinearity) that have been proposed in auditory models (Ewert and Dau, 2000; McWalter and Dau, 2017). The multilayered SincNet architecture effectively allows the network to learn temporal cascaded envelope representations in the context of AV signals.

Our previous work found that a linear CCA model can learn speech envelope filters and face decompositions based on envelope and facial landmark inputs (Chapter 2). We exploited the fact that a linear combination of FIR filtered speech envelopes (learned by CCA) is itself a filtered speech envelope. The analysis showed that different rates of acoustic envelope information in the 1-7 Hz range are correlated with motion in different facial components. A linear model like CCA with predefined audio and video features allows transparency and interpretability of the learned features but also potentially neglects many details of AV signal statistics. In this work, we therefore instead aim to learn AV representations with non-linear neural networks directly from the raw audio and video inputs. If the learned signal transformations can be interpreted in terms of their spatio-temporal filter characteristics, this may yield a more detailed description of natural AV speech signal statistics. Compared to linear methods, neural networks may also learn more detailed AV features that are useful for applications such as speech separation, as discussed in (Chapter 2).

Here, we use multilayered SincNets to learn AV representations from raw video and audio inputs. The network consists of a video and audio part and an AV fusion part. We implement a joint AV attention mechanism following Harwath et al. (2018) that can be visualized in matchmaps. We use the matchmaps to analyze the filtering and the frequency transformations performed by the audio branch of the network and the spatial attention performed by the visual branch. To train the network, we employ a training strategy similar to the one proposed by Chung et al. (2020), that uses multi-way cross-entropy loss to correctly identify which audio corresponds to the video. Furthermore, we introduce a novel correlation method to maximize the correlation between temporally aligned audio and video segments while minimizing the correlation between mismatching segments.

3.2 Methods

The model used in the work consists of a visual network and an audio network that are fused to learn correlated AV features from raw video and audio signals. The networks are inspired by the self-supervised networks proposed by Harwath et al. (2018). This approach is specifically chosen as it allows for great interpretability, and the learned representations are distributed both spatially and temporally, enabling our models to directly co-localize events within both modalities. However, in contrast to their work, where three different similarity scores was compared and their network was optimized with a ranking-based criterion (Karpathy et al., 2014), we introduce a probability-like correlation measure and, inspired by Chung et al. (2020), use multiway-cross entropy loss to train the model for speaker identification. Both multi-way and pairwise losses enforce high similarity between representations of matching AV segments relative to mismatching AV segments. However, Chung et al. (2020) showed that multi-way losses lead to more stable learning because the networks are presented with more mismatching segments. During training, the model is presented with one video segment and N audio segments, as illustrated in figure 3.1. A video representation is obtained from the visual network, and similarly N audio representation are obtained from the shared audio network. The correlation is then calculated between the video-representation and each of the N audio representation, resulting in N correlation values. As the cross-entropy loss takes probability-like inputs, we convert the correlation values from -1 to 1 into probabilities. Each correlation value is scaled to the range 0 to 1: $((1+\text{corr})/2))^3$. The cube term is added to account for the undesirable consequence that zero correlation is much more preferable than anti-correlations, thus making higher correlation values even more preferable. Lastly, each scaled correlation value is converted to probabilities by dividing them by the sum of all correlation values before using them as input to the multi-way cross-entropy loss (see figure 3.1). The model then has to identify the audio segment that matches the video segment. In other words, we directly use a probability-like correlation measure to train a network to correctly identify which of the N audio segments is matching the input video. Ideally, the correlations between matching segments should be as high as possible while being close to zero for non-matching segments. This would make the representation ideal for downstream tasks like speaker identification, speaker recognition, and speech separation.

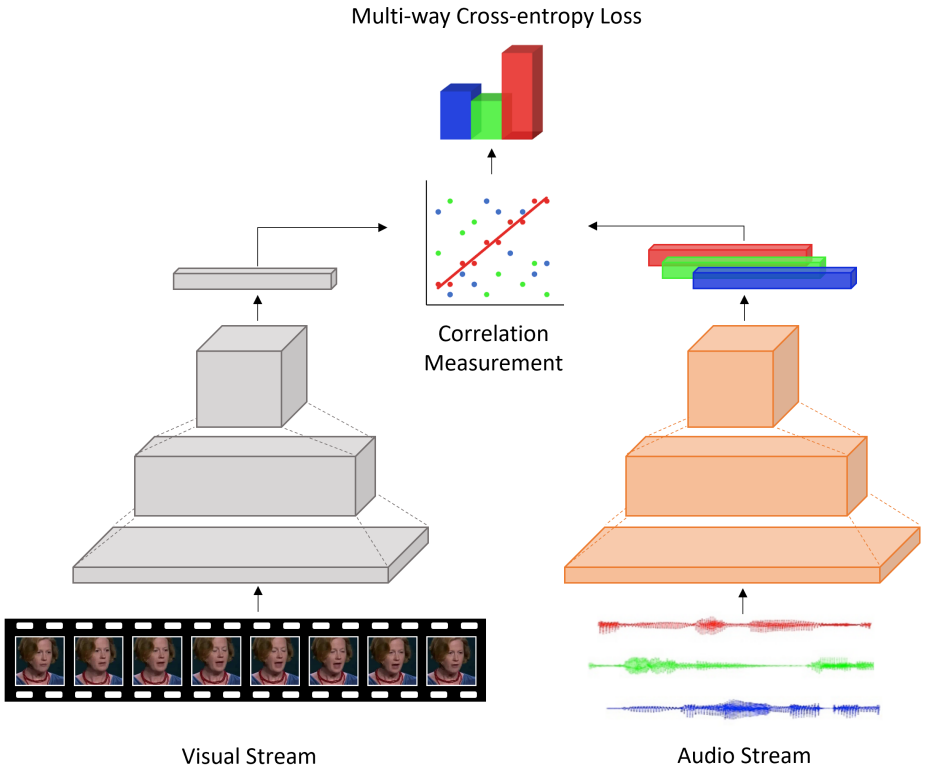


Figure 3.1: Schematic of the self-supervised approach used to train the AV neural networks. Presented with a video segment and three audio segments, the visual branch and the shared audio branch are trained to maximize the correlation between matching AV segments and minimize the correlation between mismatching segments.

We propose and compare two different models. In both cases, the visual network remains the same, whereas we compare two audio networks; a network with sinc-based convolution layers (Ravanelli and Bengio, 2018a,b) in the first two layers and a network based on standard convolutional layers.

3.2.1 Visual network

The visual network branch, (see table 3.1), is a spatio-temporal VGG-like (Chatfield et al., 2014) structure comprised of 3D convolutional blocks. All blocks consist of a 3D convolution with ReLU activation functions, followed by a batch normalization layer. However, in some blocks, a max-pooling layer is also added. The network retains the temporal resolution of the input while reducing the spatial resolution of the input. Similar to Zhou et al. (2016), we apply global average pooling to the output of the last Conv3dBlock as it enables the recovery

of spatial activation maps while providing a good accuracy. The output of the global average pooling layer is finally flattened such that it can be correlated with the feature representations of the audio network.

3.2.2 Audio network

We compare two different audio networks, (see table 3.2). Both networks take as input single channel raw audio and are based on 1D convolutional blocks. The 1D convolutional blocks consist of a 1D convolutional layer followed by a ReLU activation function and a batch normalization. Similar to the visual network, some blocks also contain a max-pooling layer. The two networks differ only in the first two layers. The first audio network, audio-SincNet, contains two successive sinc-based convolution layers (Ravanelli and Bengio, 2018a,b) in the very first layers. The sinc-filters provide for a more straightforward interpretation than standard convolutions, and they converge faster. The second audio network, audio-ConvNet, contains two standard 1D-convolution layers instead of sinc-based convolutions, but the filter lengths are equal to the filter lengths of the audio-SincNet. Following the first two convolutional blocks, both networks are comprised of four 1D convolutional blocks based on standard 1D-convolution layers. The temporal resolution of the input audio is lowered throughout the network to eventually match the temporal resolution of the video input. Lastly, the outputs are flattened, making it possible to calculate the correlation between the audio and video outputs.

Layer	**Input**	**Output**
Conv3DBlock	(b,50,224,224,3)	(b,50,112,112,64)
Conv3DBlock	(b,50,112,112,64)	(b,50,112,112,64)
Conv3DBlock	(b,50,112,112,64)	(b,50,56,56,128)
Conv3DBlock	(b,50,56,56,128)	(b,50,56,56,128)
Conv3DBlock	(b,50,56,56,128)	(b,50,56,56,128)
Conv3DBlock	(b,50,56,56,128)	(b,50,28,28,256)
Conv3DBlock	(b,50,28,28,256)	(b,50,28,28,256)
Conv3DBlock	(b,50,28,28,256)	(b,50,28,28,256)
Conv3DBlock	(b,50,28,28,256)	(b,50,14,14,256)
AveragePooling3D	(b,50,14,14,256)	(b,50,256)
Flatten	(b,50,256)	(b,12800)

Table 3.1: Video encoder.

Layer *SincNet / ConvNet*	**Input**	**Output**
SincBlock2 / Conv1DBlock	(b,32000,1)	(b,32000,64)
mpSincBlock2 / mpConv1DBlock	(b,32000,64)	(b,10666,64)
mpConv1DBlock	(b,10666,64)	(b,1777,128)
mpConv1DBlock	(b,1777,128)	(b,296,128)
Conv1DBlock	(b,296,128)	(b,148,256)
Conv1DBlock	(b,148,256)	(b,50,256)
Flatten	(b,50,256)	(b,12800)

Table 3.2: Audio encoder.

3.3 Experiments

Our goal is to train a deep neural network that, given raw video and audio inputs, can learn to extract maximally correlated representations in cases where the video and audio are temporally aligned and minimally correlated when they are misaligned. This is a desirable property and would make the extracted representation very suitable for downstream tasks like speech separation. Moreover, the network structure should enable transparent interpretations. Specifically, we want to explore which features the networks learn given the task of AV alignment. What does the visual network learn to focus on in the input videos, and which type of frequency selectivity does the network learn in the audio domain?

3.3.1 Dataset

The LRS3 dataset (Afouras et al., 2018a), which contains videos with natural speech extracted from TED and TEDx talks in English, was used to train the models. Overall, 56,430 videos were used out of the 118,516 videos from the predefined `pre-train` dataset. The 56,302 video clips corresponding to approximately 194 hours of video data come from 4,402 different speakers, and the clips vary from two seconds to six minutes in duration. To test the model, we used the predefined `test` dataset, consisting of 1,321 videos. All videos have a frame rate of 25 fps, and each frame has dimensions of (224, 224, 3). The audio is given at a sample rate of 16 kHz.

3.3.2 Training scheme

The networks are trained to identify the correct audio segment out of N presented audio segment. Both models were trained on two-second video and audio segments, corresponding to 50 frames or 32,000 audio samples. To train the models, we use three audio streams: one where the audio is temporally aligned with the video, one where the audio is from the same video clip but temporally misaligned, and one audio segment from a different video. This approach enforces the networks not only to focus on differences in pitch and frequency content (male versus female) but also ensures that the extracted representation contains valuable temporal information. The misaligned audio segments from the same videos were shifted by a minimum of five frames or 0.2 seconds. Shifting the audio by 0.2 seconds ensures that small misalignment's in the original data do not negatively influence the training process. Pixel values in the videos inputs were normalized according to a global mean and variance. The videos were randomly flipped during training. The audio was converted to a mono channel and scaled to the range -1 and 1. To train the network, we used a Stochastic Gradient Descent (SGD), with a momentum of 0.8 and an initial learning rate of 0.001 that was lowered by a factor of 0.1 if no progress was observed on the validation set within three epochs.

3.4 Results

3.4.1 Correlations

Generally, it is desirable if the representations learned by the models are highly correlated for matching AV segments and close to zero for non-matching segments. Indeed, as shown in table 3.3, we found that the representations learned by both models, audio-ConvNet, and audio-SincNet, are highly correlated in cases where the AV segments match while they are close to zero in cases where the AV segments are not aligned. Notably, the correlation values are surprisingly close to zero in the condition where the imposter audio segments come from different speakers. The same tendency is observed in the mismatch condition where the imposter audio segments are temporally shifted but originate from the same talker. The large difference in correlation between the imposter and matching audio segments suggests that the learned representations are appropriate for downstream tasks. With downstream tasks like speaker identification

and speaker separation in mind, it is interesting to focus on the results where the imposter audio segments are extracted from different talkers. It is, however, also relevant to dwell on the performance where the imposter audio segments are temporally shifted. The noticeable correlation difference indicates that the learned representations will perform well in cases where reverberation exist in the audio.

Model	**Match**	**Mismatch: different video**	**Mismatch: same video**
SincNet	68.86 ± 13.00	-0.04 ± 22.08	-5.10 ± 20.38
ConvNet	69.42 ± 14.44	-0.36 ± 21.17	-4.41 ± 20.11

Table 3.3: Comparison of correlation results. For each model the average correlation between AV features is reported for three different scenarios. Match is the scenario where the AV segments are temporally aligned and come from the same video. Mismatch is the scenario where the audio is either temporally misaligned but comes from the same video or simply comes from another video.

Model	**Accuracy: different video**	**Accuracy: same video**
SincNet	98.56 ± 0.67	99.11 ± 0.67
ConvNet	98.39 ± 0.80	99.17 ± 1.00

Table 3.4: Accuracies from a three-speaker identification task for the two models. Presented with a video segment and three audio segments (on temporally aligned and two imposter audios), the correct audio should be identified. In one case, the two imposter audio segments are from different videos. In the other case, they are from the same video and talker but temporally misaligned.

3.4.2 Speaker identification

Table 3.4 summarizes the networks' ability to perform speaker identification in two different conditions. In the first condition, the networks are presented with two audio segments from different videos along with the correct audio segment. In the second condition, the temporally aligned audio segment is presented along with two temporally misaligned audio segments originating from the same video. The table compares the performance of our two different audio networks, audio-ConvNet and audio-SincNet, on both tasks. In both conditions, both models yield close to 100 percent accuracy at identifying the corresponding audio segment. Interestingly, both networks perform a little better in the more difficult condition where all audio segments originate from the same video and same speaker. The speaker identification accuracies are

directly comparable to the results presented in (Chapter 2), where Canonical Correlation Analysis (CCA) was used to identify linear relationships between facial landmarks and modulation filtered audio signals on the LRS3 dataset. For two-second segments, the speaker identification accuracy in three talker scenarios was reported to be approximately 70 %. As anticipated, the speaker identification accuracies reported here are far superior to the results obtained using CCA. Not only do we use richer data inputs that can add valuable and hidden information not captured by the landmarks, but the linear CCA approach is also incapable of capturing the non-linear relationships that exist between visual articulators and the produced speech signal (Scholes et al., 2020; Yehia et al., 2002).

3.4.3 Interpretation of the models: matchmaps

To create matchmaps between the AV segments and hence visualize the spatiotemporal focus of the visual network, we extracted representations from the last convblock in both the audio network and the visual network, (see table 3.2 and 3.1). Thus, given a video input of size (50, 224, 224, 3), the output representation of the visual network, V, is of size (50, 14, 14, 256), meaning that it retains a 14 by 14 spatial feature map across 256 channels for each frame. From the audio networks, the audio representations, A, are of size (50, 256) given an input of size (32000, 1), where 50 corresponds to the number of input frames. The matchmaps, M, are derived by multiplying the representations along the feature dimension: $M = VA^T$, resulting in a matchmap of size (50, 14, 14, 50), where the first 50 is the temporal dimension of the video and the latter 50 is the temporal dimension of the audio. To obtain the visual matchmap, we compute the mean of the temporal audio dimension followed by the absolute value, leaving us with 50 matchmaps, one for each frame. The spatial matchmaps can then be mapped back to the original input and plotted on top of the original frame. Similarly, the audio matchmaps are computed by taking the mean of the temporal and spatial dimensions of the video, leaving us with a one-dimensional vector of length 50 that can be plotted together with the original audio.

Figure 3.2 illustrates the matchmaps in both modalities for both of our models. The SincNet results are highlighted in blue, whereas the ConvNet results are highlighted in orange. Generally, most visual attention maps from other AV speech networks tend to focus more broadly on the entire face (Afouras

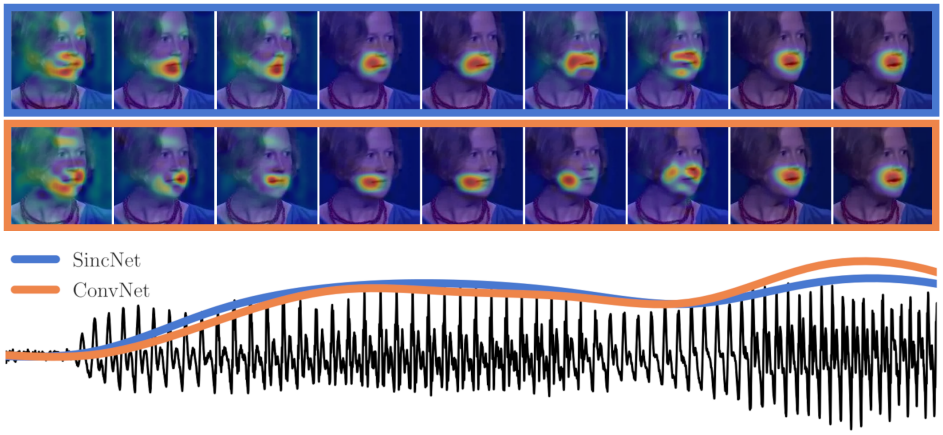


Figure 3.2: AV matchmaps from SincNet (blue) and ConvNet (orange). The two top rows show the focus from the visual networks in a given frame segment. The lower plot shows the corresponding audio along with the focus of the audio networks.

et al., 2020; Cheng et al., 2020; Owens and Efros, 2018; Sharma et al., 2020). In contrast, our models' attention maps are more narrowly focused, thus making them more interpretable. As expected, both of the visual networks mainly attend to the mouth. However, it is worth noting that during silent periods in the audio, the visual networks also focus on other facial regions, like the eyes, the jaw, and in some frames it learns to attend to both the upper and the lower parts of the lips. Also, we found that the networks have learned to focus on the mouth even when speakers are viewed from the side.

Below the visualizations of the visual attention, the figure illustrates the focus of the audio networks. Here we see that the focus of both audio networks resembles the audio envelope. Interestingly, the focus of both networks is strikingly similar, suggesting that the extracted signals must be very prominent. The magnitude spectra of the audio matchmaps are displayed in Figure 3.3. We observe that the two magnitude spectra are close to identical and that they mainly capture information in the 0-5 Hz range with a peak at 2 Hz. Generally, the magnitude spectra correspond well with the findings presented in chapter 2 of this thesis, where we also found frequencies below 5 Hz to be associated with different facial movements. Moreover, we also found larger facial movements to be associated with 2 Hz modulations in the audio.

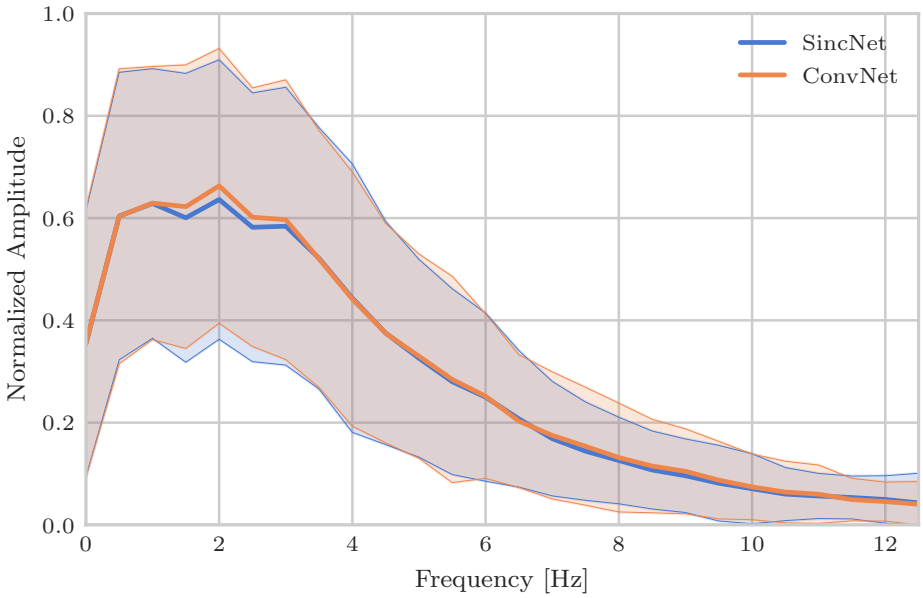


Figure 3.3: Magnitude spectra of the audio matchmaps from both networks.

3.4.4 Analysis of audio network: filters

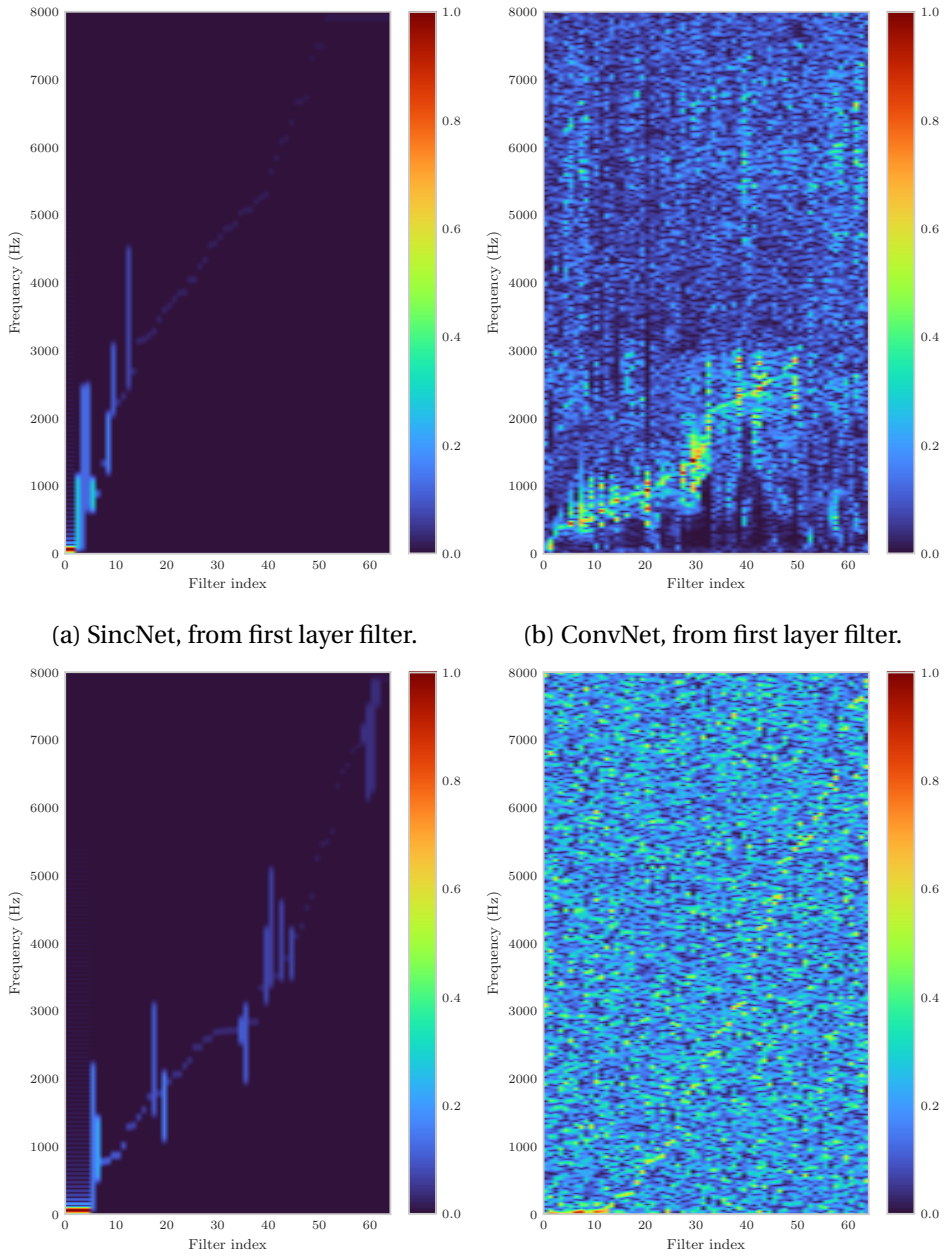
As mentioned earlier, the only difference between the two audio networks, audio-ConvNet and audio-SincNet, is the constrained formulation of the convolution filter in the first two layers. A comparison of sinc filters and standard convolution filters was presented by Ravanelli and Bengio (2018a), where the networks were trained for speaker identification. The setup presented here is in many ways similar, but in contrast to their work, we apply our self-supervised training scheme in an AV context. The visual modality adds an interesting constraint, as the learned audio filters ultimately will depend on the visual input. The magnitude frequency responses for the first layer filters are displayed in figure 3.4a and 3.4b for each network. The plots reveal that both networks seem to focus on frequency content below 3000 Hz. Also, the bandpass nature of the sinc filters makes the filters easier to interpret, in contrast to the standard convolution filters that, in many instances, have multi-band shapes. Additionally, it is worth noting that especially the sinc network learns to focus on some very low-frequency content. Figure 3.5a shows the cumulative frequency response for the first layer filters in each network. Similarly, the individual second layer filters are displayed (see figure 3.4c and 3.4d). Here we observe that both networks mainly focus on low-frequency content. This becomes even clearer when investigating

the cumulative frequency response shown in figure 3.5b. As expected, we see that the networks learn to focus on frequency content below 100 Hz. Recall that a ReLu activation (half-wave rectifier) is used in the first convolutional block. The networks therefore essentially learn to extract filtered signal envelopes in the second layer. This aligns well with the matchmaps, which showed that the network had learned to extract are envelope-like features.

3.5 Discussion

In the current study, we compared and analyzed feature representations from two AV neural networks. The two networks both consist of an audio branch and a visual branch. The visual branch has similar architecture in both networks. The same is true for the two audio networks, audio-ConvNet and audio-SincNet, except for the first two layers. The two networks differ in that standard 1D-convolutions are used in the audio-ConvNet, whereas sinc-based convolutions are used in the audio-SincNet. Further, we introduce a novel correlation measure that can be used directly with the multi-way cross-entropy loss to optimize the models in a self-supervised fashion. Ideally, the fully trained models should learn highly correlated representations in cases where audio and video segments match while being uncorrelated when the segments do not match.

The performance of the two networks was measured through a speaker identification task, where the network was presented with a video segment and three audio segments. Even when all three audio segments were from the same speaker, but two of them were temporally misaligned, both models achieved close to 100 % accuracy. Additionally, we found that in cases where the audio and video segments matched, the average correlation was approximately 70 % in contrast to approximately 0 % in mismatching cases. The idea of maximizing the correlation between representations in a multimodal setting is not new. One often used approach to analyze AV data is Canonical Correlation Analysis (CCA), which uses linear transformations to the input data to maximize the correlation between two views in the latent space. Although the standard CCA provides very interpretable results, it is limited because it only explores linear relationships between the inputs. Chandar et al. (2016) presented a Correlation Neural Network (CorrNet) that to some extent can be thought of as a neural network extension of CCA. Like CCA, the objective function of CorrNet is to maximize the correlation between different views. In contrast to both CCA and CorrNet, our proposed



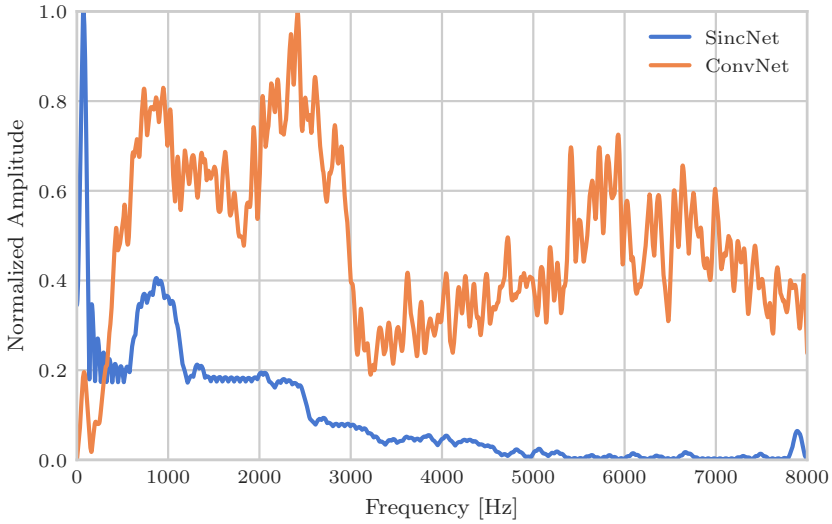
(a) SincNet, from first layer filter.

(b) ConvNet, from first layer filter.

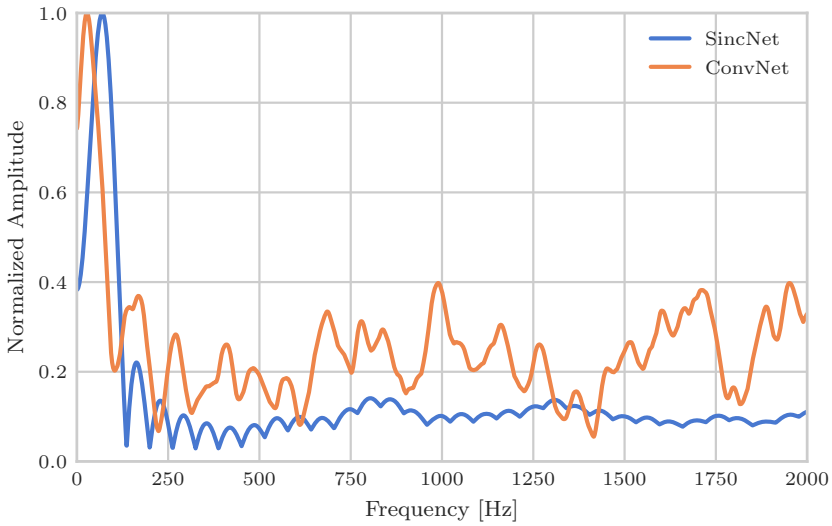
(c) SincNet, from second layer filter.

(d) ConvNet, from second layer filter.

Figure 3.4: Filter transfer functions from the SincNet model and the ConvNet model from the first layer filters, a) and b), and the second layer filters, c) and d).



(a) Summary transfer function of the first layer filters.



(b) Summary transfer function of the second layer filters.

Figure 3.5: Summary transfer functions of the first layer filters (a), and second layer filter (b) from both networks.

probability-like correlation measure more aggressively penalizes lower correlation values and rewards higher correlation values because of the scaling term. Preliminary results suggest that our approach enforces a bigger difference in correlation between matching and mismatching pairs than other approaches. Moreover, since our approach works directly with the multi-way cross-entropy loss, which allows for more stable learning (Chung et al., 2020), it is an attractive alternative for learning correlated representations in multi-modal settings.

The architecture of the networks makes it possible to create matchmaps that allow us to map the output of the networks back to the input space. Unsurprisingly, the matchmaps from both visual networks revealed that they primarily focus on the mouth region, but also that other facial areas such as the eyes and jaw seem to carry relevant information. These findings are consistent with findings in other studies that have shown that movements of the eyes tend to be associated with prosodic events as a way to put emphasis on specific words or when people want to emphasize a specific point (Guaïtella et al., 2009; Kim et al., 2014). Furthermore, many of the same facial areas were identified in our analysis of the temporal modulations in AV speech presented in chapter 2.

Besides the visual focus, we also analyzed the focus of the audio networks. By back projecting the output of the audio-ConvNet and audio-SincNet to the raw speech waveform input, we found that the focus of the networks is remarkably similar to the audio envelope. Magnitude spectra of the audio matchmaps, Figure 3.3, furthermore shows that the audio matchmaps tend to focus on the frequency range from 0-5 Hz with a peak at 2 Hz. These findings align well with the findings presented in chapter 2, where we also observed a strong 2 Hz component associated with larger facial movements, while other components captured facial movements correlated with modulations around 3-4 Hz. Overall the matchmaps align well with previous studies (Alexandrou et al., 2016; Chandrasekaran et al., 2009) that showed that the audio envelope is correlated with mouth movements, but the results also assist the findings presented in chapter 2.

Analyses of the first-layer audio filters reveal that both networks primarily learn to focus on frequency content below 3000 Hz. This frequency range captures most of the first and second formant frequencies of vowels produced by men and women and some third formant frequencies as well (Hillenbrand et al., 1995). We also observe a noticeable peak in the frequency range from 600-1200 Hz in both networks. This frequency range covers first formants of

vowels like /ɛ/, /æ/, /a/, /ɔ/, and /ʌ/, and second formants like /ɔ/, /o/, /u/ (Hillenbrand et al., 1995). However, some differences exist between the filters learned by the two networks. For example, the filters from the audio-SincNet have a dominant peak at 100-200 Hz, thus capturing fundamental frequencies, whereas the same peak is less prominent in audio-ConvNet filters. In many ways, our first-layer filters are in good agreement with the first-layer filters learned by audio-only sinc and convolution networks optimized for speaker identification (Ravanelli and Bengio, 2018a).

Besides training our networks on AV data, we also expand on the work of Ravanelli and Bengio (2018a) by analyzing the filters in the second layer of the audio networks. The second layer filters are surprisingly similar across the two models. In both models, the networks learn to focus on frequency content (well) below 100 Hz. As a ReLu activation function (a half-wave rectifier) is added to the output of the first layer, the low-pass nature of the second layer filters enables the network to perform envelope filtering. As acoustic envelope extraction is believed to play a vital role in combining audio and visual information in humans (Yuan et al., 2020), it intuitively makes sense that the networks also learn to extract envelopes. Furthermore, it is well known that envelope extraction is crucial for providing temporal cues in the auditory system, which the networks most likely also benefit from when performing speaker identification.

The approach presented here not only enables comparable performance in speaker identification scenarios but also facilitates interpretation of the networks. Both models presented here, SincNet and ConvNet, learn to extract visual (mouth, jaw, upper lip) and audio envelope-like representations that resemble information thought to play a crucial role in combing AV sensory inputs in humans. Furthermore, the high correlations between audio and video representations hold promise that the representations would be ideal for other downstream tasks than speaker identification, such as speech separation.

3.6 Conclusion

This study proposed a self-supervised learning approach and a novel correlation scheme to train interpretable AV-based CNNs, optimized to extract correlated AV features from raw audio and video input. Two AV-fusion models were trained, evaluated, and compared. In a three-speaker identification task, both models achieved close to 100 % accuracy, and for both models, the average correlation

between AV features was close to 70 % for matching AV segments, while being close to 0 % for temporally misaligned AV segments. Investigation of audio filters from the two first layers of both models showed that the sinc-based convolutions used in one of the models assisted interpretation, in contrast to the standard 1D convolutions employed in the other model. Notably, the networks' architecture allowed for backtracking of the AV features to the input space, thus allowing for interpretation of the features. We found that both networks primarily learned to focus on mouth movements during speech and on extraoral movements during silence periods. Furthermore, we found that the focus of both audio networks was related to envelope fluctuations in the range from 0-5 Hz. We have presented a framework that allowed for the extraction of highly correlated AV features that could potentially be useful in downstream speech tasks, such as acoustic speech separation. Moreover, the presented approach facilitated the interpretation of both filters and AV features, making it compelling as a tool for analysis of AV speech.

4

Audiovisual Speech Separation with Multisensory Features[a]

Abstract

We present a two-stage approach for training speaker-independent audiovisual (AV) speech separation models to extract target speech streams from single-channel speech mixtures. In contrast to audio-only speech separation models, AV models can utilize the visual speaker information to guide the speech separation process. In this study, we first present three AV-fusion models, all trained to extract visual speaker cues from talkers correlated with audio speech features. Second, visual cues from the different AV-fusion models are used to guide the speech separation process of three speech separation models. We show that, when evaluated on two speaker speech mixtures from unseen video data, our best performing model on average achieves an signal-to-distortion ratio (SDR) of 9.81 as opposed to an SDR of 9.9 achieved by a more computational heavy model on the same dataset. In contrast to many other speech separation systems that only work in non-causal settings, our proposed model performs well in causal settings. The proposed method enables the training of computationally efficient AV speech separation models that work in causal settings, making the approach attractive for real-time and memory-efficient devices.

[a] This chapter is based on Pedersen, N. F., Dau, T., Wen, C., Ceolini, E., and Hjortkjær, J. "*Audiovisual Speech Separation with Multisensory Features*" (in prep).

4.1 Introduction

Speech separation is the task of isolating a target speech stream while attenuating or, ideally, canceling out background interference, such as speech from other talkers or environmental noise. Single-channel separation of speech-on-speech mixtures is a particularly difficult task due to the similarity of the statistics of the constituent speech streams. Recently, AV speech separation methods have shown great promise and outperform audio-only speech separation systems (Ephrat et al., 2018; Michelsanti et al., 2021). While audio-only speaker-independent speech separation models suffer from the "source permutation problem" that arises when the separated speech signals are inconsistently assigned to the sources, AV speech separation models can utilize visual speaker cues to guide the separation process and alleviate the permutation problem. The visual information generally provides a reliable guiding signal, as the talkers tend to be visually separated, and the visual cues are unaffected by noise in the acoustic scene.

In recent years, deep learning-based AV speech separation methods have outperformed the more classical statistical AV speech separation approaches (Michelsanti et al., 2021). However, many deep learning-based AV speech separation models are computationally expensive and employ Bidirectional Long Short-Term Memory (BLSTM) networks, where the output at a given time step is dependent on both past and future observations. While this property is advantageous in non-causal settings, it inherently limits such models from functioning in causal settings such as real-time applications.

Ceolini et al. (2020) introduced a neural network for brain-informed speech separation from single-channel speech. The approach alleviates the need for prior information about the number of speakers, as the attended speech envelope can be decoded from the brain signals, electroencephalography via (EEG), and used to inform the speech separation system about the target speech. The authors showed that their proposed method works in causal settings while requiring less computational power than many similar systems. However, their approach is limited by the quality and stability of the EEG signal and is person-specific due to the variations in individual peoples' EEG signals.

In contrast to the subject *dependent* models needed to estimate audio target envelopes from EEG signals, recent studies investigating AV speech correspondences across many speakers have shown that the face and head movements

are correlated with slow audio envelope information (Chapter 2, Chapter 3). Thus, visual speaker cues, correlated with target audio, can be reliably extracted using a single and speaker-independent AV-fusion model.

Here, we adapt the speaker-dependent brain-informed speech separation method proposed by Ceolini et al. (2020) to train speaker-independent AV speech separation networks. Instead of using EEG signals to estimate the audio envelopes, we use visual features correlated with the target speech to guide the speech separation process of a speech separation network. We first present three different AV-fusion models of different sizes and complexities to extract visual features correlated with audio: (i) a regularized Canonical Correlation Analysis (rCCA) model (Chapter 2), (ii) a neural network extension of CCA based on Correlational Neural Networks (CorrNet) (Chandar et al., 2016), and (iii) AV-SincNet, a self-supervised model optimized for maximizing the correlation between matching AV segments from raw inputs (Chapter 3). To this end, we present three speech separation networks that rely on visual features from the AV-fusion models to extract target speech from single-channel speech mixtures. Furthermore, we compare the performance of the speech separation models in both a causal and non-causal setting to investigate real-time perspectives. We show that our approach allows comparable performance to more complex speech separation models.

4.2 Methods

This section provides an overview of the different AV-fusion strategies and the speech separation model used in this work. We introduce three different AV-fusion strategies, rCCA, CorrNet, and AV-SincNet. As illustrated in figure 4.1, all three AV-fusion models are trained to extract correlated AV features.

Next, we present a speech separation model that uses visual *hints* to extract the target speech from a single-channel speech mixture. A schematic of the speech separation approach is shown in figure 4.2. The network takes two inputs: the complex spectrogram of the speech mixture and the visual hint obtained from a fusion model. The speech separation model's output is a complex-valued mask that is used to extract the target speech from the complex spectrogram mixture.

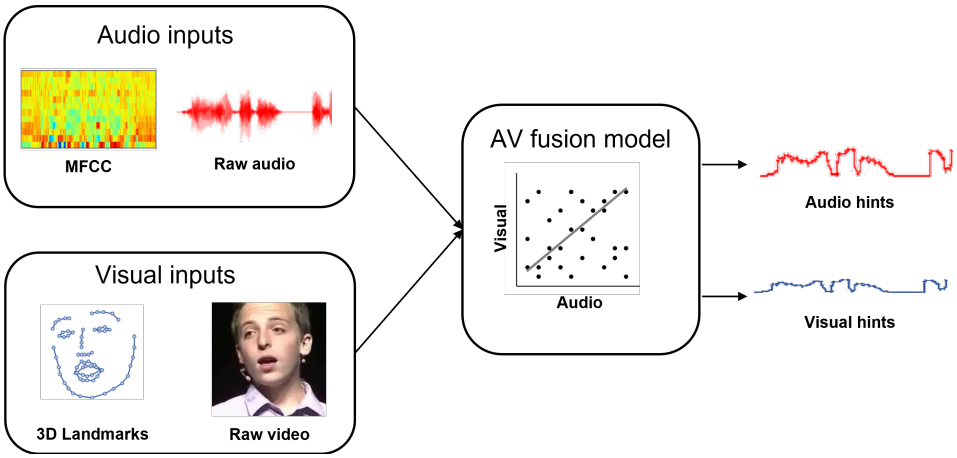


Figure 4.1: Schematic of the general fusion model strategy. Using audio and visual inputs, the fusion models learn to extract correlated AV features.

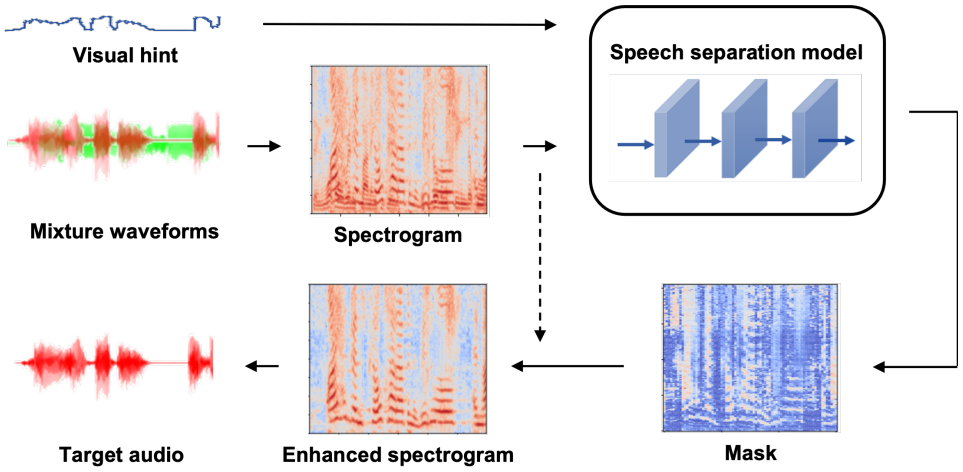


Figure 4.2: Illustration of the speech separation approach. The speech separation approach used in this work, take as inputs a complex spectrogram of the speech mixture and a visual hint from a fusion model. Using information from the visual hint about the target speaker, the speech separation model outputs a complex-valued mask that used to extract the target speech from the complex spectrogram mixture.

4.2.1 Audiovisual Fusion Strategies

Regularized Canonical Correlation Analysis

Presented with audio and video the rCCA-based AV fusion model is trained to learn linear mappings of each modality to a shared space, where they are maximally correlated. The rCCA approach is thoroughly explained in chapter 2, and the reader is hence referred to that chapter for further information. In

this study, we train the rCCA model on 3D-facial landmarks and mel-frequency cepstral coefficients (MFCCs), and then we use the weights of the first five components to extract correlated AV features.

Correlational Neural Networks

Chandar et al. (2016) proposed CorrNet, an approach that uses multimodal autoencoders to maximize correlation among the views in a projected space. Whereas the rCCA approch only explores linear relationships between the inputs, multiple layers with non-linearities can be stacked in CorrNets. Besides maximizing for correlation, CorrNets also include a self-reconstruction loss as well as a cross-reconstruction loss. Let $\mathbf{X} \in \mathbb{R}^{n \times t}$ and $\mathbf{Y} \in \mathbb{R}^{m \times t}$ be two views of some data $\mathbf{Z}$, where $z_i = (x_i, y_i)$. Then $\mathbf{Z}'$ represents the reconstructed data $\mathbf{X}'$ and $\mathbf{Y}'$. The training objective of CorrNet is to find parameters θ that minimize the loss function:

$$C(\theta) = \sum_{i=1}^{T} L(z_i, z_i') + L(z_i, x_i') + L(z_i, y_i')) - \lambda \text{corr}(h(X), h(Y)) \quad (4.1)$$

$$\text{corr(h(X), h(Y))} = \frac{\sum_{i=1}^{T} (h(x_i) - \overline{h(X)})(h(y_i) - \overline{h(Y)})}{\sqrt{\sum_{i=1}^{T} (h(x_i) - \overline{h(X)})^2 \sum_{i=1}^{T} (h(y_i) - \overline{h(Y)})^2}}, \quad (4.2)$$

where h is the hidden representation, $\overline{h(X)}$ and $\overline{h(Y)}$ are the mean of the hidden representations for each view, and λ is a scaling parameter.

The CorrNet used in this study, is trained on 3D-facial landmarks and MFFCs. Both the audio and visual branches of the CorrNet encoder consist of three fully-connected layers. The output from each branch in the encoder is concatenated and a fully connected layer of size 40 is then used to compute the hidden representation. Similarly, the decoder block also consists of three fully connected layers that from the hidden representation tries to reconstruct the audio and visual features. It is the hidden representation extracted from the visual input that is used as a hint of the target speech by the speech separation network. To extract the hidden representations from one modality only, an all-zero-input is used as input to the other modality.

AV-SincNet

Our AV-SincNet is an AV-fusion network consisting of two branches: an audio branch and a visual branch. The network is trained directly on raw audio and visual inputs in a self-supervised manner. The training objective is to maximize the correlation between embeddings from both branches when the AV segments match while minimizing correlation when the segments are misaligned. As the network is described in detail in chapter 3, we will refer readers to that chapter for further information.

4.2.2 Speech Separation

Audio Mixture

The audio mixture, $y(t)$, is obtained by summing the target speaker speech $s_t(t)$ with interfering speech or noise $s_i(t)$,

$$y(t) = s_t(t) + s_i(t) \tag{4.3}$$

where t is the time index. The complex spectrogram $Y \in \mathbb{C}^{F \times L}$ is obtained using the short-time Fourier transform (STFT) of $y(t)$,

$$Y(f,l) = \text{STFT}(y(t)) = S_t(f,l) + S_i(f,l) \tag{4.4}$$

where f is the frequency bin indices and l is the time bin indices. Following (Ephrat et al., 2018), the dynamic range of the complex spectrogram is reduced with a compression factor of 0.3,

$$Y^c = Y^{0.3}, \tag{4.5}$$

where the resulting compressed complex spectrogram $Y^c \in \mathbb{C}^{F \times L}$.

Complex Valued Mask

The approximated target speech is obtained from the complex spectrogram mixture by estimating and applying a complex-valued mask, M. The mask is applied to the complex mixture Y^c using element-wise multiplication,

$$\hat{S}_t^c = M \odot Y^c. \tag{4.6}$$

To obtain the target speech $\hat{s}_t$, the estimated target spectrogram is first decompressed and inverted using the inverse STFT (iSTFT):

$$\hat{S}_t = (\hat{S}_t^c)^3, \tag{4.7}$$

$$\hat{s}_t = \text{iSTFT}(\hat{S}_t). \tag{4.8}$$

Speech Separation: Model Architecture

To train the speech separation network, we use a network architecture similar to the one proposed by Ceolini et al. (2020). The speech separation network consists of three stages: (i) a hint fusion stage that combines the visual hint $H(l)$ at the target speech stream with the complex spectrogram mixture $Y^c(f,l)$; (ii) A stage consisting of stacks of identical and thereby modular dilated convolutional layers, that processes the output of the hint fusion stage; lastly (iii) a stage that applies a complex mask, M, to the compressed complex spectrogram mixture, Y^c, to extract the estimated target speech, $\hat{s}_t$.

In the hint fusion stage, the mixed waveform, $y(t)$, is first transformed into the compressed complex spectrogram, Y^c, using the STFT. To extend the channel number of Y^c of size 2 × F × L, a 1 × 1 2D convolution with K channels is applied, resulting in a tensor of size K × F × L. Likewise, the hint is extended and a 1 × 1 2D convolution is applied to obtain a 3D tensor of size 1 × F × L. Lastly, the two tensors are concatenated along the channel axis to the resulting in a tensor of size (K + 1) × F × L.

Following the hint fusion stage, a series of modular stacks of 2D dilated convolutional layers are applied. The use of dilated convolutional layers is inspired by the relatively recent success of fully convolutional networks for speech separation, (Luo and Mesgarani, 2019). Furthermore, dilated CNNs have proven effective in handling long signals due to a wide receptive field while using considerably fewer parameters compared to recurrent neural networks. Each of the S modular stacks consists of N convolutional blocks that each is made up by a convolution step. Each convolution step consists of three 2D convolutional layers: a 1 × 1 convolution, a 3 × 3 dilated convolution with a dilation factor i, and a 1 × 1 convolution, followed by a batch normalization layer. A ReLu activation is used after each of the first two convolutional layers. Each convolutional block has two inputs, the output of the previous layer and a

skip connection input. Besides the skip connections, the convolutional blocks also contain a residual connection. Each block has two outputs; one is the skip connection input summed with the output of the convolution step, the other is the output of the convolution step summed with the residual connection. Throughout this stage, the network retains the spatial size of the feature input, thus the output tensor is of size (K + 1) × F × L.

Lastly, in the reconstruction stage, we first derive the complex mask *M*. We do that by first reshaping the feature embedding from (K + 1) × F × L to 2 × F × L using a 2D convolutional layer. We then apply a hyperbolic tangent function to map the values in the range [-1,1] and obtain the complex mask. Finally, the target speech can be estimated using eq. (4.6), (4.7), and (4.8).

The three speech separation networks presented in this study all consisted of two stacks, $S = 2$, with six convolutional blocks in each stack, $N = 6$.

SDR and SI-SDR

The signal-to-distortion ratio (SDR) (Vincent et al., 2006) is a commonly used measurement to evaluate the performance of speech separation models, and is calculated as:

$$\text{SDR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2}, \tag{4.9}$$

where s_{target} denotes the true source, and e_{interf}, e_{noise}, and e_{artif} are error terms for interference, noise, and artifacts, respectively. Although SDR is a widely used evaluation metric, situations can arise where the SDR values are artificially inflated due to the way that the noise terms are estimated. Le Roux et al. (2019) introduced a more robust scale-invariant SDR (SI-SDR), where the amplitude scaling dependence of SDR is mitigated, leading to a more stable evaluation metric. The SI-SDR is defined as:

$$\text{SI}-\text{SDR} = 10\log_{10}\frac{||\alpha s||^2}{||\alpha s - \hat{s}||^2}, \tag{4.10}$$

where s is the clean speech signals, $\hat{s}$ estimated speech signals, both with zero mean, and α is a normalization term defined as:

$$\alpha = \frac{\hat{s}^\top s}{||s||^2}. \tag{4.11}$$

4.3 Experiments

The overall aim of our approach is first to train AV-fusion models to learn broadly applicable AV features that can be beneficial in the downstream task of speech separation. In the speech separation system, we can then use the learned visual features as guiding signals to inform the speech separation system about the target audio. We hypothesize that more correlated AV features should translate to better performance in the speech separation system.

Dataset

To train the AV-fusion models and the speech separation models, we used the LRS3 dataset (Afouras et al., 2018a), which contains videos with natural speech extracted from TED and TEDx talks in English. The talker is visible at all times during the video. To train the rCCA model and the CorrNet model, we extracted three-second AV-segments from 74,511 videos corresponding to 62 hours of video data, from the predefined `pre-train` and `trainval` datasets. The AV-SincNet model was trained with approximately 194 hours of video data from the `pre-train` dataset. To test the performance of the fusion models and the speech separation models, we used the predefined `test` dataset, consisting of 1,321 videos. All the videos have a frame rate of 25 fps, and each frame has a dimension of (224, 224, 3). The audio is given at a sample rate of 16 kHz.

4.3.1 AV-fusion models

Audiovisual Features

Both the CorrNet and rCCA fusion models were trained using AV segments of three-second duration. First, the audio was downsampled to 8,000 Hz before computing 40-dimensional MFCCs, extracted every 40 ms using a window length of 64 ms, resulting in a feature dimension of 75×40. 3D-face landmarks similar to the approach presented in chapter 2 were used as visual inputs to the CorrNet and rCCA model. The landmarks were first low-passed filtered at 8 Hz to remove jitter originating from the frame-to-frame estimation of the landmark positions. The visual features of size $75 \times 68 \times 3$ were then flattened on the third dimension resulting in a feature size 75×204.

The AV-SincNet was trained on the raw audio waveforms and video pixels. AV segments of two-second duration were used to train the AV-SincNet.

4.3.2 Speech Separation

For each of the three pretrained AV fusion models, we trained a speech separation model that uses visual representations from the fusion model to perform target speaker extraction.

Audiovisual Features

We use single-channel speech mixtures and AV features from the individual fusion models to train and evaluate the speech separation systems. Two-second segments were used to evaluate the AV-SincNet-based speech separation model, whereas three-second segments were used for rCCA-based and CorrNet-based speech separation models. Hanning windows with a window size of 512 and a step size of 320 are used to compute the input spectrograms for all three models.

Training strategy

Following Ceolini et al. (2020), we employ a curriculum training scheme to train the speech separation model. The idea is that since the learned visual features are only partially correlated with the target audio, the visual features can be regarded as "noisy" audio features. Therefore, we start by training the speech separation model with the "clean" audio features and gradually increase the amount of noise injected into the audio representation. The noisy audio hint is given by:

$$H_{AN} = H_A + \lambda H_N \tag{4.12}$$

where H_{AN} is the noisy hint, H_A is the clean hint, λ is the noise factor, and H_N is the hint noise. Lastly, we train the speech separation models on the visual feature representations.

The noisy hints are computed using two different approaches. The noise added to the audio hints from the rCCA-based model is found by computing the distribution of the residuals between audio and visual features. We then use a zero-centered Gaussian distribution with $\sigma = 0.3$ to approximate the distribution of the residuals. During training, the λ value is increased in steps of 0.1 from 0.05 to 0.55. For both the CorrNet model and AV-SincNet model, the distribution of the residuals cannot be approximated by a Gaussian distribution. Instead, we add a random phase noise to the audio hint in the frequency domain

and transform the hint back to the time domain. In both cases, the injected noise increases in steps of 0.25 from 0.25 to 1.0 before training on the clean visual hint.

To train the model, we used the SI-SDR as the objective function. The networks were trained using the Adam optimizer with an initial learning rate of 0.001. If no improvements were observed for three consecutive epochs on the validation loss, the learning rate was reduced by 10.

4.4 Results

4.4.1 AV-fusion models

	Audio feature	**Visual feature**	**Input length**	**AV correlation**
AV-SincNet	Raw audio	Raw video	2 seconds	0.69
CorrNet	MFCCs	3D-landmarks	3 seconds	0.57
rCCA	MFCCs	3D-landmarks	3 seconds	0.39

Table 4.1: Fusion strategies.

4.4.2 Comparison of AV fusion models

The fusion models can be compared in terms of the averaged Pearson's correlation between audio and video features learned by the models on the test data. The correlation results are compared in table 4.1. The average correlation for the linear rCCA approach was 0.39, whereas the non-linear CorrNet and AV-SincNet approaches obtained average correlations at 0.57 and 0.69, respectively. As expected, we see that the two neural network approaches, CorrNet and SincNet, learned to extract more correlated features than the simpler rCCA approach. Both neural network approaches learn non-linear projections of the audio and video input features, and the higher correlations indicate that linear-only approaches might be limited in capturing important AV correspondences. Furthermore, we see the AV-SincNet approach yields higher correlations between AV features than the CorrNet approach. The likely explanation is that both the 3D-landmarks and the MFCCs (used as input features in the CorrNet) are reduced representations compared to the raw waveform and pixel data and that relevant information may be discarded when computing the features. For instance, MFFCs do not contain phase information.

4.4.3 Speech Separation

Here we present the results of the three different speaker-independent speech separation models. The speech separation results for two-talker mixtures are compared in table 4.2. As can be seen, the best performing model is based on AV-SincNet features which achieved an SI-SDR of 8.98. Speech separation of two talkers based on CorrNet features achieved an almost comparable SI-SDR of 8.09, whereas the rCCA based model achieved a considerably lower SI-SDR of 5.58. The results align well with the fact that the representations learned by AV-SincNet and CorrNet are considerably more correlated than those learned by rCCA. Figure 4.3 shows an example of the speech separation process with AV-SincNet in a non-causal setting. As can be seen from the figure, the model correctly learned to attenuate the unwanted speech while convincingly retaining the target speech. Besides the offline or non-causal setting, we also trained an AV-SincNet model and CorrNet model in a causal setting. The causal setting ensured that the outputs at time step t only depend on the previous time steps, expressed as $P(y_t|x_1, x_2, ...x_{(t-1)})$. The results for the causal setting are also shown in table 4.2, here we again observed that the AV-SincNet model performed better than the CorrNet. Compared to the non-causal setting a performance drop of almost 1 dB was found for the AV-SicNet based speech separation model and approximately 0.6 dB for the CorrNet based speech separation model.

We also tested the speech separation performance in even more complex real-world scenarios. To simulate more complex real-world scenarios, background noise (e.g. aircraft, kindergarten, etc.) was added to the two speaker audio mixture. Similar to the two previous cases, we found a performance drop when changing from the non-causal to the causal setting. In line with the previous results, we found the AV-SincNet based speech separation model to perform best. Lastly, we compared the performance (see table 4.3) of the CorrNet and the AV-SincNet based speech separation models to a model proposed by Ochiai et al. (2019). Similar to our models, it was also trained on the LRS3 dataset. The results in table 4.3 are reported in SDR to make them comparable. The best-performing model, AV-SincNet, achieved comparable performance with 9.81 dB SDR, as opposed to 9.9 dB. This slight performance increase of their model comes at the expense of a much more computational-heavy model that only works in non-causal settings. In contrast, our approach showed comparable performance even with the computational efficient CorrNet fusion approach,

which produces 'lightweight' features. Furthermore, our approach works in both causal and non-causal settings making it more broadly applicable, even in real-time applications.

Setting	Scenario	AV-SincNet	CorrNet	rCCA
Non-causal	2spk	8.98	8.09	5.58
	2spk + noise	8.12	7.64	-
Causal	2spk	8.00	7.52	-
	2spk + noise	7.31	7.18	-

Table 4.2: Average SI-SDR in dB for three models, tested in different scenarios and settings.

Setting	Scenario	AV-SincNet	CorrNet	Ochiai et al. (2019)
Non-causal	2spk	9.81	8.95	9.9
Causal	2spk	8.88	8.56	-

Table 4.3: Comparison between the AV-SincNet model, CorrNet model and a speech separation model proposed by Ochiai et al. (2019). All results are obtained using the LRS3 test set and is reported in average SDR in dB.

4.5 Discussion

In the first part of this study, we trained three different AV-fusion models: rCCA, CorrNet, and AV-SincNet. Presented with audio and video features, the models all learned correlated audio and video representations. Whereas raw audio and video inputs were used for the AV-SincNet model, we used MFFCs and 3D-facial landmarks as inputs to both the CorrNet model and the rCCA model. The performance of the fusion models was evaluated via Pearson's correlation between the learned AV features. The AV-SincNet model achieved the highest average correlation of 0.69 between the extracted AV representations. The average correlation value for the CorrNet model and rCCA model was 0.57 and 0.39, respectively. It is worth noting that while both the CorrNet model and AV-SincNet model use non-linearities, the rCCA model solely relies on learning linear mappings between the AV input features. As some of the interrelationships between the visual articulator and the speech acoustic in key aspects are non-linear (Scholes et al., 2020; Yehia et al., 2002), the rCCA model is inherently limited in capturing these relationships. This limitation is probably the main reason why the rCCA

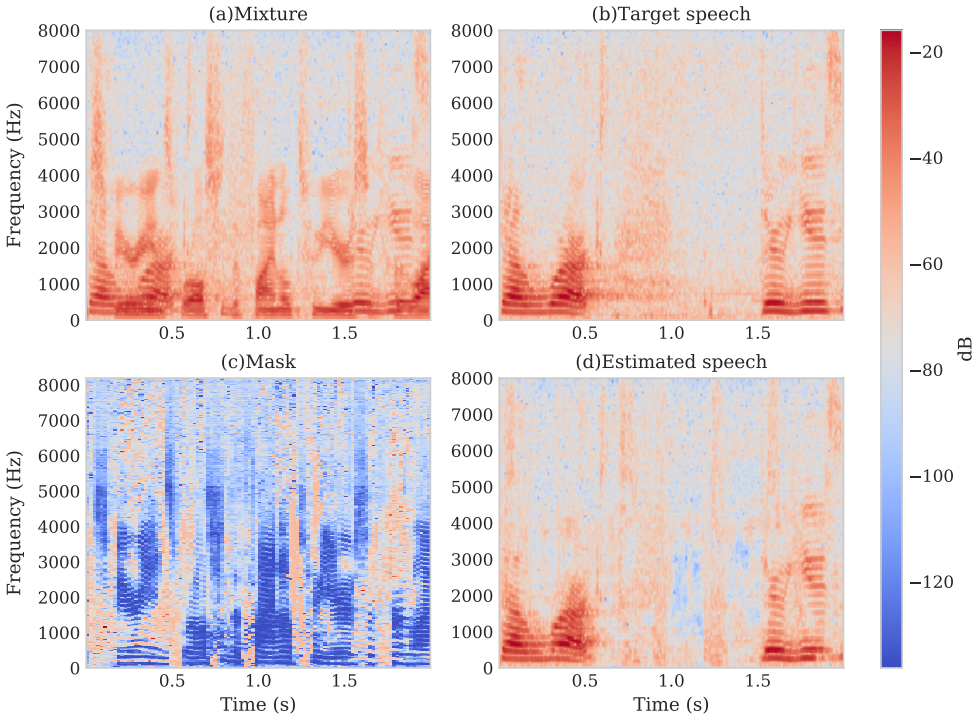


Figure 4.3: Two-speaker speech separation example with AV-SincNet in a non-causal setting. (a) shows an audio mixture of two speakers and (b) shows the target speech. (c) displays the mask that is applied to the mixture to obtain the estimated target speech (d).

model performs worse than the two non-linear methods.

While the AV-SincNet model achieves the highest AV correlation, it is also the most complex model with respect to the number of parameters. Whereas the AV-SincNet fusion model has 3.9 M parameters, the CorrNet has 0.5 M parameters, and the rCCA model has only 1,220 weights as parameters. Moreover, as the AV-SincNet model is trained directly on the raw audio and video input, it can learn to extract features that are directly optimized for maximizing the correlation. The AV-SincNet model can learn to extract phase information, which is disregarded when computing MFCCs and thus not available for the CorrNet and rCCA fusion models. Considering these advantages, it is not surprising that the AV-SincNet model yields the highest correlation values.

Features from the three different fusion models were used to train three AV speech separation models. The DNN-based speech separation system utilizes the visual feature information to extract the target speech from a single-channel audio mixture. Visual embeddings from the three AV fusion models were used to provide visual information to the speech separation systems, as these were

optimized to be correlated with the target audio. Our approach shows that target speech can be extracted from audio mixtures with two speakers and with two speakers and real-world background noise. We also showed that the models can be used in non-causal/offline settings and causal settings with only a minor decrease in performance (0.6-1 SI-SDR in dB). We found that the speech separation model based on AV-SincNet in all test cases performed best, and generally, we observed a strong relationship between speech separation quality, SI-SDR, and how correlated the visual embeddings are with the audio.

In contrast to other approaches with slightly better performing and more computational heavy models (Ephrat et al., 2018; Ochiai et al., 2019), our approach showed comparable performance while performing well in causal settings. Especially, the speech separation model based on the CorrNet fusion model offers good performance while being computationally efficient, making it a compelling option for use in both low consumption devices and real-time applications. Compared to the original idea of using EEG signals as the guiding signal to do brain-informed speech enhancement Ceolini et al. (2020), our AV strategy allows for better performance while additionally being speaker-independent.

4.6 Conclusion

In this study, we proposed a two-stage approach to train AV speech separation models, where the AV fusion stage and the speech separation stage can be optimized independently of each other.

Three speaker-independent AV-fusion models of different complexity were trained on natural AV speech data to extract correlated AV features. Using visual speaker cues from the AV-fusion models as guiding signals, three speech separation models were trained to extract the target speeches from single-channel audio mixtures. We found that the correlation of the visual speaker cues with the target audio was directly related to the performance of the speech separation model. Further, we observed that even if the speech separation models are based on relatively simple AV-fusion models, the performance was still comparable to that of much more complex speech separation models. Importantly, the proposed model performed well in causal or real-time settings, making it an appealing and potentially widely applicable approach.

5

General discussion

This thesis presented different data-driven approaches for identifying correlated audio and visual cues in natural speech signals and investigated the usefulness of these cues in AV speech separation.

First, using linear analysis of AV speech data from thousands of speakers, we identified a set of generic facial movements associated with speech production and the amplitude envelope rates associated with these movements (*Chapter 2*). Next, the linear analysis was extended with deep neural network-based AV fusion models, trained to extract correlated AV feature embeddings directly from raw audio and video data (*Chapter 3*). Building upon the findings in *Chapter 2* and *Chapter 3*, we finally presented an AV speech separation model that used visual cues to perform acoustic source separation (*Chapter 4*). The presented separation model can work in causal settings while also achieving comparable performance to computational "heavier" models, raising perspectives for real-time applications.

5.1 Summary of main results

In our first study (*Chapter 2*), we used a CCA model to analyze natural AV speech from many talkers. We estimated 3D facial landmarks directly from videos of single talkers, allowing us to capture facial motions on a much larger cohort of talkers than what would have been possible with the traditional and more cumbersome manual data collection methods. Presented with filtered audio envelopes and the 3D facial landmarks extracted from the videos, the CCA model learned speech envelope filterings correlated with facial motion patterns. Our results revealed two primary temporal ranges of envelope fluctuations related to facial motion across speakers. The first is distributed around 3-4 Hz and relates to mouth openings. The second range of modulations peaks around 1-2 Hz and relates to more global face and head motion. In both cases, we found the envelope information correlated with landmarks distributed across the face,

reflecting that natural speech involves highly coordinated motor activity. This implies that envelope cues are not only available from mouth movements but can also be retrieved from extraoral parts of the face and head. Notably, the derived AV correlations were predictive across speakers implying that these temporal cues are consistent in natural AV speech statistics. The relatively simple AV input features and the linear nature of CCA made the analysis straightforward and transparent. However, at the same time, the approach could potentially have overlooked essential non-linear aspects of AV signal statistics.

To address the limitations of the CCA approach, we next analyzed the AV speech data through the "dissection" of two Convolutional Neural Networks (CNNs). The networks were trained in a self-supervised manner directly on raw audio and video to extract correlated AV feature embeddings (*Chapter 3*). Both networks achieved close to 100 % accuracy when evaluated in a three-speaker identification task (compared to 76 % for CCA), and for both networks, the average correlation value between matching AV segments was close to 70 % (compared to 22 % for CCA), whereas, for non-matching AV segments, it was near 0 %. Examination of the audio filters learned by the networks revealed that the networks learned to extract features akin to envelopes of the speech audio, providing the models with basic temporal audio information. Moreover, it was shown that both networks in the visual domain primarily tended to focus heavily on the mouth region during speech production.

In our third study (*Chapter 4*), we used the correlated AV feature embeddings of the AV fusion models presented in the two previous chapters to train an AV speech separation model that utilized visual speaker information to extract the corresponding target speech from single-channel audio mixtures. We observed a strong relationship between the correlation of the AV feature embeddings and the performance of the speech separation models. Furthermore, we found the speech separation models to perform well in causal and non-causal settings under various acoustic conditions.

5.2 Discussion

5.2.1 Analysis of audiovisual speech

Previous studies on AV speech have either focused on facial motion tracking from a limited number of subjects (Lucero et al., 2005; Lucero and Munhall, 2008;

Vatikiotis-Bateson et al., 1996; Yehia et al., 2002) or only considered specific facial regions such as the mouth area (Chandrasekaran et al., 2009). In contrast, the approach presented in chapter 2 allowed for the analysis of natural speech across thousands of speakers by capturing facial movements across the entire face using a deep neural network (Bulat and Tzimiropoulos, 2017) to estimate 3D facial landmarks.

While our analyses highlighted the well-known synchronized mouth-envelope modulations in the 4 Hz range (Chandrasekaran et al., 2009), it also identified more global face and head motions related to envelope modulations around 1-2 Hz. Interestingly, the correspondences between larger head motions and slower envelope modulations were only identified when analyzing natural speech across many speakers from the LRS3 dataset (Afouras et al., 2018a) but not when analyzing the simpler GRID dataset (Cooke et al., 2006). As both datasets contained approximately 30 hours of AV speech data, the differences in the results must instead arise from the fact that the LRS3 dataset includes many more individual speakers, a more diverse and complex vocabulary, and that the speakers move freely. The differences between the two analyses, thus, demonstrate the importance of choosing the speech material with care, as the findings ultimately will depend on the speech task in the speech material.

Another important aspect of the first study is that we intentionally restricted the search for AV cues to focus on envelope amplitude modulation rates correlated with facial movements. Although this made sense to better understand how modulations are related to facial movements, it also restricted the model from inspecting audio and visual cues that potentially are even more correlated. In particular, non-linear aspects of AV speech, such as the relation between visible articulators and the produced speech signal, are non-linear in essential aspects (Scholes et al., 2020; Yehia et al., 2002) and can not be captured in the proposed method.

To capture non-linear relationships, we analyzed AV speech through the analysis of deep neural networks in chapter 3. Like in chapter 2, the objective of the models presented in chapter 3 was to identify correlated audio and visual cues from natural AV speech, but the two approaches are conceptually different. In contrast to chapter 2, we imposed few restrictions on the AV fusion models in the second study, enabling the models to freely learn information from the raw input modalities that would result in correlated audio and visual representations. The approach undoubtedly made the analysis process less transparent

and harder to interpret. However, the proposed method still allowed for interpretation of the first two layers of audio filters and enabled backtracking of the AV cues to the input spaces. Especially, the use of sinc-based convolutions (Ravanelli and Bengio, 2018a) allowed for interpreting the audio filters and which audio frequency content yielded the most correlated audio cues. Moreover, we could conclude that the proposed models learned to focus specifically on the mouth but also on extra-oral parts of the face, supporting the findings from chapter 2 that visual cues correlated with audio cues are distributed across the entire face.

While the neural networks can be more challenging to interpret, the extracted AV cues are significantly more correlated than those obtained with the CCA approach in chapter 2 (69% for SincNet vs. 22% for CCA). The fact that the AV cues are more correlated also makes them more attractive for use in a downstream task like AV speech separation. The approach presented in chapter 3 also resembles an emerging strategy that heavily relies on complex models to learn from massive datasets, rather than carefully controlling every step of the analysis (Aldeneh et al., 2021; Ravanelli and Bengio, 2018a). Interpretability and transparency are, to some extent, compromitted on behalf of better performance when using neural networks. However, the neural network's ability to learn from that vast amount of available video data combined with clever model architectures that allow for some interpretation might be a way to identify unknown AV relationships that are otherwise omitted for the sake of transparency and simplicity.

5.2.2 Speech separation

Chapter 4 introduced a speaker-independent AV speech separation network that used visual target speaker cues to extract the corresponding target speech from single-channel audio mixtures. We adapted the speaker-dependent brain-informed speech separation network by Ceolini et al. (2020) to rely on visual speaker cues rather than speech envelopes derived from EEG signals. The use of visual cues is attractive for several reasons. First, the visual speaker cues are relatively easy to capture (with video cameras) and naturally correlate with the audio, making them favorable as guiding signals. Further, the visual scene is not corrupted by acoustic noise in the auditory scene, thus providing a reliable guiding signal. Last, visual cues can be obtained from all speakers using a single pretrained model, alleviating the need to train person-specific models, which is

needed to estimate speech envelopes from EEG signals.

Whereas several AV speech separation studies rely on the memory-intensive Bidirectional Long-Short-Term-Memory (BLSTM) networks (Ephrat et al., 2018; Ochiai et al., 2019), our proposed method only uses 2D convolutions, which decrease the number of parameters in the model, thus making it more computationally efficient. Besides being memory efficient and speaker-independent, the model also performs well in causal settings, suggesting that it could be a promising approach for real-time applications and low-resource devices such as hearing aids.

However, real-time implementation is to some extent hindered by the temporal window required to compute the short-time Fourier transform (STFT). An alternative approach could be to replace the STFT representation with time-domain representations (Luo and Mesgarani, 2019), which would resolve the latency issue. It might also be possible to increase the performance and roboustness of our system by including multi-channel audio input, as it can provide complementary spatial information which has been shown to benefit other AV speech separation systems, particularly in reverberant settings and when the talkers face is occluded (Gu et al., 2020; Tan et al., 2020).

5.3 Perspectives

In chapter 2 and chapter 3, we based our analyses on the LRS3 dataset (Afouras et al., 2018a), which consists of videos captured during TED talks. While this speech material can be considered as "wild" data or natural speech (Michelsanti et al., 2021), it does not include conversational speech. We know from several studies that the speech dynamics change during conversations, as conversational speech involves turn takings (Donnarumma et al., 2017) and non-verbal movements related to social interactions (Latif et al., 2014). This type of speech material has so far been unavailable. However, the EGO4D Consortium has recently announced that they soon will publish Ego4D (Grauman et al., 2021), a massive-scale egocentric video dataset containing AV speech from first-person views. Analyses of this type of data would most likely reveal movements specifically related to the dynamics of conversational speech, which are not captured in our analyses. It would therefore be interesting to investigate and compare differences between the datasets. Moreover, the Ego4D dataset is collected from nine different countries worldwide, and it would be interesting to investigate

the differences in conversations based on regions. Such information might be useful for region- or language-specific speech separation models.

The 3D facial landmarks used in our studies consisted of 68 points per face. Recently, Grishchenko et al. (2020) published a deep neural network that reliably estimates facial meshes consisting of 468 facial points while working in real-time on mobile devices. With the additional information that more landmarks could provide in real-time implementation, it would be an appealing feature in future analyses and AV speech separation systems.

Today, standard hearing aids do not utilize visual inputs to improve speech intelligibility. However, with the emergence of smart glasses with built-in video cameras, microphones, and eye-trackers, it is possible to imagine hearing-aid glasses where the user's eye-gaze controls auditory feedback. More computational power available in smaller devices in combination with efficient speech separation models as proposed here may pave the way for such devices in a not too distant future.

Bibliography

Afouras, T., J. S. Chung, and A. Zisserman (2018a). “LRS3-TED: a large-scale dataset for visual speech recognition”. In:

Afouras, T., J. S. Chung, and A. Zisserman (2018b). “The conversation: Deep audio-visual speech enhancement”. In: *arXiv preprint arXiv:1804.04121*.

Afouras, T., A. Owens, J. S. Chung, and A. Zisserman (2020). “Self-supervised learning of audio-visual objects from video”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, pp. 208–224.

Aldeneh, Z. et al. (2021). “On The Role of Visual Cues in Audiovisual Speech Enhancement”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. Elsevier, pp. 8423–8427.

Alexandrou, A. M., T. Saarinen, J. Kujala, and R. Salmelin (2016). “A multimodal spectral approach to characterize rhythm in natural speech”. In: *The Journal of the Acoustical Society of America* 139.1, pp. 215–226.

Arnold, P. and F. Hill (2001). “Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact”. In: *British Journal of Psychology* 92.2, pp. 339–355.

Bennett, J. W., P. van Lieshout, and C. M. Steele (2007). “Tongue control for speech and swallowing in healthy younger and older subjects”. In: *International Journal of Orofacial Myology and Myofunctional Therapy* 33.1, pp. 5–18.

Bernstein, L. E., E. T. Auer Jr, and S. Takayanagi (2004). “Auditory speech detection in noise enhanced by lipreading”. In: *Speech Communication* 44.1-4, pp. 5–18.

Brown, S., Y. Yuan, and M. Belyk (2021). “Evolution of the speech-ready brain: The voice/jaw connection in the human motor cortex”. In: *Journal of Comparative Neurology* 529.5, pp. 1018–1028.

Bulat, A. and G. Tzimiropoulos (2017). “How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)”. In: *International Conference on Computer Vision*.

Butterworth, B. and U. Hadar (1989). “Gesture, speech, and computational stages: A reply to McNeill.” In:

Ceolini, E. et al. (2020). “Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception”. In: *NeuroImage* 223, p. 117282.

Chandar, S., M. M. Khapra, H. Larochelle, and B. Ravindran (2016). “Correlational neural networks”. In: *Neural computation* 28.2, pp. 257–285.

Chandrasekaran, C., A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar (2009). “The Natural Statistics of Audiovisual Speech”. In: *PLoS Computational Biology* 5.

Chatfield, K., K. Simonyan, A. Vedaldi, and A. Zisserman (2014). “Return of the devil in the details: Delving deep into convolutional nets”. In: *arXiv preprint arXiv:1405.3531*.

Cheng, Y., R. Wang, Z. Pan, R. Feng, and Y. Zhang (2020). “Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning”. In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3884–3892.

Cheveigné, A. de, M. Slaney, S. A. Fuglsang, and J. Hjortkjaer (2021). “Auditory stimulus-response modeling with a match-mismatch task”. In: *Journal of Neural Engineering* 18.4, p. 046040.

Cheveigné, A. de, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjaer, M. Slaney, and E. Lalor (2018). “Decoding the auditory brain with canonical component analysis”. In: *NeuroImage* 172, pp. 206–216.

Chung, S.-W., J. S. Chung, and H.-G. Kang (2020). “Perfect match: Self-supervised embeddings for cross-modal retrieval”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.3, pp. 568–576.

Cooke, M., J. Barker, S. Cunningham, and X. Shao (2006). “An audio-visual corpus for speech perception and automatic speech recognition”. In: *The Journal of the Acoustical Society of America* 120.5, pp. 2421–2424.

Dau, T., D. Püschel, and A. Kohlrausch (1996). “A quantitative model of the “effective”signal processing in the auditory system. I. Model structure”. In: *The Journal of the Acoustical Society of America* 99.6, pp. 3615–3622.

Delgutte, B., B. M. Hammond, and P. A. Cariani (1998). “Neural coding of the temporal envelope of speech: relation to modulation transfer functions”. In: *Psychophysical and physiological advances in hearing*, pp. 595–603.

Ding, N., L. Melloni, H. Zhang, X. Tian, and D. Poeppel (2016). “Cortical tracking of hierarchical linguistic structures in connected speech”. In: *Nature neuroscience* 19.1, pp. 158–164.

Ding, N., A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel (2017). “Temporal modulations in speech and music”. In: *Neuroscience & Biobehavioral Reviews* 81, pp. 181–187.

Ding, N. and J. Z. Simon (2014). “Cortical entrainment to continuous speech: functional roles and interpretations”. In: *Frontiers in human neuroscience* 8, p. 311.

Doelling, K. B., L. H. Arnal, O. Ghitza, and D. Poeppel (2014). “Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing”. In: *Neuroimage* 85, pp. 761–768.

Dolata, J. K., B. L. Davis, and P. F. MacNeilage (2008). “Characteristics of the rhythmic organization of vocal babbling: Implications for an amodal linguistic rhythm”. In: *Infant Behavior and development* 31.3, pp. 422–431.

Donnarumma, F., H. Dindo, P. Iodice, and G. Pezzulo (2017). “You cannot speak and listen at the same time: a probabilistic model of turn-taking.” In: *Biological cybernetics* 111.2.

Dupoux, E. (2018). “Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner”. In: *Cognition* 173, pp. 43–59.

Edwards, E. and E. F. Chang (2013). “Syllabic (2–5 Hz) and fluctuation (1–10 Hz) ranges in speech and auditory processing”. In: *Hearing research* 305, pp. 113–134.

Ejiri, K. and N. Masataka (1999). “Synchronization between preverbal vocal behavior and motor action in early infancy: II. An acoustical examination of the functional significance of the synchronization.” In: *Japanese Journal of Psychology*.

Ejiri, K. and N. Masataka (2001). “Co-occurences of preverbal vocal behavior and motor action in early infancy”. In: *Developmental Science* 4.1, pp. 40–48.

Elliott, T. M. and F. E. Theunissen (2009). “The modulation transfer function for speech intelligibility”. In: *PLoS computational biology* 5.3, e1000302.

Ephrat, A. et al. (2018). “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation”. In: *arXiv preprint arXiv:1804.03619*.

Erber, N. P. (1975). “Auditory-visual perception of speech”. In: *Journal of speech and hearing disorders* 40.4, pp. 481–492.

Esteve-Gibert, N. and P. Prieto (2014). “Infants temporally coordinate gesture-speech combinations before they produce their first words”. In: *Speech Communication* 57, pp. 301–316.

Ewert, S. D. and T. Dau (2000). “Characterizing frequency selectivity for envelope fluctuations”. In: *The Journal of the Acoustical Society of America* 108.3, pp. 1181–1196.

Fuchs, S. and P. Perrier (2005). “On the complex nature of speech kinematics”. In: *ZAS papers in Linguistics* 42, pp. 137–165.

Ghazanfar, A. A. and D. Y. Takahashi (2014a). “Facial expressions and the evolution of the speech rhythm”. In: *Journal of cognitive neuroscience* 26.6, pp. 1196–1207.

Ghazanfar, A. A. and D. Y. Takahashi (2014b). “The evolution of speech: vision, rhythm, cooperation”. In: *Trends in cognitive sciences* 18.10, pp. 543–553.

Ghazanfar, A. A., D. Y. Takahashi, N. Mathur, and W. T. Fitch (2012). “Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics”. In: *Current Biology* 22.13, pp. 1176–1182.

Ginosar, S., A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik (2019). “Learning individual styles of conversational gesture”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506.

Giraud, A.-L. and D. Poeppel (2012). “Cortical oscillations and speech processing: emerging computational principles and operations”. In: *Nature neuroscience* 15.4, pp. 511–517.

Girin, L., J.-L. Schwartz, and G. Feng (2001). “Audio-visual enhancement of speech in noise”. In: *The Journal of the Acoustical Society of America* 109.6, pp. 3007–3020.

Goswami, U. and V. Leong (2013). “Speech rhythm and temporal structure: converging perspectives?” In: *Laboratory Phonology* 4.1, pp. 67–92.

Graf, H. P., E. Cosatto, V. Strom, and F. J. Huang (2002). “Visual prosody: Facial movements accompanying speech”. In: *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*. IEEE, pp. 396–401.

Grant, K. W., B. E. Walden, and P. F. Seitz (1998). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration". In: *The Journal of the Acoustical Society of America* 103.5, pp. 2677–2690.

Grauman, K. et al. (2021). "Around the World in 3,000 Hours of Egocentric Video". In: *CoRR* abs/2110.07058.

Greenberg, S., H. Carvey, L. Hitchcock, and S. Chang (2003). "Temporal properties of spontaneous speech—a syllable-centric perspective". In: *Journal of Phonetics* 31.3-4, pp. 465–485.

Grimme, B., S. Fuchs, P. Perrier, and G. Schöner (2011). "Limb versus speech motor control: A conceptual review". In: *Motor control* 15.1, pp. 5–33.

Grishchenko, I., A. Ablavatski, Y. Kartynnik, K. Raveendran, and M. Grundmann (2020). "Attention Mesh: High-fidelity Face Mesh Prediction in Real-time". In: *arXiv preprint arXiv:2006.10962*.

Gu, R., S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu (2020). "Multi-modal multi-channel target speech separation". In: *IEEE Journal of Selected Topics in Signal Processing* 14.3, pp. 530–541.

Guaïtella, I., S. Santi, B. Lagrue, and C. Cavé (2009). "Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation". In: *Language and speech* 52.2-3, pp. 207–222.

Hadar, U., T. J. Steiner, E. C. Grant, and F. C. Rose (1984). "The timing of shifts of head postures during conservation". In: *Human Movement Science* 3.3, pp. 237–245.

Hadar, U., T. J. Steiner, E. C. Grant, and F. C. Rose (1983). "Kinematics of head movements accompanying speech during conversation". In: *Human Movement Science* 2.1-2, pp. 35–46.

Harwath, D., A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass (2018). "Jointly discovering visual objects and spoken words from raw sensory input". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 649–665.

Haufe, S. et al. (2014). "On the interpretation of weight vectors of linear models in multivariate neuroimaging". In: *Neuroimage* 87, pp. 96–110.

Head, T. et al. (Mar. 2018). *scikit-optimize/scikit-optimize: v0.5.2*. Version v0.5.2.

Helfer, K. S. (1997). "Auditory and auditory-visual perception of clear and conversational speech". In: *Journal of Speech, Language, and Hearing Research* 40.2, pp. 432–443.

Hiiemae, K. M., J. B. Palmer, S. W. Medicis, J. Hegener, B. S. Jackson, and D. E. Lieberman (2002). “Hyoid and tongue surface movements in speaking and eating”. In: *Archives of Oral Biology* 47.1, pp. 11–27.

Hillenbrand, J., L. A. Getty, M. J. Clark, and K. Wheeler (1995). “Acoustic characteristics of American English vowels”. In: *The Journal of the Acoustical Society of America* 97.5, pp. 3099–3111.

Houtgast, T. and H. J. M. Steeneken (1973). “The modulation transfer function in room acoustics as a predictor of speech intelligibility”. In: *Acta Acustica United with Acustica* 28.1, pp. 66–73.

Ideli, E., B. Sharpe, I. V. Bajić, and R. G. Vaughan (2019). “Visually assisted time-domain speech enhancement”. In: *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, pp. 1–5.

Inbar, M., E. Grossman, and A. N. Landau (2020). “Sequences of Intonation Units form a˜ 1 Hz rhythm”. In: *Scientific reports* 10.1, pp. 1–9.

Iverson, J. M. and M. K. Fagan (2004). “Infant vocal–motor coordination: precursor to the gesture–speech system?” In: *Child development* 75.4, pp. 1053–1066.

Iverson, J. M. and S. Goldin-Meadow (1998). “Why people gesture when they speak”. In: *Nature* 396.6708, pp. 228–228.

Iverson, J. M. and E. Thelen (1999). “Hand, mouth and brain. The dynamic emergence of speech and gesture”. In: *Journal of Consciousness studies* 6.11-12, pp. 19–40.

Jacewicz, E., R. A. Fox, C. O’Neill, and J. Salmons (2009). “Articulation rate across dialect, age, and gender”. In: *Language variation and change* 21.2, p. 233.

Johnston, A., B. B. Brown, and R. Elson (2021). “Synchronous facial action binds dynamic facial features”. In: *Scientific Reports* 11.1, pp. 1–10.

Jørgensen, S. and T. Dau (2011). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing”. In: *The Journal of the Acoustical Society of America* 130.3, pp. 1475–1487.

Karpathy, A., A. Joulin, and L. Fei-Fei (2014). “Deep fragment embeddings for bidirectional image sentence mapping”. In: *arXiv preprint arXiv:1406.5679*.

Keitel, A., J. Gross, and C. Kayser (2018). “Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features”. In: *PLoS biology* 16.3, e2004473.

Kim, J., E. Cvejic, and C. Davis (2014). “Tracking eyebrows and head gestures associated with spoken prosody”. In: *Speech Communication* 57, pp. 317–330.

Krahmer, E. and M. Swerts (2007). “The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception”. In: *Journal of memory and language* 57.3, pp. 396–414.

Kuratate, T., K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia (1999). “Audio-visual synthesis of talking faces from speech production correlates”. In: *Sixth European Conference on Speech Communication and Technology*.

Kuratate, T., E. Vatikiotis-Bateson, and H. C. Yehia (2005). “Estimation and animation of faces using facial motion mapping and a 3D face database”. In: *Computer-graphic facial reconstruction*, pp. 325–346.

Latif, N., A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. Munhall (2014). “Movement coordination during conversation”. In: *PLoS one* 9.8, e105036.

Le Roux, J., S. Wisdom, H. Erdogan, and J. R. Hershey (2019). “SDR–half-baked or well done?” In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 626–630.

LeCun, Y., Y. Bengio, and G. Hinton (2015). “Deep learning”. In: *nature* 521.7553, pp. 436–444.

Lindblad, P., S. Karlsson, and E. Heller (1991). “Mandibular movements in speech phrases—A syllabic quasiregular continuous oscillation”. In: *Scandinavian Journal of Logopedics and Phoniatrics* 16.1-2, pp. 36–42.

Lucero, J. C., S. T. R. Maciel, D. A. Johns, and K. G. Munhall (2005). “Empirical modeling of human face kinematics during speech using motion clustering”. In: *The Journal of the Acoustical Society of America* 118.1, pp. 405–409.

Lucero, J. C. and K. G. Munhall (2008). “Analysis of facial motion patterns during speech using a matrix factorization algorithm”. In: *The Journal of the Acoustical Society of America* 124.4, pp. 2283–2290.

Luo, H., Z. Liu, and D. Poeppel (2010). “Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation”. In: *PLoS biology* 8.8, e1000445.

Luo, Y. and N. Mesgarani (2019). “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation”. In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8, pp. 1256–1266.

MacLeod, A. and Q. Summerfield (1987). "Quantifying the contribution of vision to speech perception in noise". In: *British journal of audiology* 21.2, pp. 131–141.

MacNeilage, P. F. (1998). "The frame/content theory of evolution of speech production". In: *Behavioral and brain sciences* 21.4, pp. 499–511.

Mariooryad, S. and C. Busso (2012). "Generating human-like behaviors using joint, speech-driven models for conversational agents". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.8, pp. 2329–2340.

Matsuo, K. and J. B. Palmer (2010). "Kinematic linkage of the tongue, jaw, and hyoid during eating and speech". In: *Archives of oral biology* 55.4, pp. 325–331.

McClave, E. (1998). "Pitch and manual gestures". In: *Journal of Psycholinguistic Research* 27.1, pp. 69–89.

McGurk, H. and J. MacDonald (1976). "Hearing lips and seeing voices". In: *Nature* 264.5588, pp. 746–748.

McNeill, D. (1992). *Hand and mind*. De Gruyter Mouton.

McWalter, R. and T. Dau (2017). "Cascaded amplitude modulations in sound texture perception". In: *Frontiers in neuroscience* 11, p. 485.

Michelsanti, D. et al. (2021). "An overview of deep-learning-based audio-visual speech enhancement and separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Montavon, G., W. Samek, and K.-R. Müller (2018). "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing* 73, pp. 1–15.

Moore, C. A., A. Smith, and R. L. Ringel (1988). "Task-specific organization of activity in human jaw muscles". In: *Journal of Speech, Language, and Hearing Research* 31.4, pp. 670–680.

Müller, E. and G. MacLeod (1982). "Perioral biomechanics and its relation to labial motor control". In: *The Journal of the Acoustical Society of America* 71.S1, S33–S33.

Müller, P., G. A. Kalberer, M. Proesmans, and L. Van Gool (2005). "Realistic speech animation based on observed 3-D face dynamics". In: *IEE Proceedings-Vision, Image and Signal Processing* 152.4, pp. 491–500.

Munhall, K. G., J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson (2004). "Visual prosody and speech intelligibility: Head movement improves auditory speech perception". In: *Psychological science* 15.2, pp. 133–137.

Munhall, K. G. and E. Vatikiotis-Bateson (1998). "The moving face during speech communication". In: *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*, pp. 123–139.

Nagrani, A., J. S. Chung, W. Xie, and A. Zisserman (2020). "Voxceleb: Large-scale speaker verification in the wild". In: *Computer Speech & Language* 60, p. 101027.

Nelson, P. C. and L. H. Carney (2004). "A phenomenological model of peripheral and central neural responses to amplitude-modulated tones". In: *The Journal of the Acoustical Society of America* 116.4, pp. 2173–2186.

Ochiai, T., M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani (2019). "Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-Visual Speaker Clues." In: *INTERSPEECH*, pp. 2718–2722.

Ohala, J. J. (1975). "The temporal regulation of speech". In: *Auditory analysis and perception of speech*, pp. 431–453.

Owens, A. and A. A. Efros (2018). "Audio-visual scene analysis with self-supervised multisensory features". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648.

Patterson, R. D., I. Nimmo-Smith, J. Holdsworth, and P. Rice (1987). "An efficient auditory filterbank based on the gammatone function". In: *A meeting of the IOC Speech Group on Auditory Modelling at RSRE*. Vol. 2. 7.

Pellegrino, F., C. Coupé, and E. Marsico (2011). "A cross-language perspective on speech information rate". In: *Language*, pp. 539–558.

Poeppel, D. and M. F. Assaneo (2020). "Speech rhythms and their neural foundations". In: *Nature reviews neuroscience* 21.6, pp. 322–334.

Potamianos, G., C. Neti, G. Gravier, A. Garg, and A. W. Senior (2003). "Recent advances in the automatic recognition of audiovisual speech". In: *Proceedings of the IEEE* 91.9, pp. 1306–1326.

Pouw, W., S. J. Harrison, and J. A. Dixon (2020a). "Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony." In: *Journal of Experimental Psychology: General* 149.2, p. 391.

Pouw, W., A. Paxton, S. J. Harrison, and J. A. Dixon (2020b). "Acoustic information about upper limb movement in voicing". In: *Proceedings of the National Academy of Sciences* 117.21, pp. 11364–11367.

Ramsay, J. O., K. G. Munhall, V. L. Gracco, and D. J. Ostry (1996). "Functional data analyses of lip motion". In: *The Journal of the Acoustical Society of America* 99.6, pp. 3718–3727.

Ravanelli, M. and Y. Bengio (2018a). “Interpretable convolutional filters with sincnet”. In: *arXiv preprint arXiv:1811.09725*.

Ravanelli, M. and Y. Bengio (2018b). “Speaker recognition from raw waveform with sincnet”. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 1021–1028.

Reisberg, D., J. Mclean, and A. Goldfield (1987). “Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli.” In:

Rimmele, J. M., D. Poeppel, and O. Ghitza (2021). “Acoustically Driven Cortical δ Oscillations Underpin Prosodic Chunking”. In: *Eneuro* 8.4.

Risueno-Segovia, C. and S. R. Hage (2020). “Theta synchronization of phonatory and articulatory systems in marmoset monkey vocal production”. In: *Current Biology* 30.21, pp. 4276–4283.

Roberts, S. G., F. Torreira, and S. C. Levinson (2015). “The effects of processing and sequence organization on the timing of turn taking: a corpus study”. In: *Frontiers in psychology* 6, p. 509.

Rosenblum, L. D. (2008). “Speech perception as a multimodal phenomenon”. In: *Current Directions in Psychological Science* 17.6, pp. 405–409.

Ross, L. A., D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe (2007). “Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments”. In: *Cerebral cortex* 17.5, pp. 1147–1153.

Rouger, J., S Lagleyre, B. Fraysse, S Deneve, O. Deguine, and P Barone (2007). “Evidence that cochlear-implanted deaf patients are better multisensory integrators”. In: *Proceedings of the National Academy of Sciences* 104.17, pp. 7295–7300.

Sargin, M. E., Y. Yemez, E. Erzin, and A. M. Tekalp (2007). “Audiovisual synchronization and fusion using canonical correlation analysis”. In: *IEEE Transactions on Multimedia* 9.7, pp. 1396–1403.

Scholes, C., J. I. Skipper, and A. Johnston (2020). “The interrelationship between the face and vocal tract configuration during audiovisual speech”. In: *Proceedings of the National Academy of Sciences* 117.51, pp. 32791–32798.

Schorr, E. A., N. A. Fox, V. van Wassenhove, and E. I. Knudsen (2005). “Auditory-visual fusion in speech perception in children with cochlear implants”. In: *Proceedings of the National Academy of Sciences* 102.51, pp. 18748–18750.

Schroeder, C. E. and J. Foxe (2005). “Multisensory contributions to low-level, ‘unisensory’ processing”. In: *Current opinion in neurobiology* 15.4, pp. 454–458.

Schwartz, J.-L., F. Berthommier, and C. Savariaux (2004). “Seeing to hear better: evidence for early audio-visual interactions in speech identification”. In: *Cognition* 93.2, B69–B78.

Schwartz, J.-L. and C. Savariaux (2014). “No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag”. In: *PLoS Comput Biol* 10.7, e1003743.

Sharma, R., K. Somandepalli, and S. Narayanan (2020). “Crossmodal learning for audio-visual speech event localization”. In: *arXiv preprint arXiv:2003.04358*.

Shinn-Cunningham, B. G. and V. Best (2008). “Selective attention in normal and impaired hearing”. In: *Trends in amplification* 12.4, pp. 283–299.

Sigg, C., B. Fischer, B. Ommer, V. Roth, and J. Buhmann (2007). “Nonnegative CCA for audiovisual source separation”. In: *2007 IEEE Workshop on Machine Learning for Signal Processing*. IEEE, pp. 253–258.

Singh, N. C. and F. E. Theunissen (2003). “Modulation spectra of natural sounds and ethological theories of auditory processing”. In: *The Journal of the Acoustical Society of America* 114.6, pp. 3394–3411.

Slaney, M. and M. Covell (2001). “Facesync: A linear operator for measuring synchronization of video facial images and audio tracks”. In: *Advances in Neural Information Processing Systems*, pp. 814–820.

Smith, A. and H. N. Zelaznik (2004). “Development of functional synergies for speech motor coordination in childhood and adolescence”. In: *Developmental psychobiology* 45.1, pp. 22–33.

Stevenson, R. A. and M. T. Wallace (2013). “Multisensory temporal integration: task and stimulus dependencies”. In: *Experimental brain research* 227.2, pp. 249–261.

Sumby, W. H. and I. Pollack (1954). “Visual contribution to speech intelligibility in noise”. In: *The journal of the acoustical society of america* 26.2, pp. 212–215.

Tan, K., Y. Xu, S.-X. Zhang, M. Yu, and D. Yu (2020). “Audio-visual speech separation and dereverberation with a two-stage multimodal network”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.3, pp. 542–553.

Thomas, S. M. and T. R. Jordan (2004). “Contributions of oral and extraoral facial movement to visual and audiovisual speech perception.” In: *Journal of Experimental Psychology: Human Perception and Performance* 30.5, p. 873.

Trujillo, J. P., S. C. Levinson, and J. Holler (2021). "Visual Information in Computer-Mediated Interaction Matters: Investigating the Association Between the Availability of Gesture and Turn Transition Timing in Conversation". In: *International Conference on Human-Computer Interaction*. Springer, pp. 643–657.

Varnet, L., M. C. Ortiz-Barajas, R. G. Erra, J. Gervain, and C. Lorenzi (2017). "A cross-linguistic study of speech modulation spectra". In: *The Journal of the Acoustical Society of America* 142.4, pp. 1976–1989.

Vatikiotis-Bateson, E., K. G. Munhall, Y. Kasahara, F. Garcia, and H. Yehia (1996). "Characterizing audiovisual information during speech". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Vol. 3. IEEE, pp. 1485–1488.

Viemeister, N. F. (1979). "Temporal modulation transfer functions based upon modulation thresholds". In: *The Journal of the Acoustical Society of America* 66.5, pp. 1364–1380.

Vincent, E., R. Gribonval, and C. Févotte (2006). "Performance measurement in blind audio source separation". In: *IEEE transactions on audio, speech, and language processing* 14.4, pp. 1462–1469.

Vitkovitch, M. and P. Barber (1996). "Visible speech as a function of image quality: Effects of display parameters on lipreading ability". In: *Applied cognitive psychology* 10.2, pp. 121–140.

Wagner, P., Z. Malisz, and S. Kopp (2014). "Gesture and speech in interaction: An overview". In: *Speech Communication* 57, pp. 209–232.

Walsh, B. and A. Smith (2002). "Articulatory Movements in Adolescents: Evidence for Protracted Development of Speech Motor Control Process". In: *Journal of Speech, Language, and Hearing Research* 45.6, pp. 1119–1133.

Wang, D. and J. Chen (2018). "Supervised speech separation based on deep learning: An overview". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10, pp. 1702–1726.

Wu, J. et al. (2019). "Time domain audio visual speech separation". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 667–673.

Yehia, H. C., T. Kuratate, and E. Vatikiotis-Bateson (2002). "Linking facial animation, head motion and speech acoustics". In: *Journal of Phonetics* 30.3, pp. 555–568.

Yuan, Y., R. Wayland, and Y. Oh (2020). “Visual analog of the acoustic amplitude envelope benefits speech perception in noise”. In: *The Journal of the Acoustical Society of America* 147.3, EL246–EL251.

Zhang, Y. S. and A. A. Ghazanfar (2020). “A hierarchy of autonomous systems for vocal production”. In: *Trends in neurosciences* 43.2, pp. 115–126.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba (2016). “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.
External examiners: Mark Lutman, Stefan Stenfeld

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.
External examiners: Brian Moore, Kathrin Krumbholz

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.
External examiners: Michael Akeroyd, Armin Kohlrausch

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.
External examiners: Jesko Verhey, Steven van de Par

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.
External examiners: Björn Hagerman, Ejnar Laukli

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.
External examiners: Inga Holube, Birgitta Larsby

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.
External examiners: Birger Kollmeier, Ray Meddis

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.
External examiners: David Kemp, Stephen Neely

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.
External examiners: Bernhard Seeber, Michael Vorländer

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.
External examiners: Christopher Plack, Christian Lorenzi

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.
External examiners: Joost Festen, Jürgen Tchorz

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.
External examiners: Bob Burkard, Stephen Neely

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.
External examiners: Stuart Rosen, Christian Lorenzi

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.
External examiners: Michael Stone, Oded Ghitza

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.
External examiners: John Culling, Martin Cooke

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.
External examiners: Lawrence Rosenblum, Matthias Gondan

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.
External examiners: Shihab Shamma, Guy Brown

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.
External examiners: Sascha Spors, Ville Pulkki

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.
External examiners: Bernhard Seeber, Steven van de Par

Vol. 20: *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.
External examiners: Christopher Plack, Enrique Lopez-Poveda

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.
External examiners: Steven van de Par, John Culling

Vol. 22: *Federica Bianchi*, Pitch representations in the impaired auditory system and implications for music perception, 2016.
External examiners: Ingrid Johnsrude, Christian Lorenzi

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.
External examiners: Judy Dubno, Martin Cooke

Vol. 24: *Johannes Käsbach*, Characterizing apparent source width perception, 2016.
External examiners: William Whitmer, Jürgen Tchorz

Vol. 25: *Gusztáv Löcsei*, Lateralized speech perception with normal and impaired hearing, 2016.
External examiners: Thomas Brand, Armin Kohlrausch

Vol. 26: *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.
External examiners: Laurel Carney, Bob Carlyon

Vol. 27: *Henrik Gerd Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.
External examiners: Volker Hohmann, Piotr Majdak

Vol. 28: *Richard Ian McWalter*, Analysis of the auditory system via synthesis of natural sounds, speech and music, 2017.
External examiners: Maria Chait, Christian Lorenzi

Vol. 29: *Jens Cubick*, Characterizing the auditory cues for the processing and perception of spatial sounds, 2017.
External examiners: Ville Pulkki, Pavel Zahorik

Vol. 30: *Gerard Encina-Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.
External examiners: Roland Schaette, Ian Bruce

Vol. 31: *Christoph Scheidiger*, Assessing speech intelligibility in hearing-impaired listeners, 2018.
External examiners: Enrique Lopez-Poveda, Tim Jürgens

Vol. 32: *Alan Wiinberg*, Perceptual effects of non-linear hearing aid amplification strategies, 2018.
External examiners: Armin Kohlrausch, James Kates

Vol. 33: *Thomas Bentsen*, Computational speech segregation inspired by principles of auditory processing, 2018.
External examiners: Stefan Bleeck, Jürgen Tchorz

Vol. 34: *François Guérit*, Temporal charge interactions in cochlear implant listeners, 2018.
External examiners: Julie Arenberg, Olivier Macherey

Vol. 35: *Andreu Paredes Gallardo*, Behavioral and objective measures of stream segregation in cochlear implant users, 2018.
External examiners: Christophe Micheyl, Monita Chatterjee

Vol. 36: *Søren Fuglsang*, Characterizing neural mechanisms of attention-driven speech processing, 2019.
External examiners: Shihab Shamma, Maarten de Vos

Vol. 37: *Borys Kowalewski*, Assessing the effects of hearing-aid dynamic-range compression on auditory signal processing and perception, 2019.
External examiners: Brian Moore, Graham Naylor

Vol. 38: *Helia Relaño Iborra*, Predicting speech perception of normal-hearing and hearing-impaired listeners, 2019.
External examiners: Ian Bruce, Armin Kohlrausch

Vol. 39: *Axel Ahrens*, Characterizing auditory and audio-visual perception in virtual environments, 2019.
External examiners: Pavel Zahorik, Piotr Majdak

Vol. 40: *Niclas A. Janssen*, Binaural streaming in cochlear implant patients, 2019.
External examiners: Tim Jürgens, Hamish Innes-Brown

Vol. 41: *Wiebke Lamping*, Improving cochlear implant performance through psychophysical measures, 2019.
External examiners: Dan Gnasia, David Landsberger

Vol. 42: *Antoine Favre-Félix*, Controlling a hearing aid with electrically assessed eye gaze, 2020.
External examiners: Jürgen Tchorz, Graham Naylor

Vol. 43: *Raul Sanches-Lopez*, Clinical auditory profiling and profile-based hearing-aid fitting, 2020.
External examiners: Judy R. Dubno, Pamela E. Souza

Vol. 44: *Juan Camilo Gil Carvajal*, Modeling audiovisual speech perception, 2020.
External examiners: Salvador Soto-Faraco, Kaisa Maria Tippana

Vol. 45: *Charlotte Amalie Emdal Navntoft*, Improving cochlear implant performance with new pulse shapes: a multidisciplinary approach, 2020
External examiners: Andrew Kral, Johannes Frijns

Vol. 46: *Naim Mansour*, Assessing hearing device benefit using virtual sound environments, 2021.
External examiners: Virginia Best, Pavel Zahorik

Vol. 47: *Anna Josefine Munch Sørensen*, The effects of noise and hearing loss on conversational dynamics, 2021.
External examiners: William McAllister Whitmer, Martin Cooke

Vol. 48: *Thirsa Huisman*, The influence of vision on spatial localization in normal-hearing and hearing-impaired listeners, 2021.
External examiners: Steven van de Par, Christopher Stecker

Vol. 49: *Florine Lena Bachmann*, Subcortical electrophysiological measures of running speech, 2021.
External examiners: Samira Anderson, Tobias Reichenbach

Vol. 50: *Nicolai Fernández Pedersen*, Audiovisual speech analysis with deep learning, 2021.
External examiners: Hani Camille Yehia, Zheng-Hua Tan

The end.

To be continued...

It is well known that seeing a talker's face can improve the comprehension of auditory speech compared to listening without visual inputs. This is especially observed in noisy settings such as "cocktail-party" scenarios. However, there is a lack of knowledge of the audiovisual (AV) cues and how the two modalities are related. This thesis aimed to contribute to a better understanding of the relationship between auditory and visual cues created during speech production. The AV relationship was analyzed across thousands of speakers. This being possible due to recent advances in computer vision and data-driven approaches. Using canonical correlation analysis we identified two primary temporal ranges of envelope fluctuations related to facial motions across speakers. Using a self-supervised learning approach, we trained interpretable nonlinear neural networks to extract highly correlated AV features. Lastly, we presented an AV speech separation model that used visual cues to perform acoustic source separation.
Overall, this thesis provided new insights into how auditory and visual speech cues are related and showed their usefulness in AV speech separation.

DTU Health Tech
Department of Health Technology

Ørsteds Plads
Building 352
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 39 50
www.dtu.dk