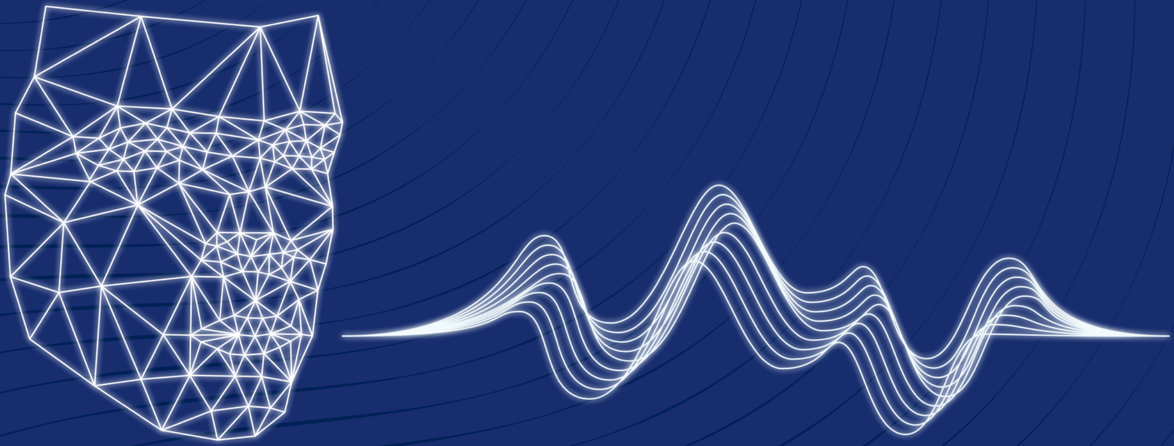
# Audio-Visual Speech Enhancement Based on Deep Learning

Michelsanti, Daniel



# AUDIO-VISUAL SPEECH ENHANCEMENT BASED ON DEEP LEARNING

BY
DANIEL MICHELSANTI

DISSERTATION SUBMITTED 2020

AALBORG UNIVERSITY
DENMARK

# Audio-Visual Speech Enhancement Based on Deep Learning

PhD Thesis

Daniel Michelsanti

2020

| | |
|---|---|
| Dissertation submitted: | August 31st, 2020 |
| PhD supervisor: | Professor Zheng-Hua Tan<br>Aalborg University, Denmark |
| Assistant PhD supervisor: | Professor Jesper Jensen<br>Aalborg University, Denmark<br>Oticon A/S, Denmark |
| PhD committee: | Associate Professor Flemming Christensen (chairman)<br>Aalborg University<br>Professor John H. L. Hansen<br>University of Texas<br>Research Scientist Xavier Alameda-Pineda<br>Université Grenoble Alpes |
| PhD Series: | Technical Faculty of IT and Design, Aalborg University |
| Department: | Department of Electronic Systems |

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-698-4

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2020

# About the Author

Daniel Michelsanti

Daniel Michelsanti received the B.Sc. degree in Computer Science and Electronic Engineering (cum laude) at the University of Perugia, Italy, and the M.Sc. degree in Vision, Graphics and Interactive Systems at Aalborg University, Denmark, in 2014 and 2017, respectively. He is currently a PhD fellow at the Department of Electronic Systems, Aalborg University, Denmark. His research interests are in the areas of multimodal speech enhancement and machine learning (specifically deep learning).

This page intentionally left blank.

# Abstract

*Speech communication* is often challenged by several sources of disturbance that surround a speaker and a listener involved in a conversation. One example is a cocktail party, in which a listener is immersed in an acoustically noisy environment, generally consisting of a target speaker, competing speakers, reverberations, and background noise. In this situation, two phenomena usually occur. On the one hand, the target speaker manifests a clear change in the way of speaking to maintain their speech intelligible; a tendency known as *Lombard effect*. On the other hand, the listener focuses their auditory attention on the speech of interest, while filtering out the other sounds; a phenomenon called *cocktail party effect*.

When the background noise level is sufficiently high and/or the listener is hearing impaired, the two aforementioned mechanisms do not guarantee an effective communication. The listener may benefit from using hearing aids, devices that, besides amplification, perform *speech enhancement*, which consists of extracting the speech of interest from a given degraded speech signal. Speech enhancement is traditionally addressed with techniques that consider only acoustic signals. However, important information can be extracted from the lip movements and the facial expressions of the target speaker, which are reliable cues even in presence of high levels of background noise. Therefore, speech enhancement systems that use both acoustic and visual information are able to outperform audio-only approaches.

In this thesis, we study the problem of *audio-visual speech enhancement based on deep learning* using one microphone and one camera. In particular, we propose a new taxonomy and perform an experimental analysis of training targets and objective functions for audio-visual speech enhancement systems based on deep learning. Furthermore, we investigate the impact of Lombard effect on a deep-learning-based speech enhancement approach from an acoustic and a visual perspective. Additionally, we propose a new algorithm to reconstruct the speech of interest from the silent video of the target speaker. Finally, we provide a systematic survey of audio-visual speech datasets, evaluation methods and audio-visual speech enhancement systems.

This page intentionally left blank.

# Resumé

*Talekommuniktation* bliver ofte besværliggjort af flere forskellige støjelementer, som omgiver taleren og lytteren involveret i en samtale. Et eksempel er et cocktail party, hvor lytteren bliver udsat for et akustisk støjfyldt miljø generelt bestående af en ønsket taler, konkurrende talere, rumklang og baggrundsstøj. I denne situation opstår der to fænomener. På den ene side udviser ønskede taleren en tydelig ændring af sin tale for at opretholde deres taleforståelse; en tendens kendt som *Lombardeffekten*. På den anden side fokuserer lytteren sin auditoriske opmærksomhed på den ønskede tale imens andre lyde bliver filtreret fra; et fænomen der kaldes *cocktailparty-effekten*.

Desværre når baggrundsstøjen er tilstrækkelig høj og/eller lytteren er hørehæmmet, garanterer de to førnævnte mekanismer ikke for en effektiv kommunikation. Lytteren kan få gavn af at bruge høreapparater, anordninger der er designet til at forstærke talesignal samt udføre *taleforbedring*, som består af udtrækkering en ønsket taler fra et givent forringet talesignal. Taleforbedringsteknologi bliver traditionelt håndteret med teknikker som kun betragter akustiske signaler. Dog kan vigtig information blive udtrukket fra læbebevægelser og ansigtsudtryk fra den ønskede taler, som er pålidelige signaler selv når niveauet af baggrundsstøjen er højt. Derfor kan taleforbedringssystemer som anvender både akustiske og visuelle informationer give bedre præstation hvis sammenlignet med kun lyddrevne metoder.

I denne afhandling studerer vi problemet *audiovisuel taleforbedring baseret på dyb læring* ved brug af én mikrofon og ét kamera. Særligt foreslår vi en ny taksonomi og udfører en eksperimentel analyse af træningsmål og kostfunktioner til audiovisuel taleforbedringssystemer baseret på dyb læring. Desuden undersøger vi påvirkningen af Lombardeffekten på taleforbedringsteknologi baseret på dyb læring fra et akustisk og et visuelt perspektiv. Derudover foreslår vi en ny algoritme til at rekonstruere den ønskede tale alene ud fra videobillederne uden lyd. Afslutningsvis foreslår vi en systematisk gennemgang af audiovisuelle taledatasæt, evalueringsmetoder og audiovisuel taleforbedringssystemer.

This page intentionally left blank.

# Contents

This page intentionally left blank.

# List of Abbreviations

| | |
|---|---|
| **AO** | Audio-Only |
| **AO-SE** | Audio-Only Speech Enhancement |
| **AO-SS** | Audio-Only Speech Separation |
| **ASR** | Automatic Speech Recognition |
| **AV** | Audio-Visual |
| **AVCDCN** | Audio-Visual Codebook Dependent Cepstral Normalisation |
| **AV-SE** | Audio-Visual Speech Enhancement |
| **AV-SS** | Audio-Visual Speech Separation |
| **BiLSTM** | Bi-directional Long Short-Term Memory |
| **CNN** | Convolutional Neural Network |
| **DNN** | Deep Neural Network |
| **DOA** | Direction Of Arrival |
| **ESTOI** | Extended Short-Time Objective Intelligibility |
| **F0** | Fundamental Frequency |
| **FFNN** | Feedforward Fully-connected Neural Network |
| **GMM** | Gaussian Mixture Model |
| **GRU** | Gated Recurrent Unit |
| **HMM** | Hidden Markov Model |
| **LSTM** | Long Short-Term Memory |
| **MAE** | Mean Absolute Error |
| **MLP** | Multilayer Perceptron |
| **MOS** | Mean Opinion Score |
| **MSE** | Mean Squared Error |

| | |
|---|---|
| **PESQ** | Perceptual Evaluation of Speech Quality |
| **PSD** | Power Spectral Density |
| **RNN** | Recurrent Neural Network |
| **SE** | Speech Enhancement |
| **SGD** | Stochastic Gradient Descent |
| **SNR** | Signal-to-Noise Ratio |
| **STFT** | Short-Time Fourier Transform |
| **STOI** | Short-Time Objective Intelligibility |
| **TF** | Time-Frequency |
| **VAD** | Voice Activity Detector |
| **VAE** | Variational Auto-Encoder |

# List of Publications

The Part II of this thesis consists of the following publications:

[A] **D. Michelsanti**, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "On Training Targets and Objective Functions for Deep-Learning-Based Audio-Visual Speech Enhancement", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 8077–8081, 2019.

[B] **D. Michelsanti**, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Effects of Lombard Reflex on the Performance of Deep-Learning-Based Audio-Visual Speech Enhancement Systems", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 6615–6619, 2019.

[C] **D. Michelsanti**, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-Learning-Based Audio-Visual Speech Enhancement in Presence of Lombard Effect", *Speech Communication*, vol. 115, pp. 38–50, 2019.

[D] **D. Michelsanti**, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen, "Vocoder-Based Speech Synthesis from Silent Videos", *Proceedings of Interspeech (to appear)*, 2020.

[E] **D. Michelsanti**, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation", Submitted to *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

Besides papers [A]-[E], the author also worked on three conference papers before the start of his PhD project. A list of them is reported below to showcase the additional research activities in which the author has been involved throughout his career.

[F] A. A. Sangüesa, A.-D. Ene, N. K. Jørgensen, C. A. Larsen, **D. Michelsanti** and M. Kraus, "Pyramid Algorithm Framework for Real-Time Image Effects in Game Engines", *Interactivity, Game Creation, Design, Learning, and Innovation*, pp. 289–296, 2016.

[G] **D. Michelsanti**, A.-D. Ene, Y. Guichi, R. Stef, K. Nasrollahi and T. B. Moeslund, "Fast Fingerprint Classification with Deep Neural Networks", *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5, pp. 202–209, 2017.

[H] **D. Michelsanti** and Z.-H. Tan, "Conditional Generative Adversarial Networks for Speech Enhancement and Noise-Robust Speaker Verification", *Proceedings of Interspeech*, pp. 2008–2012, 2017.

# Preface

This thesis documents the scientific work conducted as part of the PhD project "Audio-Visual Speech Enhancement Based on Deep Learning". The thesis is submitted to the Technical Doctoral School of IT and Design at Aalborg University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The project was carried out within the Centre for Acoustic Signal Processing Research (CASPR), at the Section for Signal and Information Processing, Department of Electronic Systems, Aalborg University, Aalborg, Denmark. Parts of the work were conducted during a secondment at the Image Processing Group, Department of Information & Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain.

The thesis consists of an extended introduction and a collection of scientific papers. The purpose of the introduction is to provide the reader with the fundamental coordinates to understand the research area, the objectives and the contributions of the project. The papers present the contributions in details, elaborating on the methods and discussing the results.

It is a common belief that doing a PhD is a solo journey. However, like every journey in life, it would not last long without the support of others. For this reason, the author would like to thank all the people who, more or less directly, contributed to this PhD project. First and foremost, the author is extremely grateful to Zheng-Hua Tan and Jesper Jensen, that were ideal supervisors: their motivation, support and guidance had a great impact not only on this work, but also on the author's view concerning the way to conduct research, which is something that he will keep with him for the rest of his life. Then, the author would like to thank Sigurdur Sigurdsson, for the useful insights that he provided during the first part of this project. Special thanks go to Gloria Haro, Emilia Gómez, and Olga Slizovskaia, for their hospitality and the successful collaboration during the author's stay at Universitat Pompeu Fabra. Furthermore, the author wish to extend his thanks to Dong Yu, Shi-Xiong Zhang, Yong Xu, and Meng Yu, for the very pleasant collaboration throughout the final part of this project. Finally, the author would like to thank his colleagues, his friends, and his family, who, in different ways, made this three-year-long journey even more enjoyable.

*"I had great ambitions and extravagant dreams, but so did the errand boy and the seamstress, for everyone has dreams; the only difference is whether or not we have the strength to fulfil them or a destiny that will fulfil them through us.*

*When it comes to dreams, I'm no different from the errand boy and the seamstress. The only thing that distinguishes me from them is that I can write. Yes, that's an activity, a real fact about myself that distinguishes me from them. But in my soul I am just the same."*

Bernardo Soares

F. Pessoa, *The Book of Disquiet*,
M. Jull Costa (translation)

# Part I

# Introduction

# Introduction

*Speech* is one of the preferred ways that we use to communicate with others. As such, speech communication has been the foundation of the progress of humanity, because it allows to share ideas, thoughts and feelings, which are the basis for spreading knowledge and building relationships. Given this premise, it is not a surprise that humans have always tried to make speech communication easier in challenging situations, taking advantage of technology. Examples include videoconference systems, which allow to have a conversation even though the conference attendees are separated by a long distance, and hearing aid systems, which are specifically designed to help people with a hearing loss. With this thesis we address a problem that is central for both videoconference and hearing aid systems: the potential degradation of the *speech of interest*, or *target speech*, due to *background noise*.

Consider the situation illustrated in Figure 1. Two people are having a conversation during a cocktail party. In the specific moment depicted in the figure, the person on the couch, the *target speaker*, is telling a story to the person on the armchair, the *listener*. Meanwhile, other people are having their own conversations in the immediate surrounding area and some background music is played by a harpist. Despite the presence of all these sources of disturbance, the human auditory system of the listener is able, with some effort, to selectively focus on what the target speaker is saying, while filtering

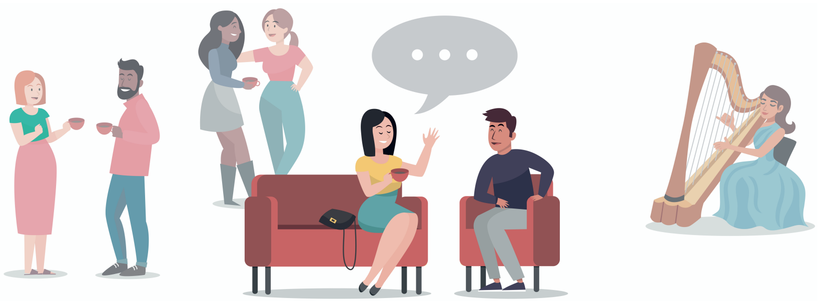


**Fig. 1:** Cocktail party. Some images designed by pikisuperstar and macrovector / Freepik.

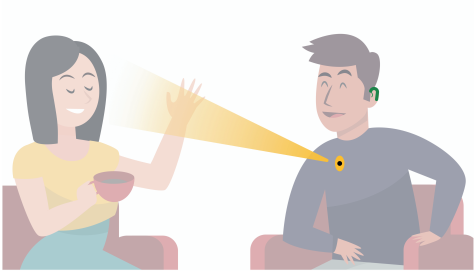


**Fig. 2:** The potential next-generation hearing aid system that we have in mind as an application of the work performed in this thesis. The system consists of a camera device (in yellow) and a traditional digital hearing aid system (in green) that are able to jointly process audio-visual information with the purpose of enhancing the speech of interest. Some images designed by pikisuperstar / Freepik.

out the other sounds [12, 112]. The problem of recognising the speech of interest in this scenario is known as the *cocktail party problem* [13].

Designing an automatic system to extract the speech of interest in a cocktail party scenario is important in many applications. Consider, for example, the case in which the listener in Figure 1 is hearing impaired. The background noise makes it extremely difficult for him to follow the conversation. One strategy that he could adopt is to focus his attention on the mouth region of the speaker and try to read her lips, because the movements of visible articulatory organs are immune to the acoustic noise. However, the listener would greatly benefit from using hearing aids, that could suppress all the unwanted sounds and deliver the clean target speech, i.e. perform Speech Enhancement (SE). With this thesis, we envision algorithms that, thanks to the flexibility of *deep learning*, could exploit visual information from the target speaker for SE and be employed in next-generation hearing aid systems (cf. Figure 2). In particular, we study, design and validate techniques that can be used in cocktail party scenarios. Furthermore, we provide a survey of relevant deep-learning-based Audio-Visual (AV) approaches in the literature.

In the rest of the Introduction, we review some basic concepts about speech science, deep learning and Audio-Visual Speech Enhancement (AV-SE) that are relevant to this thesis. Finally, we summarise the contributions of our work and identify some possible future research directions.

# 1 Speech Science

In this Section, we deal with fundamentals of *speech science*, which is the field of study concerning three aspects: *speech production*, *speech transmission* and *speech perception*. In the literature, the process involving these aspects is referred as *speech chain* [22, 39], illustrated in Figure 3.

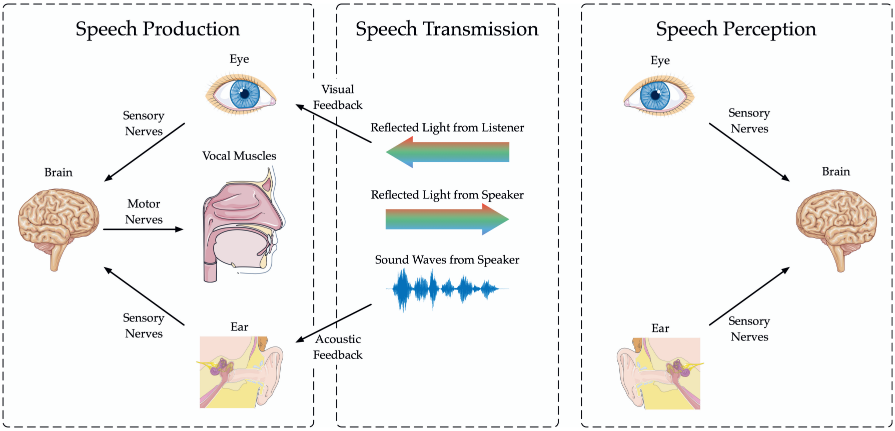


**Fig. 3:** Illustration of the speech chain, inspired by [22, 39]. Some images from Servier Medical Art Creative Commons Attribution 3.0 Unported License (`http://smart.servier.com`).

The speech chain is mainly a feed-forward process consisting of several stages [22]. The thoughts of a speaker are first converted into a linguistic structure by the speaker's brain, which then sends signals to vocal muscles with the goal of producing speech. After that, speech, generally consisting of an acoustic component (i.e. sound waves) and a visual component (i.e. mouth movements and facial expressions), is transmitted to a listener through a communication channel. At this point, the listener's sensory organs, specifically ears and eyes, capture the sound waves and the reflected light coming from the speaker and send this information to the brain, which converts it into meaning. As illustrated in Figure 3, the process involves also two feedback loops: the speaker monitors and adjusts their speech based on an acoustic feedback, provided by listening to their own voice, and a visual feedback, obtained by looking at the expressions of the listeners, who might have problems in understanding the speech.

## 1.1 Speech Production

### Fundamentals of Speech Production

Speech sounds may be grouped into two categories: *voiced* and *unvoiced* [21]. To produce voiced sounds, such as *vowels*, the airflow coming from the lungs sets the vocal folds into a vibratory state. The rate of the vibration is known as Fundamental Frequency (F0)[1] and it depends on the size and the tension of the vocal folds [21]. Then, the airflow passes through the pharyngeal, the

[1] Sometimes, F0 is called *pitch*, although the latter should be used to indicate the F0 perceived by a listener.

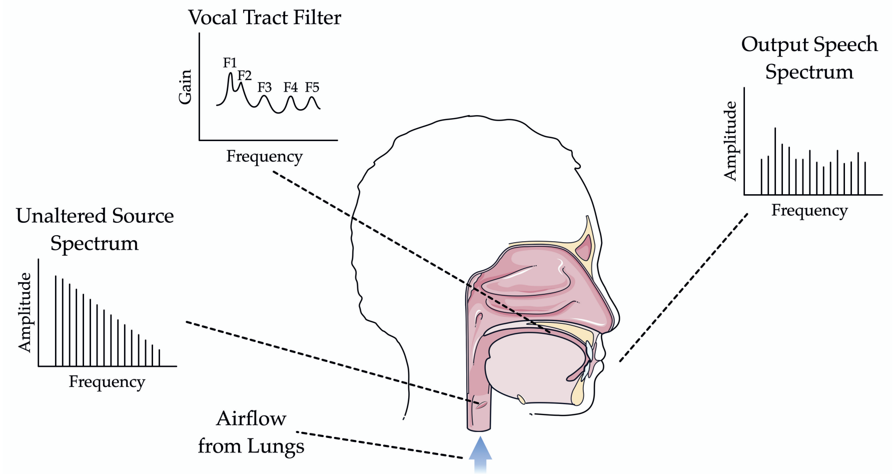


**Fig. 4:** Illustration of the source-filter model, inspired by [2, 67]. F1, F2, . . . , F5 indicate the formants. Some images from Servier Medical Art Creative Commons Attribution 3.0 Unported License (`http://smart.servier.com`).

oral and the nasal cavities, before being expelled through the mouth opening and the nostrils. On the other hand, unvoiced sounds are produced by either inducing a turbulence in the airflow with some constrictions in the vocal tract, in case of *fricatives*, like /f/, and /s/, or suddenly releasing an air pressure caused by a complete closure of some parts of the vocal tract, in case of *plosives*[2], such as /p/, /t/, and /k/.

The production of the acoustic component of speech is often modelled within a linear framework known as *source-filter model* [27] (cf. Figure 4). In this model, the sound produced by the vibration of the vocal folds represents an *excitation signal*, or *source*, whose *spectrum*, i.e. a representation of the frequency content of the signal, is harmonically rich. When the sound propagates through the vocal tract, the spectrum changes according to a transfer function, which, in frequency domain, is characterised by several local maxima, known as *formants*, and local minima, known as *antiformants*. Therefore, the final speech spectrum can be seen as a filtered version of the source spectrum. The source-filter model has the potential disadvantage of making assumptions that might not be valid. For example, it does not model the interaction between source and vocal tract [71, 122]. However, it is sufficiently good in most applications.

Since speech is dynamic, its changes over time cannot be captured in details by the speech spectrum. Therefore, a Time-Frequency (TF) representation known as *spectrogram* is often used (cf. Figure 5). A spectrogram allows

[2] There are also voiced fricative and plosive sounds, like /z/ and /b/.

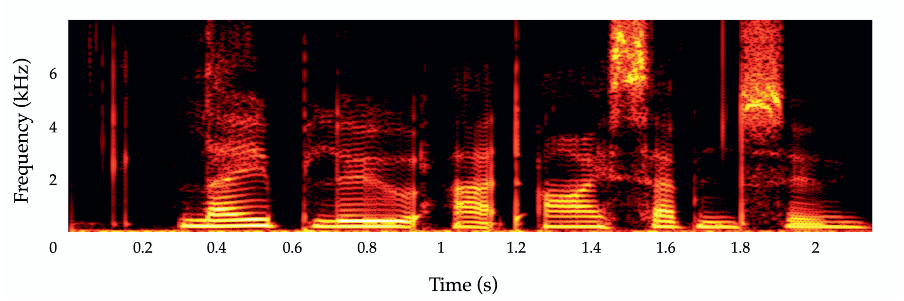


**Fig. 5:** Spectrogram of the sentence "lay blue at i seven soon" pronounced by a female speaker in the GRID dataset [17].

to visually inspect how the frequency characteristics of the speech signal (in particular formants and *harmonics*) varies in time.

Besides an acoustic component, speech has also a visual component. During a conversation, some articulatory organs of the speaker are hit by the light from the surrounding environment, which can be seen as an electromagnetic wave. As a consequence, the articulatory organs of the speaker absorb the light of certain wavelengths and reflect the rest. This process can be referred as production of visual speech.

### Speech Analysis and Synthesis with Vocoders

Throughout the years, several techniques have been developed to emulate the speech production process with a machine, for example by using *vocoders* [24, 111]. A vocoder (a word derived from the fusion of VOice and CODER [111]) is a system consisting of an analysis step, which decomposes a speech signal into parameters that describe some relevant aspects of speech, and a synthesis step, which is used to reconstruct the original signal with the parameters estimated in the analysis stage. A vocoder is useful in a range of different applications, such as transmission [111], synthesis [66] and modification [4, 66] of speech.

In our work, we use a vocoder called WORLD [91] to reconstruct speech from the silent video of a talking face (cf. paper [D] in the Part II of this thesis for a more detailed discussion of the approach). WORLD is based on a source-filter model of speech production with mixed excitation [84] and allows to perform high-quality speech synthesis in real time [90, 91]. Specifically, the analysis module of WORLD decomposes a speech signals into the following three time-varying parameters:

- F0, which we have already mentioned when we talked about vocal folds vibration.

- Spectral envelope, which can be seen as a smooth curve that links the peaks of the speech magnitude spectrum.

- Aperiodicity, defined as the power ratio between the speech signal and its aperiodic component [89].

The synthesis module of WORLD is fed with a sequence of the previously mentioned parameters and produces a speech waveform as output.

## 1.2 Speech Transmission

In order to reach the listener, either one or both acoustic and visual components of speech have to propagate through a communication channel. In a generic conversation setting, like the one illustrated in Figure 1, the speaker and the listener are close to each other. Therefore, the sound waves and the reflected light coming from the speaker reach the listener through air. However, even in this simple setting, speech can be impacted acoustically by background noise and reverberations, and visually by poor illumination and occlusions.

Sometimes, we would like the communication channel to modify the acoustic signal to improve the listener experience. A good example is the one illustrated in Figure 2, where the listener is wearing a next-generation hearing aid system to compensate for their hearing loss. This situation is more complex than the previous one, because the acoustic and the visual components of speech follow two paths: on the one hand, they reach the listener, as observed before; on the other hand, they also pass through the hearing aid device which processes acoustic and visual information and delivers an enhanced and hearing-loss compensated acoustic signal to the listener's ears.

Other AV common scenarios are also possible. If the people involved in a conversation are separated by a long distance, then the communication may occur via a video call. In this case, the acoustic and visual signals are captured by cameras and microphones and transmitted to the listener's device, which has a screen and a loudspeaker. Similarly, the audio and the video from the listener are transmitted to the speaker's device, providing feedback information. During the signal transmission, several operations take place, such as analog-to-digital and digital-to-analog conversions, speech and video coding, and network packets transmission. Issues occurring in any of these stages might cause an information loss, lowering the reliability of the communication channel.

In this thesis, our focus is on algorithms that can be applied in devices constituting part of the communication channel, with the final goal of attenuating the background noise and improving the speech understanding.

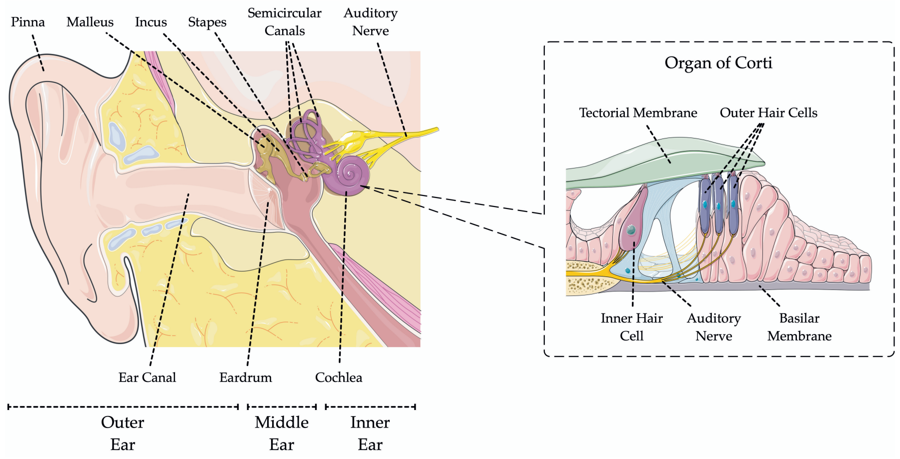


**Fig. 6:** Cross section of the human ear. Some images from Servier Medical Art Creative Commons Attribution 3.0 Unported License (`http://smart.servier.com`).

## 1.3 Speech Perception

### Fundamentals of Speech Perception

The two human sensory organs responsible for the reception of the acoustic and the visual components of speech are the *ear* and the *eye*, respectively. Here, we present some basic elements of the physiology of ear and eye along with some fundamentals of AV speech perception. Further details can be found in [19, 83, 97].

The human auditory periphery may be divided into three parts (cf. Figure 6): *outer*, *middle* and *inner* ear. The sound waves coming from the speaker first hit the *pinna*, which reflects them towards the *ear canal* in a way that helps the listener to determine the speech direction of arrival [11]. The configuration of the ear canal amplifies the frequencies around 3 kHz, which are important in speech perception [102]. The outer ear ends with the *eardrum*, which has the role of transforming the motion of acoustic waves into mechanical vibrations. Then, the middle ear, consisting of three ossicles, namely *malleus*, *incus*, and *stapes*, transfers the mechanical vibrations to the inner ear. This transfer is most efficient between 0.5 and 5 kHz, as reported in [5, 101]. In the inner ear, the *cochlea* converts the mechanical movements from the middle ear into electrical signals, which propagate towards the brain via the *auditory nerve*. Specifically, the cochlea has a spiral shape filled with fluids and is divided into three tubings by the *Reissner's membrane* and the *basilar membrane*. When the stapes moves, a pressure difference inside the cochlea sets the basilar membrane into movement. Given the mechanical properties of the basilar membrane, sounds at different frequencies cause displacements

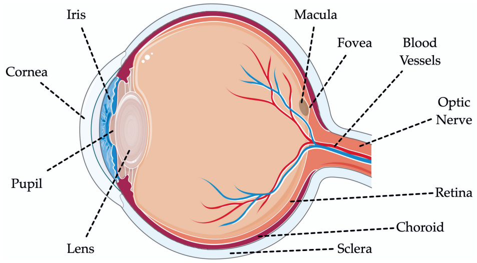


**Fig. 7:** Cross section of the human eye. Some images from Servier Medical Art Creative Commons Attribution 3.0 Unported License (`http://smart.servier.com`).

of the membrane at different locations: high frequency sounds generate basilar membrane fluctuations close to the base of the cochlea, i.e. the part that is in contact with the stapes, while low frequency sounds generate fluctuations close to the apex, i.e. the other end of the cochlea. Because of this property, the cochlea is often modelled as a bank of bandpass filters [88]. The mechanical movements of the basilar membrane are converted into neural activity by the *inner hair cells*, which are part of the so-called *organ of Corti* (cf. right side of Figure 6). This organ also contains *outer hair cells*, which have a motor function influencing the mechanics of the cochlea [88].

The elements and the mechanisms described above allow humans to perceive sounds having a frequency between 0.02 and 20 kHz with a large dynamic range of about 120 dB [102]. Damages or problems concerning one or more parts of the auditory system might impact the perception of sounds. In general, we distinguish between problems of the outer or middle ear, causing a *conductive hearing loss*, and problems of the inner ear, causing a *sensorineural hearing loss* [102]. External devices, such as hearing aids, can reduce the discomfort of conductive hearing loss by amplifying specific frequencies of a speech signal. While amplification can also reduce the discomfort of mild to medium sensorineural hearing loss, other invasive treatments may be necessary for more severe conditions. For example, if the cochlea is damaged, a cochlear implant may be used to mimic the spectral decomposition of the cochlea and provide electrical signals to the auditory nerve.

While the human ear is a receptor of sounds, the human eye (cf. Figure 7) is the sensory organ used for vision, and in our case it is relevant for the perception of the visual component of speech. When the reflected light from the speaker reaches the listener, it is refracted by the front part of the eye, known as *cornea*, and passes through a hole called *pupil*. The pupil is surrounded by the *iris*, whose role is to control the size of the pupil depending on the intensity of the light. Behind the pupil, a *lens* further refracts the light to maintain the focus on an object although its distance from the eye changes. At this point, the light reaches the *retina* which is a layer that converts the image projected on its surface into electrical signals. In particular, the retina consists of two kinds of photoreceptor cells: *rods* and *cones* [19]. Rods populate mostly the periphery of the retina and allow vision at low levels of light. On the other hand, cones are in the central part of the retina, the *fovea*, and are divided into three types: short or blue, middle or green, and long or red based on the wavelength of light which they are sensitive to [102]. The presence of three types of cones allows the brain to interpret different wavelengths of light as colours. When the light hits rods and cones, it triggers a chemical reaction that generates electrical impulses. These impulses are transmitted by the *optic nerve* to the brain, which interprets them and allows humans to perceive a scene in the so-called *visible spectrum*, which is the portion of the electromagnetic spectrum ranges from approximately 400 to 700 nm [87]. Problems concerning the visual system might result in a visual impairment. Some medical conditions that impact vision, like myopia, can be easily addressed with external corrective lenses, but in other cases, such as retinal detachment, more invasive treatments are required.

Some situations that we experience in real life provide anecdotal evidence of the importance of vision in speech perception [83]. For example, when we watch dubbed movies, we can feel some discomfort due to the mismatch between the lip movements of the speakers and the acoustic stimulus that reaches our ears. Researchers have investigated more thoroughly the role of vision in speech perception throughout the years [83]. In particular, Sumby and Pollack [116] conducted several experiments in which subjects were presented with Audio-Only (AO) or AV stimuli representing words in noise picked from different vocabulary sets and were instructed to select the words that they perceived from a given list. The most important finding from the Sumby and Pollack's work was that the relative contribution of the visual information to the word recognition performance[3] was essentially independent of the Signal-to-Noise Ratio (SNR), although the absolute contribution of vision decreased when the SNR increased. Another study [85] showed that vision affected speech perception even when the acoustic component was not degraded by background noise. Specifically, when exposed to stimuli that had a mismatch between the acoustic and the visual components, the listener tended to perceive a plausible compromise. For example, if the acoustic component of a stimulus was [ba], while the visual component was [ga], then the listener perceived the entire stimulus as [da]. This perceptual phenomenon is known as *McGurk effect*.

Given the importance of vision, in this thesis we investigate systems that incorporate visual information to process speech signals and deliver acoustic signals to a human receiver.

[3] Here, the relative contribution of visual information is defined as the ratio between the actual contribution of vision and the maximum achievable improvement over AO performance.

### Speech Quality and Intelligibility Estimation

In order to assess the performance of speech processing approaches, like the ones that we study in this thesis, researchers usually refer to two perceptual aspects: *speech quality* and *speech intelligibility*. Speech quality consists of all the attributes concerning *how* a speaker produces an utterance, such as naturalness [21, 79]. On the other hand, speech intelligibility concerns *what* a speaker says, i.e. the content of an utterance [79].

Since quality and intelligibility are perceptual aspects, ideally, we would like to involve humans in the assessment of speech signals. Many paradigms to perform *subjective listening tests* have been proposed [21, 79]. In parts of our work, we conduct experiments with panels of listeners to assess AV signals (cf. paper [C]). However, listening tests are often costly and time consuming. Therefore, researchers have developed algorithms, known as *objective measures*, that bypass a listening test with human subjects and estimate speech quality and intelligibility. In the following, we present the objective measures that we use in our studies.

The Perceptual Evaluation of Speech Quality (PESQ) measure [61–63, 104] is the most widely used estimator of speech quality. Although it was originally designed to evaluate speech coding approaches used in telephone networks, more recently, PESQ has been shown to be a reliable measure also to assess the overall quality of speech signals processed with common SE systems [56]. In particular, PESQ approximates the results of a Mean Opinion Score (MOS) test [58–60], which is a listening test where human subjects are asked to rate speech signals using a five-point discrete scale. In order to estimate the human response to an acoustic stimulus, PESQ considers some basic psychoacoustic principles, such as:

- The non-uniform frequency resolution of the human auditory system [115].
- The non-linear human loudness perception [133].
- Masking effects, which might prevent the perception of weak sounds [35].

The Short-Time Objective Intelligibility (STOI) measure [118] is used to estimate the speech intelligibility of an acoustic signal. It computes a correlation coefficient between the short-time overlapping temporal envelope segments of the clean speech signal and the ones of the processed signal. Several studies have shown that STOI correlates well with the results of listening experiments, e.g. for speech processed by SE algorithms [26, 118, 130]. Recently, Extended Short-Time Objective Intelligibility (ESTOI) [64] was proposed to more accurately handle speech degraded by highly modulated noise sources.

## 1.4 Lombard Effect

As we have previously described for the cocktail party scenario, speech communication may occur in presence of background noise. In this case, it is possible to observe changes in the speech production of the speaker, a phenomenon called *Lombard effect*, in honour of Étienne Lombard, the French otolaryngologist who first reported it in 1911 [80].

There is a general consensus in assuming that the two following processes are involved in Lombard effect [74]:

- An *internal* (or *private*) *loop* occurs when the speaker changes their way of speaking, i.e. their *speaking style*, as a consequence of hearing their own speech in noise.
- An *external* (or *public*) *loop* takes place when the speaker regulates their speech based on the feedback received from the listener.

Although Lombard effect manifests in lots of different ways from speaker to speaker [65, 82], several studies [32, 49, 65, 98, 117] identified some common traits that characterise the change from neutral to Lombard speech. First, speakers tend to increase their vocal levels based on the noise level [65, 99, 117]. This energy increment affects especially higher frequencies in the speech spectrum [117]. In addition, increases in F0 have been reported [65, 117] along with changes in formant centre frequencies. In particular, mean F1 frequencies generally increase [65, 98, 117], while F2 frequencies show smaller changes [117] and may either increase [65] or decrease [98]. Another general tendency is an overall increment in word duration [50, 65, 114], which has been reported to be influenced by the linguistic content: words that are semantically relevant in a sentence appear more elongated if compared with other words [96]. Changes due to Lombard effect can be observed also for the visual component of speech: generally, the speaker exhibits larger face and head movements [33, 34, 123].

All these characteristics have an impact on speech perception. Specifically, studies showed that changes caused by Lombard effect in acoustic and visual parameters have a general benefit on speech intelligibility, even when the energy of neutral and Lombard speech is normalised [23, 29, 30, 69, 99, 117]. Since humans are better at recognising Lombard speech, one might think that better performance can be achieved also by automatic systems. However, several works reported a performance degradation of Automatic Speech Recognition (ASR) systems when they did not take Lombard effect into account [51, 65, 81, 82, 114].

Given the impact that Lombard effect might have on systems deployed at low SNRs, in parts of our work (cf. papers [B] and [C]) we study the influence that a mismatch between Lombard and neutral speech has on deep-learning-based AV-SE.

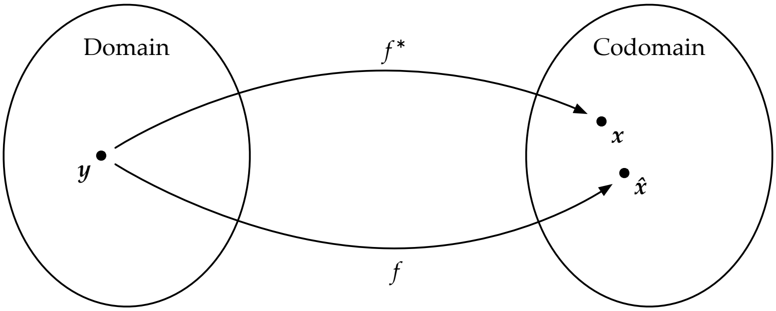


**Fig. 8:** Assuming that a mapping function, $f^*$, from an input vector, $\boldsymbol{y}$, to an output vector, $\boldsymbol{x}$, exists, in a regression problem we want to find a function $f$ which approximates $f^*$ sufficiently well for a specific application.

# 2 Deep Learning

In this Section, we present some basic concepts of deep learning, since the methods used in this thesis are based on it. For further details on the topic, the reader can refer to e.g. [47].

## 2.1 Fundamentals of Deep Learning for Regression Problems

Deep learning is a family of machine learning techniques that can be used to solve a wide range of different problems. In the particular context of this thesis, we use deep learning for *regression* (cf. Figure 8), i.e. to approximate a mapping function, $f^*$, between an input vector[4], $\boldsymbol{y}$, and an output vector, $\boldsymbol{x}$, given a set of input-output pairs, $D$, known as *training set*. A deep learning model is characterised by a set of parameters, $\theta$, that need to be learned to best approximate $f^*$, also for samples not belonging to the training set, an ability known as *generalisation* to unobserved samples, often achieved by using a large training set.

The process of updating the parameters $\theta$ is known as *training phase*, and it consists of finding a solution to the following optimisation problem [72]:

$$\theta^* = \underset{\theta}{\arg\min} \sum_{D} J(f(\boldsymbol{y},\theta),\boldsymbol{x}), \quad \text{with } (\boldsymbol{y},\boldsymbol{x}) \in D, \tag{1}$$

where $f(\boldsymbol{y},\theta)$ represents the deep learning model and $J(\cdot,\cdot)$ is an *objective function*, which is used to measure the distance between the output of the model, $\hat{\boldsymbol{x}} = f(\boldsymbol{y},\theta)$, and the so-called *training target*, $\boldsymbol{x}$. An approximation of the solution to the problem of Equation (1) is generally obtained with Stochastic Gradient Descent (SGD) [68, 105] or one of its variants, such as

[4] We use this notation to be aligned with the signal model that we will introduce in Section 3.1.

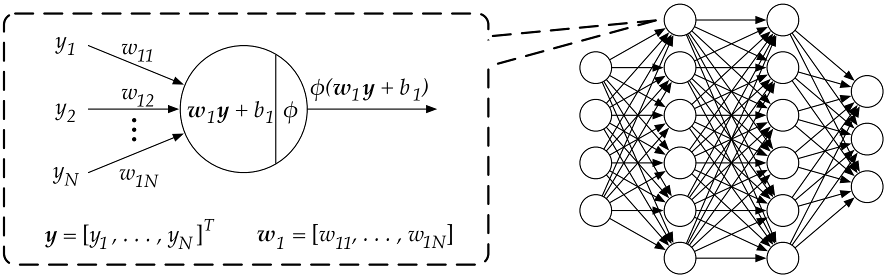


**Fig. 9:** Illustration of a multilayer perceptron with two hidden layers. On the left, there is a representation of the perceptron, the basic unit of a deep learning model. Inspired by [78].

Adam [70], which iteratively update the parameters $\theta$ using the *backpropagation* algorithm to compute the gradient [107, 126].

The particular choice of training target and objective function may be critical for the performance of a deep-learning-based system. Therefore, in paper [A], we perform a study of these two elements for the problem of our interest, i.e. AV-SE.

## 2.2 Deep Learning Architectures

Throughout the years, several deep learning architectures have been proposed. Here, we present the ones used in this thesis.

### Multilayer Perceptron

The basic unit of a deep learning model is the *perceptron* [106], also called *node* or *artificial neuron*, because it coarsely models a biological neuron. A perceptron (cf. left side of Figure 9) can be expressed as:

$$\hat{x}_1 = f(\boldsymbol{y}, \theta) = \phi(\boldsymbol{w}_1 \boldsymbol{y} + b_1), \quad \text{with } \theta = \{\boldsymbol{w}_1, b_1\}, \tag{2}$$

where $\phi(\cdot)$ is a, usually non-linear, function called *activation function*, $\boldsymbol{w}_1$ is a vector of *weights* applied to the input, and $b_1$ is an additional parameter, called *bias*, that allows the model to best fit the data.

If we want to learn a mapping function between an input vector $\boldsymbol{y}$ and an output vector $\boldsymbol{x}$ having more than one coordinate, we can extend Equation (2) obtaining the general formulation of a single-layer perceptron:

$$\hat{\boldsymbol{x}} = f(\boldsymbol{y}, \theta) = \phi(\boldsymbol{W}\boldsymbol{y} + \boldsymbol{b}), \quad \text{with } \theta = \{\boldsymbol{W}, \boldsymbol{b}\}, \tag{3}$$

where the matrix of weights, $\boldsymbol{W}$, and the bias vector, $\boldsymbol{b}$, are the parameters of the model. The main limitation of a perceptron is that its adaptive part is a linear model. Therefore, it cannot learn a non-linear transformation of the

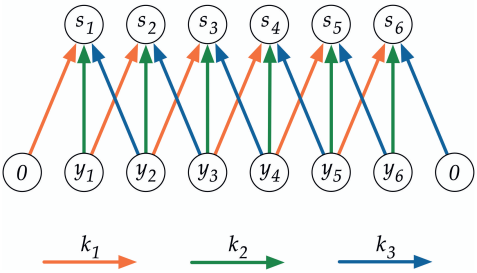


**Fig. 10:** Illustration of a convolutional layer for a kernel having 3 elements. Zero-padding is applied to the input vector.

input vector. One way to overcome this problem is to stack multiple perceptrons to obtain an architecture known as Multilayer Perceptron (MLP) [47] or Feedforward Fully-connected Neural Network (FFNN). A MLP (cf. right side of Figure 9) can be expressed as [72]:

$$\hat{\boldsymbol{x}} = f^{(P)}(\cdots f^{(2)}(f^{(1)}(\boldsymbol{y}, \theta_1), \theta_2) \cdots, \theta_P), \tag{4}$$

where $f^{(p)}(\cdot, \theta_p)$ indicates a single-layer perceptron with parameters $\theta_p$, and $P$ denotes the number of layers of the model, i.e. its *depth*. It has been shown that choosing $P = 2$ is sufficient to obtain a *univeral approximator*, meaning that a MLP can approximate any function to any degree of accuracy [53]. In practice, the number of layers is often increased obtaining Deep Neural Networks (DNNs), models whose hierarchical structure allows to provide representations of the data at different levels of abstraction and learn the desired mapping function more efficiently.

### Convolutional Neural Networks

In a MLP each node shares a connection with each and every node of the previous layer. This means that, if the dimensionality of the input is high, the model needs a very large number of parameters. Convolutional Neural Networks (CNNs) [77] are models that use *convolution*[5] in one or more of their layers and are able to deal with high dimensional data very efficiently.

Given an input vector $\boldsymbol{y} = [y_1, y_2, \cdots, y_N]$ and a weighting vector $\boldsymbol{k} = [k_{-M}, k_{-M+1}, \cdots, k_{M-1}, k_M]$, often called *kernel* or *filter*, with $N > 2M + 1$, a convolutional layer[6] in a neural network (cf. Figure 10) can be seen as a

[5] The operation implemented in a CNN is usually the cross-correlation, not the convolution. The two terms are often used interchangeably, and here we adopt the same convention.

[6] A convolutional layer is generally obtained by applying multiple kernels to the same input vector.

vector $\boldsymbol{s} = [s_1, s_2, \cdots, s_N]$, where each element,

$$s_i = \phi\Bigg(\underbrace{\sum_{m=-M}^{M} y_{i+m} k_m}_{\text{Convolution}} + b\Bigg), \tag{5}$$

with $b$ denoting the bias, is essentially a perceptron. If we consider Equation (5), not every $s_i$ is defined (for example, if we assume $M = 1$, $y_0$ and $y_{N+1}$ are undefined). Therefore, a *zero-padding* of the input vector is sometimes applied, i.e. some zero-valued samples are appended at the beginning and at the end of $\boldsymbol{y}$. It is straightforward to extend the definition of Equation (5) to the two-dimensional case, e.g. to use images as input of the model.

Due to the nature of convolution, each node of a layer is connected only with a small[7] neighbourhood of nodes of the previous layer (cf. Figure 10). In addition, since many nodes share the same kernel, CNNs are *translation equivariant* [47] and allow to considerably reduce the number of model parameters if compared to MLPs.

Another operation that is often performed in CNNs is *pooling*. The main idea of pooling is to make the model invariant to some transformations of the input by downsampling the output of a convolutional layer. While this property may be desirable for some tasks, such as detection, in other cases, e.g. when we want to preserve the structure of the input, pooling might be detrimental because it determines an information loss.

### Recurrent Neural Networks

While MLPs and CNNs are feed-forward neural network architectures, since the information flows in one direction from the input to the output layers, Recurrent Neural Networks (RNNs) introduce feedback connections. Due to their structure, RNNs are particularly suitable to process sequential data. Let $\boldsymbol{y}^{(t)}$ indicate the input vector at time $t$. Then, a RNN can be defined as [47]:

$$\boldsymbol{h}^{(t)} = \phi_h(\boldsymbol{W}\boldsymbol{y}^{(t)} + \boldsymbol{V}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}), \tag{6}$$

$$\hat{\boldsymbol{x}}^{(t)} = \phi_x(\boldsymbol{U}\boldsymbol{h}^{(t)} + \boldsymbol{c}), \tag{7}$$

where $\boldsymbol{W}$, $\boldsymbol{V}$, and $\boldsymbol{U}$ are the weight matrices, $\boldsymbol{b}$ and $\boldsymbol{c}$ are the bias vectors, and $\phi_h(\cdot)$ and $\phi_x(\cdot)$ are two activation functions. The vector $\boldsymbol{h}^{(t)}$ represents the *state*, also known as *memory*, of the model.

In practice, RNNs as defined in Equations (6) and (7) are rarely used. The reason is that they are generally affected by the *vanishing gradient problem* [47], i.e. the backpropagated gradient may tend to zero. Therefore, gated RNNs

[7] Here, we assume that the number of elements of the kernel is way smaller than the number of elements of the input.

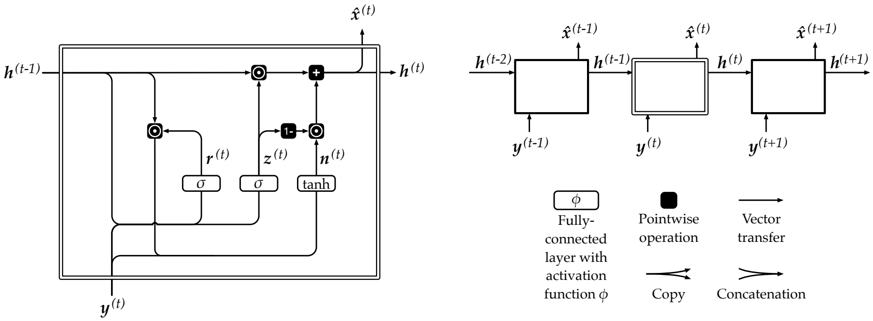


**Fig. 11:** Illustration of a gated recurrent unit. Inspired by the LSTM illustration by Colah [93].

were introduced. Here, we consider Gated Recurrent Units (GRUs) [14], since we use them in our work (cf. paper [D]). However, other gated RNN architectures, like Long Short-Term Memory (LSTM) [38, 52], are also popular.

In the following, we report the update equations for a GRU (cf. Figure 11):

$$\boldsymbol{r}^{(t)} = \sigma(\boldsymbol{W}_{\text{res}}\boldsymbol{y}^{(t)} + \boldsymbol{V}_{\text{res}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_{\text{res}}), \tag{8}$$

$$\boldsymbol{z}^{(t)} = \sigma(\boldsymbol{W}_{\text{up}}\boldsymbol{y}^{(t)} + \boldsymbol{V}_{\text{up}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_{\text{up}}), \tag{9}$$

$$\boldsymbol{n}^{(t)} = \tanh(\boldsymbol{W}_{\text{cs}}\boldsymbol{y}^{(t)} + \boldsymbol{V}_{\text{cs}}(\boldsymbol{r}^{(t)} \odot \boldsymbol{h}^{(t-1)}) + \boldsymbol{b}_{\text{cs}}), \tag{10}$$

$$\boldsymbol{h}^{(t)} = (1 - \boldsymbol{z}^{(t)}) \odot \boldsymbol{n}^{(t)} + \boldsymbol{z}^{(t)} \odot \boldsymbol{h}^{(t-1)}, \tag{11}$$

$$\hat{\boldsymbol{x}}^{(t)} = \boldsymbol{h}^{(t)}, \tag{12}$$

where: $\boldsymbol{W}_{\text{res}}$, $\boldsymbol{W}_{\text{up}}$, $\boldsymbol{W}_{\text{cs}}$, $\boldsymbol{V}_{\text{res}}$, $\boldsymbol{V}_{\text{up}}$, and $\boldsymbol{V}_{\text{cs}}$ are weight matrices; $\boldsymbol{b}_{\text{res}}$, $\boldsymbol{b}_{\text{up}}$, and $\boldsymbol{b}_{\text{cs}}$ are bias vectors; $\sigma(\cdot)$ and $\tanh(\cdot)$ are sigmoid and hyperbolic tangent activation functions, respectively; "$\odot$" denotes the element-wise multiplication. The main difference between a GRU and a vanilla RNN is that the former uses two gates to reset or update the state vector [47]. In particular, the *reset gate*, $\boldsymbol{r}^{(t)}$, determines whether the old state, $\boldsymbol{h}^{(t-1)}$, should be used to compute a candidate of the next state, $\boldsymbol{n}^{(t)}$. The *update gate*, $\boldsymbol{z}^{(t)}$, on the other hand, controls which impact the candidate of the next state should have on the old state in order to get the new state, $\boldsymbol{h}^{(t)}$. With this architecture, the gradient would not vanish, because, with the update rule of Equation (11), it can flow without changes through the network.

# 3 Audio-Visual Speech Enhancement

In this Section, we present in more details the main problem that we address in this thesis: AV-SE. Specifically, we first define a signal model and formulate the problem formally. Then, we provide an overview of the AV-SE literature, covering representative works in the field.

## 3.1 Signal Model and Problem Formulation

Let $x[n]$ and $d[n]$ denote the target clean speech signal and an additive noise signal, respectively, with $n$ indicating a discrete-time index. The observed acoustic noisy speech signal can be modelled as:

$$y[n] = x[n] + d[n]. \tag{13}$$

The task of single-microphone Audio-Only Speech Enhancement (AO-SE) consists of determining an estimate, $\hat{x}[n]$, of $x[n]$, given only $y[n]$. In the case of single-microphone single-camera AV-SE, $x[n]$ is estimated from $y[n]$ and an additional two-dimensional visual signal, $v[m]$, with $m$ indicating a discrete-time index usually different from $n$, since acoustic and visual signals are generally sampled with different sampling rates. Sometimes, $y[n]$ is not accessible, hence $x[n]$ is estimated solely from $v[m]$. In this particular case, we talk about *speech reconstruction from silent videos*.

The signal model of Equation (13) is often expressed in the TF domain as:

$$Y(k,l) = X(k,l) + D(k,l), \tag{14}$$

where $k$ and $l$ denote a frequency bin index and a time frame index, respectively, while $Y(k,l)$, $X(k,l)$ and $D(k,l)$ indicate the Short-Time Fourier Transform (STFT) coefficients of the noisy speech, the target clean speech, and the noise signals, respectively. In this case, the goal of SE is to determine an estimate, $\hat{X}(k,l)$, of $X(k,l)$. Since STFT coefficients are complex-valued, an estimate of $X(k,l)$ requires to approximate the magnitude, $|X(k,l)|$, and the phase, $\angle X(k,l)$. The short-time target phase is usually considered less important than the short-time target magnitude [36, 37, 125]. As a consequence, most SE approaches focus on estimating only $|X(k,l)|$ and use $\angle Y(k,l)$ to reconstruct the time-domain enhanced speech signal. Exceptions exist: for example, researchers have proposed approaches to provide an estimate of $\angle X(k,l)$ [3], $X(k,l)$ [25] or $x[n]$ [128].

In some situations, the observed acoustic signal, may be a mixture consisting of different speech signals uttered by multiple speakers and an additive background noise signal. For example, for two target speech signals, $x_1[n]$ and $x_2[n]$, and an additive background noise signal, $d[n]$, the mixture can be modelled as:

$$y[n] = x_1[n] + x_2[n] + d[n]. \tag{15}$$

The task of separating the speakers in the mixture, i.e. extracting all the existing speech signals, $x_1[n]$ and $x_2[n]$ in the case of Equation (15), given $y[n]$, is known as Audio-Only Speech Separation (AO-SS). Like in SE, if some additional visual information is provided, then we deal with Audio-Visual Speech Separation (AV-SS). Figure 12 illustrates the difference between AV-SE, speech reconstruction from silent videos and AV-SS.

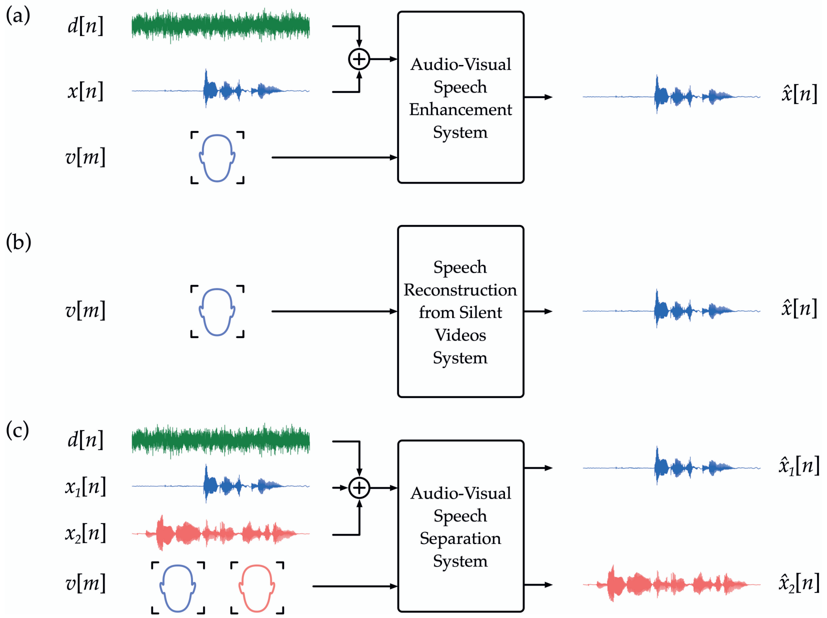

(a)
$d[n]$
$x[n]$
$v[m]$
Audio-Visual Speech Enhancement System
$\hat{x}[n]$
(b)
$v[m]$
Speech Reconstruction from Silent Videos System
$\hat{x}[n]$
(c)
$d[n]$
$x_1[n]$
$x_2[n]$
$v[m]$
Audio-Visual Speech Separation System
$\hat{x}_1[n]$
$\hat{x}_2[n]$

**Fig. 12:** Illustration of the difference between: (a) audio-visual speech enhancement; (b) speech reconstruction from silent videos; (c) two-speaker audio-visual speech separation.

## 3.2 Literature Review

Throughout the years, several approaches have been proposed and investigated to solve the problem of AV-SE. Initially, AV-SE systems were designed exploiting specific *domain knowledge* [7, 44] and, because of this, we refer to them as *knowledge-based approaches*. These approaches often assumed that the target speech and the noise signals were independent of each other and distributed according to known probabilistic distributions. These assumptions allowed to build models that worked well for specific situations, but were able to provide only marginal improvements in more complex scenarios.

Recently, AV-SE research switched from the design and the investigation of knowledge-based systems to the study of *deep-learning-based approaches*. This transition, evident from the timeline illustrated in Figure 13, was driven by the success of deep learning in several fields, like image classification [73], face verification [120], and board games [113], and occurred due to the availability of suitable computational resources and large-scale datasets. As reported in Section 2, the main idea behind deep learning is to train complex models, potentially consisting of millions of parameters, that act as universal approximators without necessarily relying on assumptions from a specific domain. Since deep learning models learn to perform a task from data, it is crucial to use large-scale training datasets with high enough variability to be

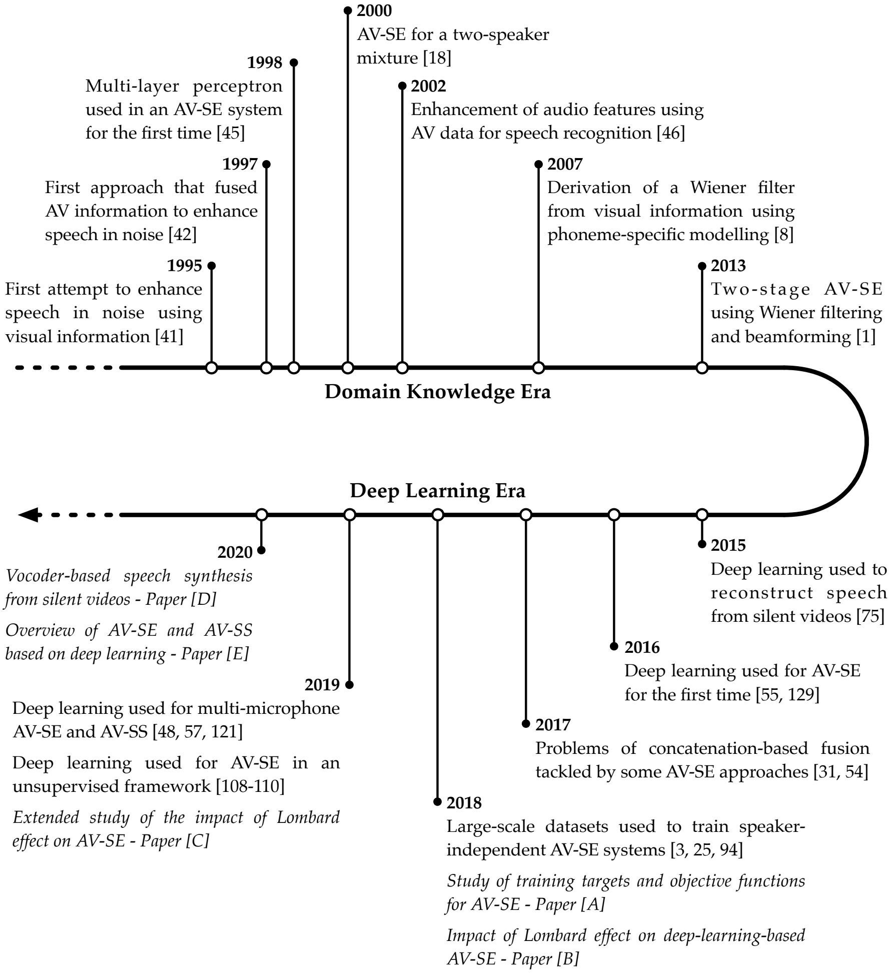

2000
AV-SE for a two-speaker mixture [18]
1998
Multi-layer perceptron used in an AV-SE system for the first time [45]
2002
Enhancement of audio features using AV data for speech recognition [46]
1997
First approach that fused AV information to enhance speech in noise [42]
2007
Derivation of a Wiener filter from visual information using phoneme-specific modelling [8]
1995
First attempt to enhance speech in noise using visual information [41]
2013
Two-stage AV-SE using Wiener filtering and beamforming [1]
Domain Knowledge Era
Deep Learning Era
2020
Vocoder-based speech synthesis from silent videos - Paper [D]
Overview of AV-SE and AV-SS based on deep learning - Paper [E]
2015
Deep learning used to reconstruct speech from silent videos [75]
2016
Deep learning used for AV-SE for the first time [55, 129]
2019
Deep learning used for multi-microphone AV-SE and AV-SS [48, 57, 121]
Deep learning used for AV-SE in an unsupervised framework [108-110]
Extended study of the impact of Lombard effect on AV-SE - Paper [C]
2017
Problems of concatenation-based fusion tackled by some AV-SE approaches [31, 54]
2018
Large-scale datasets used to train speaker-independent AV-SE systems [3, 25, 94]
Study of training targets and objective functions for AV-SE - Paper [A]
Impact of Lombard effect on deep-learning-based AV-SE - Paper [B]

**Fig. 13:** Non-exhaustive timeline of AV-SE literature. We included the papers in the Part II of this thesis in italics. Each year (in bold) indicates when the studies reported underneath were first made available to the community, either as a pre-print or as an actual publication in conference proceedings or journals.

representative of real-world scenarios.

In the rest of this Section, we review relevant AV-SE works in the literature, following the chronological development of the field as illustrated in the timeline of Figure 13. In particular, the overview is organised into two parts: first, we describe early, mainly knowledge-based, AV-SE approaches; then we present the latest advances in the field dominated by deep-learning-based techniques. Performing a systematic survey of the field is outside of the scope of this Introduction, but if the reader is interested in further details, we refer to [103, 119, 132] and paper [E].

### Domain Knowledge Era

The use of visual information to enhance the noisy speech of a target speaker can be traced back to the pioneering studies conducted by Girin et al. [40, 41]. In these works, adaptive filters, such as the Wiener filter [21, 79, 127], were applied to the noisy signal in order to obtain an estimate of the clean speech. The parameters of the filters were estimated using a linear regression method from three lip shape features (height, width, and area). Although the experiments were conducted using a very simple setup, i.e. an AV speech corpus consisting of stationary vowels that were corrupted with acoustic additive white noise, the results were promising in showing that visual information could be used to enhance noisy speech.

The method in [40, 41] did not jointly exploit AV information to estimate the filter applied to the noisy speech signal. Therefore, a new system was proposed in [42] with the aim of estimating the filter parameters from a combination of acoustic and visual features. The experiments, conducted for vocalic transitions degraded by additive white noise, highlighted the effectiveness of the approach especially at low SNRs, where AO-SE systems had particular problems.

Frameworks similar to the one in [42] were used in subsequent works to investigate AV-SE for vowel-plosive-vowel transitions [43–45]. In addition, a study of the impact of non-linear models for filter parameters estimation, like a single-hidden-layer MLP, was also performed [44, 45]. These works showed that the use of MLPs allowed to outperform linear regression in all the conditions under tests, obtaining good results for vowel sounds.

Although the studies presented above [40–45] showed that visual information could be exploited to enhance the speech of interest, the experimental setup was very basic: no continuous speech was used and the noise signal was stationary. Progress towards modelling more complex scenarios was made in [18, 28], where the speech of interest was extracted from a two-speaker mixture. There, the acoustic mixture signal, in the form of a TF representation, specifically a periodogram, and the visual signal from the target speaker, in the form of raw pixels or image motion features, in particular optical flow [9], were projected to a low-dimensional sub-space using two single-layer perceptrons, one for each modality. The perceptrons' weights were chosen in a way that maximised the mutual information between acoustic and visual signals in the projected space. The extraction of the speech of interest was driven by the following intuition: large weights of the perceptron mapping the acoustic signal to the maximally informative sub-space corresponded to the periodogram coefficients associated with the visual data from the target speaker. Therefore, a filter whose coefficients were proportional to the perceptron weights allowed to perform the enhancement. This approach based on mutual information differed from the other popular methods at the time, because it did not assume that audio and visual signals were distributed according to a parametric model, such as a Gaussian distribution.

Visual information was also used to enhance audio features for ASR. Goecke et al. [46] proposed to apply a linear filter to a vector consisting of concatenated acoustic and visual features. The filter was obtained by Mean Squared Error (MSE) estimation of clean audio features. The experiments, conducted on a speaker-independent setting using continuous speech, showed that an ASR system trained on AV enhanced features could outperform an analogous ASR system trained on noisy audio features. However, the performance of this method were substantially inferior to the performance obtained with a previously proposed AV ASR system [100]. One of the identified reasons was the use of a linear filtering technique. Therefore, a non-linear enhancement approach that exploited AV information was proposed in [20]. The idea behind this approach was to add to each frame of the noisy representation a weighted average of audio compensation terms pre-defined in a codebook. The weights assigned to the codewords were estimated from combined AV features. This method, called Audio-Visual Codebook Dependent Cepstral Normalisation (AVCDCN), allowed to outperform its AO counterpart by a large margin, showing that AV-SE could be used to effectively improve ASR systems.

Although Wiener filtering was used in some pioneering studies reported above [40, 44], it was proposed again, e.g. in [6–8, 10], to model more complex scenarios in an AV setting. A Wiener filter allows to optimally estimate the speech of interest according to a MSE criterion [21, 79]. In order to build such a filter, the Power Spectral Density (PSD) of the clean speech signal and the PSD of the noise signal are required. The most challenging task is to obtain the PSD of the clean speech signal. In [40, 44], this quantity was estimated using a linear regression method or a MLP. Almajai et al. [8], on the other hand, estimated the PSD of the clean speech either from a Gaussian Mixture Model (GMM), modelling the global correlation between acoustic and visual features, or from a set of phoneme-specific GMMs selected with a network of Hidden Markov Models (HMMs). The results showed the effectiveness of the approaches, especially the HMM-GMM one. The system was later improved in [6, 7], where an AV Voice Activity Detector (VAD) was introduced to get a more accurate estimate of the PSD of the noise signal.

A Wiener filter was adopted also by Abel et al. [1], who proposed a two-stage system that exploited vision and multi-channel acoustic information. In particular, the noisy multi-microphone signals were preprocessed with a visually derived Wiener filtering method; then, an adaptive beamformer was used to further enhance the speech signal. The approach showed good noise reduction performance at extremely low SNRs, although distortions were introduced by the beamforming technique when the SNR increased. This suggested that a better way to combine visual and multi-channel acoustic information was needed to guarantee good SE performance over a wide SNR range. Potentially, the use of, at that time, emerging deep-learning-based methods would have allowed to better process AV information, and this was what researchers started to investigate.

### Deep Learning Era

Before being used for AV-SE, deep learning was adopted to reconstruct speech from silent videos by Le Cornu and Milner [75]. Specifically, a system was designed to estimate the spectral envelope of the speech of interest using features extracted from the visual frames of the target speaker's mouth. The estimation was performed with either a GMM or a MLP having three hidden layers. In addition, fundamental frequency and aperiodicity were artificially generated without using visual information. The target speech signal was then reconstructed from spectral envelope, fundamental frequency and aperiodicity with the use of a vocoder. The results showed that the GMM systems outperformed the ones using a MLP. Some changes to the approach were later proposed [76] to improve the performance: the system was trained to classify codewords representing acoustic vectors instead of directly estimating the spectral envelope; RNNs were used to model the relationship between acoustic and visual feature sequences better. This work highlighted the potential of deep learning in exploiting visual information to generate intelligible speech.

The first deep-learning-based approaches that addressed the problem of AV-SE appeared in [55, 129]. Specifically, Wu et al. [129] proposed to use a CNN and a MLP to extract visual and acoustic features, respectively. Then, these features were concatenated and fed into a Bi-directional Long Short-Term Memory (BiLSTM) network, which estimated the log power spectrum of the clean speech signal. On the other hand, Hou et al. [55] concatenated noisy acoustic features with a compact visual feature vector obtained from 18 landmark points of the speaker's mouth. The AV feature vector was, then, used as input to a MLP in order to estimate the mel-scaled spectrogram of the clean speech.

Afterwards, researchers noticed that a concatenation of AV feature vectors did not allow to control how the fusion of multi-modal information occurred. In other words, AV-SE systems were not able to fully exploit visual information, but the enhancement was dominated by the audio or vice versa. Potential solutions to this problem were proposed in several works. For example, Hou et al. [54] forced the system to exploit visual information by using a network that needed to estimate not only the target speech signal, but also the visual frames of the speaker's mouth. Additionally, to guarantee that vision was used to learn the acoustic output, not only the visual frames, one of the input modalities was randomly selected to be zeroed out during training. This strategy, already used for other multi-modal systems [15, 92], is often called *multi-style training*. Gabbay et al. [31] proposed a different training procedure, in which the noise used to degrade the acoustic speech signal was another utterance spoken by the same target speaker. This was found to be effective, because the only way to extract the correct target speech from a mixture of speech signals belonging to the same speaker was by using vision. Today, attention-based mechanisms, which allow to attend to the parts of the input that are most important, can be considered state-of-the-art methods to fuse AV information [16, 48], because they perform a soft selection of the most useful modality in a given situation.

A clear advance in deep-learning-based AV-SE systems occurred when large-scale datasets were used for training. The first works [3, 25, 94] adopting AV speech corpora consisting of recordings from hundreds, and sometimes thousands, of speakers showed an impressive improvement over previous systems. Specifically, these approaches were able to effectively enhance speech signals from speakers not observed at training time, an achievement not reported before. Later, speaker-independent techniques were also proposed for speech reconstruction from silent videos [124].

Recently, research in deep-learning-based AV-SE moved towards other aspects. For example, novel systems were proposed to exploit information from multi-microphone recordings [48, 57, 121, 131]. In fact, vision can be used not only to enhance the speech of interest from the lip movements, but also to localise the target speaker. This is useful in order to accurately estimate the speech Direction Of Arrival (DOA) for beamforming. Another interesting research direction is the one investigated by Sadeghi et al. [108–110]: the use of deep learning to perform AV-SE in an unsupervised learning framework. Deep-learning-based approaches are generally trained to learn a mapping between pairs of synthetically generated noisy speech signals and clean speech signals in a supervised learning setting. This approach usually requires a large amount of data in order to cover different situations. On the other hand, Sadeghi et al. [110] proposed to use Variational Auto-Encoders (VAEs) to model the characteristics of the speech of interest, without the need of using multiple noise types and noise levels to guarantee good generalisation performance. This is an important step towards systems that are not limited by requirements of access to large-scale datasets.

# 4 Contributions

The main body of this thesis (Part II) consists of a collection of five papers which document the scientific contributions of our work. These contributions address specific research questions and observations that we list in Table 1.

**Table 1:** Research questions and observations addressed by the papers in Part II and respective contributions.

| Paper | Research Question (RQ) / Observation (OB) | Contribution |
|---|---|---|
| [A] | RQ: What is the effect of the choice of training targets and objective functions for AV-SE? | Experimental study of several training targets and objective functions for an AV-SE system. |
| | OB: Lack of uniformity regarding the terminology used for training targets and objective functions in the context of SE. | Proposal of a new taxonomy to unify the different terminologies in the literature. |
| [B] | RQ: What is the impact of Lombard effect on state-of-the-art AV-SE systems? | Investigation of the impact of a mismatch between training data (neutral speech) and testing data (Lombard speech) on AV-SE performance. |
| | RQ: What is the benefit of using Lombard speech to train AV-SE systems? | Experimental study on the improvement that can be achieved using Lombard speech for training. |
| [C] | RQ: What is the effect of the inter-speaker variability in Lombard condition on AV-SE systems? | Analysis of per-speaker performance in Lombard condition in relation to acoustic and visual features. |
| | OB: Lack of a Lombard-aware AV-SE system designed for a wide SNR range. | Design of an AV-SE system trained with neutral and Lombard speech to operate on a wide SNR range. |
| | OB: Lack of AV methods to assess the performance of SE systems. | Use of AV stimuli in listening tests to evaluate systems' performance. |
| [D] | RQ: Can vocoder-based systems for speech reconstruction from silent videos estimate fundamental frequency and aperiodicity? | Design of a system that could predict vocoder parameters and text information from silent videos in a multi-task learning fashion. |
| [E] | OB: Lack of a comprehensive overview on deep-learning-based AV-SE and AV-SS. | Extensive survey covering several aspects of AV-SE and AV-SS: AV speech datasets; evaluation methods; main elements of AV-SE and AV-SS systems; related research. |

[A] "On Training Targets and Objective Functions for Deep-Learning-Based Audio-Visual Speech Enhancement", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[B] "Effects of Lombard Reflex on the Performance of Deep-Learning-Based Audio-Visual Speech Enhancement Systems", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[C] "Deep-Learning-Based Audio-Visual Speech Enhancement in Presence of Lombard Effect", *Speech Communication*, 2019.

[D] "Vocoder-Based Speech Synthesis from Silent Videos", *Interspeech (to appear)*, 2020.

[E] "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation", *Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

The entire investigation spans multiple aspects of deep-learning-based AV-SE, following a logical progression. Specifically, we start by conducting a study of training targets and objective functions (paper [A]), given the unclear impact of them on the performance of AV-SE systems. Using the optimal combination of training target and objective function found in this study, we then extensively investigate the impact of Lombard effect on deep-learning-based AV-SE (papers [B] and [C]). Afterwards, we decide to put our effort on the special case of speech reconstruction from silent videos (paper [D]), a task that can be reasonably considered harder than the general AV-SE problem, because only visual information is available as input. Finally, we propose a systematic survey of the current state of the art in the field of AV-SE and AV-SS (paper [E]), with two aims: helping beginners to navigate through the large number of approaches in the literature; inspiring experts by providing insights and perspectives on current challenges and possible future research directions.

In the following, we shortly summarise the content of each paper.

### [A] On Training Targets and Objective Functions for Deep-Learning-Based Audio-Visual Speech Enhancement

In this paper, we perform an experimental study of training targets and objective functions for deep-learning-based audio-visual speech enhancement. Specifically, we compare systems that estimate either a mask or a spectrogram, optimising the mean squared error between the ground truth and the output in several domains. The analysis is conducted for both seen and unseen speakers using six kinds of additive noise (one of them not used at training time, to evaluate the generalisation performance of the approaches). In previous works, similar experiments have been performed for audio-only speech enhancement systems, but the findings may be inappropriate for audio-visual techniques, since the target estimation is mostly driven by visual cues when the signal to noise ratio is low.

The results show that the best performance in terms of estimated speech quality and intelligibility can be generally obtained by a direct estimation of a mask. The model that directly estimates the log magnitude spectrum of the speech of interest performs similarly in terms of estimated speech quality.

In addition, a new taxonomy is proposed to unify the different terminologies adopted for training targets and objective functions in the context of speech enhancement.

### [B] Effects of Lombard Reflex on the Performance of Deep-Learning-Based Audio-Visual Speech Enhancement Systems

In this paper, we investigate the impact of Lombard effect on audio-visual speech enhancement systems based on deep learning. In particular, we examine two aspects. First, we verify whether a system trained on neutral speech, which is the general practice for state-of-the-art approaches, can improve Lombard speech at several signal to noise ratios. Second, we conduct some experiments with systems trained on Lombard speech to quantify the performance improvement that can be achieved over the systems trained with neutral speech. All these aspects are important, because audio-visual techniques are expected to be deployed in situations where audio-only approaches struggle, i.e. in presence of high levels of background noise and Lombard speech.

Although systems trained with neutral speech provide an improvement over the unprocessed noisy signals in terms of estimated speech quality and intelligibility, we find that Lombard-aware systems can improve the signals even more. This confirms that the mismatch between neutral and Lombard speech should not be ignored by audio-visual speech enhancement systems.

### [C] Deep-Learning-Based Audio-Visual Speech Enhancement in Presence of Lombard Effect

This paper extends the investigation of paper [B]. Specifically, after having conducted new experiments using a cross-validation setting, we analyse the impact that the inter-speaker variability has on audio-visual speech enhancement with respect to acoustic (fundamental frequency) and geometric articulatory (mouth aperture and mouth spreading) features. Then, we propose a Lombard-aware system that can be used to enhance speech signals at low and high signal to noise ratios, by training it with Lombard and neutral speech. Finally, we conduct listening tests with audio-visual stimuli, since current objective measures used for performance assessment have the limitation of estimating speech quality and intelligibility from audio signals in isolation.

The results, using both objective measures and listening tests, confirm the findings in paper [B], highlighting the benefit of training a system with Lombard speech. In addition, we find that the performance gap between Lombard-aware and non-Lombard-aware systems is larger for female speakers. This gender difference is likely to be caused by the speech characteristics that female speakers exhibit in Lombard condition, with a large increment in fundamental frequency, which is the feature that correlates the most with the estimated speech quality and intelligibility increase among the ones investigated. Furthermore, the way we train signal-to-noise-ratio-independent systems by using Lombard speech for low signal to noise ratios and neutral speech for high signal to noise ratios is effective: a substantial improvement can be observed in Lombard conditions if compared with a non-Lombard-aware approach, while a negligible performance drop occurs at higher signal to noise ratios.

### [D] Vocoder-Based Speech Synthesis from Silent Videos

In this paper, we propose an approach to reconstruct speech from the silent video of a talker. The system is based on deep learning and is able to simultaneously perform speech reconstruction and speech recognition from raw video frames in a multi-task learning framework. We conduct experiments on both seen and unseen speakers settings.

We show that our vocoder-based system is able to reach state-of-the-art performance in terms of estimated speech quality and intelligibility. A per-speaker analysis of the performance highlights that, for unseen speakers, the scores obtained from the used evaluation measures spread over a wide range. This suggests that the network finds it hard to perform the task for speakers whose facial traits largely differ from the speakers in the training set. In addition, we report a performance trade-off between speech reconstruction and speech recognition, inherited from the used multi-task learning framework, that should be further investigated in future works.

### [E] An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation

In this paper, we present a thorough survey of audio-visual speech enhancement and separation based on deep learning. Previous overview articles focused on either audio-only approaches or knowledge-based audio-visual methods. Recently, many researchers started to use deep learning for audio-visual speech enhancement and separation, motivated by the performance that data-driven approaches allow to achieve. Therefore, there is a need for a systematic overview and discussion of recent advances in the field.

We structured the overview by introducing a signal model and a formulation of the problems. Then, we present the main audio-visual speech datasets and evaluation methods, because they are relevant for training the systems and compare them, respectively. After that, the focus shifts to the principal elements of audio-visual systems, namely visual features, acoustic features, deep learning methods, fusion techniques, training targets and objective functions. Finally, we survey the approaches for speech reconstruction from silent videos and audio-visual sound source separation for non-speech signals, given their strong link with audio-visual speech enhancement and separation. Throughout the paper, we avoid to advocate one method over another based on their performance. Instead, we try to let novel ideas emerge to inspire and stimulate new research.

# 5 Future Research Directions

This thesis explores several aspects of audio-visual speech enhancement based on deep learning. This research topic has recently received attention from the speech community, due to the many different elements that the combination of acoustic and visual information allows to explore. In the following, we report some possible research directions that would allow a substantial progress in the field.

## Target Speaker Detection

In an acoustic scene with several speakers and background noise, detecting the speaker to be treated as a target is challenging, especially when only the acoustic signals are available. Visual information makes it easier to detect the target speaker, since vision is essentially not affected by the acoustic environment. However, there are situations in which detecting the target speaker in a video is critical, for example: when the video resolution is low and computer vision algorithms cannot detect the speaker's face; when the speaker is not in the video (a typical scenario is that of dubbed movies); when there is a mismatch between the position of the speaker and the speech direction of arrival, which might occur when the speaker is using a microphone during public events and the loudspeakers are placed in a different location. In all these cases, the acoustic cues should be weighted more than the visual ones. Designing a system that intelligently handles different situations by correctly detecting the target speaker and avoiding to suppress the speech of interest is an interesting future research direction.

## Objective Functions

Deep-learning-based AV-SE systems are usually trained to minimise the MSE or the Mean Absolute Error (MAE) between their output and the ground truth. Although these objective functions allow to achieve good performance, perceptually-motivated alternatives might improve the intelligibility of the enhanced signals. Another possibility is to use an adversarial training procedure, where the network that performs the enhancement tries to fool another neural network, called discriminator, whose role is to distinguish between the enhanced speech signals and the ground truth. This framework showed its effectiveness for AO-SE [86, 95] and could be adopted also for AV-SE.

## Real-Time Models for Low Resource Devices

Processing information through deep learning models is challenging in some applications. Consider, for example, hearing aids: they are devices that need to guarantee low latency performance despite having a quite low processing power and a limited storage. Designing a system based on deep learning with these constraints is not trivial, because deep neural networks generally consist of millions of parameters, which cannot be easily handled by the embedded system of a hearing aid. This challenge exists for AO-SE systems, but it is even bigger for AV-SE, where the data from the visual feed has a higher dimensionality if compared to the acoustic signal. Reducing the number of parameters of a deep learning model without having a performance drop is highly desirable and represents a research topic that should be investigated in the future.

## Unsupervised Learning

Most deep-learning-based AV-SE systems consist of supervised learning approaches. This means that the deep learning models are trained with data that is synthetically generated to resemble real-world conditions. Usually, the clean speech is added to different kinds of noise at several SNRs and the neural network learns a mapping between the noisy and the target speech signals. This approach has several potential problems. First of all, the clean speech needs to be mixed with a large number of noise types during training, so that the model can reach good generalisation performance. Then, noise and speech are treated as independent signals, an assumption that is not correct (for example, Lombard effect occurs when the speaker is immersed in a noisy environment). Finally, collecting clean speech data requires effort and resources, because speech is usually degraded by a certain amount of noise. Although some attempts at using an unsupervised learning framework exist [110], finding a way to leverage both clean speech and actual recordings of speech in noise could make a breakthrough in the field.

## Audio-Visual Estimators of Speech Quality and Intelligibility

We have seen that speech perception is not unimodal: both acoustic and visual information contribute to the way in which humans perceive speech. Ideally, the performance of a SE system in terms of quality and intelligibility should be measured with listening tests. Unfortunately, this procedure is often time consuming and costly. Therefore, objective measures of speech quality and intelligibility are often preferred. However, current objective measures do not take visual information into account. The development of AV estimators of speech quality and intelligibility is an interesting future research topic.

# References

[1] A. Abel and A. Hussain, “Novel two-stage audiovisual speech filtering in noisy environments,” *Cognitive Computation*, vol. 6, no. 2, pp. 200–217, 2014.

[2] H. Ackermann, S. R. Hage, and W. Ziegler, “Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective,” *Behavioral and Brain Sciences*, no. 6, pp. 529–546, 2014.

[3] T. Afouras, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” *Proc. of Interspeech*, 2018.

[4] Y. Agiomyrgiannakis and O. Rosec, “ARX-LF-based source-filter methods for voice modification and transformation,” in *Proc. of ICASSP*, 2009.

[5] R. Aibara, J. T. Welsh, S. Puria, and R. L. Goode, “Human middle-ear sound transfer function and cochlear input impedance,” *Hearing Research*, vol. 152, no. 1-2, pp. 100–109, 2001.

[6] I. Almajai and B. Milner, “Effective visually-derived Wiener filtering for audio-visual speech processing,” in *Proc. of AVSP*, 2009.

[7] ——, “Visually derived Wiener filters for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[8] I. Almajai, B. Milner, J. Darch, and S. Vaseghi, “Visually-derived Wiener filters for speech enhancement,” in *Proc. of ICASSP*, 2007.

[9] P. Anandan, “A computational framework and an algorithm for the measurement of visual motion,” *International Journal of Computer Vision*, vol. 2, no. 3, pp. 283–310, 1989.

[10] F. Berthommier, “Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement,” in *Proc. of ICASSP*, 2004.

[11] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT press, 1997.

[12] A. W. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[13] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[14] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proc. of EMNLP*, 2014.

[15] J. S. Chung and A. Zisserman, “Lip reading in profile,” in *Proc. of BMVC*, 2017.

[16] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, “Facefilter: Audio-visual speech separation using still images,” *arXiv preprint arXiv:2005.07074*, 2020.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] T. Darrell, J. W. Fisher, and P. Viola, "Audio-visual segmentation and "the cocktail party effect"," in *Proc. of ICMI*, 2000.

[19] H. Davson, *Physiology of the Eye*. Macmillan International Higher Education, 1990.

[20] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization)," in *Proc. of SAM*, 2002.

[21] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 2000.

[22] P. Denes and E. Pinson, *The Speech Chain*. Macmillan, 1993.

[23] J. J. Dreher and J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *The Journal of the Acoustical Society of America*, vol. 29, no. 12, pp. 1320–1323, 1957.

[24] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[25] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[26] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[27] G. Fant, *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations*, ser. Description and analysis of contemporary standard Russian ; 2. The Hague, Netherlands: Mouton, 1960.

[28] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. of NIPS*, 2001.

[29] M. Fitzpatrick, J. Kim, and C. Davis, "Auditory and auditory-visual Lombard speech perception by younger and older adults," in *Proc. of AVSP*, 2013.

[30] ——, "The effect of seeing the interlocutor on auditory and visual speech production in noise," *Speech Communication*, vol. 74, pp. 37–51, 2015.

[31] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," *Proc. of Interspeech*, 2018.

[32] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Lœvenbruck, "An acoustic and articulatory study of lombard speech: Global effects on the utterance," in *Proc. of ICSLP*, 2006.

[33] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the Lombard effect," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 3, pp. 588–608, 2010.

[34] M. Garnier, L. Ménard, and G. Richard, "Effect of being seen on the production of visible speech cues. a pilot study on Lombard speech," in *Proc. of Interspeech*, 2012.

[35] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*. CRC Press, 2016.

[36] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement?unimportant, important, or impossible?" in *Proc. of CEEEI*, 2012.

[37] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[38] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[39] B. Gick, I. Wilson, and D. Derrick, *Articulatory phonetics*. John Wiley & Sons, 2012.

[40] L. Girin, G. Feng, and J.-L. Schwartz, "Débruitage de parole par un filtrage utilisant l'image du locuteur. une étude de faisabilité," *Traitement du Signal*, vol. 13, no. 4, pp. 319–334, 1996.

[41] ——, "Noisy speech enhancement with filters estimated from the speaker's lips," in *Proc. of EUROSPEECH*, 1995.

[42] ——, "Noisy speech enhancement by fusion of auditory and visual information: A study of vowel transitions," in *Proc. of EUROSPEECH*, 1997.

[43] L. Girin, J.-L. Schwartz, and G. Feng, "Can the visual input make the audio signal "pop out" in noise? A first study of the enhancement of noisy vcv acoustic sequences by audio-visual fusion," in *Proc. of AVSP*, 1997.

[44] ——, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[45] L. Girin, L. Varin, G. Feng, and J.-L. Schwartz, "A signal processing system for having the sound "pop-out" in noise thanks to the image of the speaker's lips: New advances using multi-layer perceptrons," in *Proc. of ICSLP*, 1998.

[46] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. of ICASSP*, vol. 2, 2002.

[47] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016.

[48] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[49] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," *Signal Processing*, vol. 17, no. 3, pp. 282–282, 1989.

[50] ——, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 1-2, pp. 151–173, 1996.

[51] P. Heracleous, C. T. Ishi, M. Sato, H. Ishiguro, and N. Hagita, "Analysis of the visual Lombard effect and automatic recognition experiments," *Computer Speech & Language*, vol. 27, no. 1, pp. 288–300, 2013.

[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[53] K. Hornik, M. Stinchcombe, H. White *et al.*, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[54] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[55] J.-C. Hou, S.-S. Wang, Y.-H. Lai, J.-C. Lin, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using deep neural networks," in *Proc. of APSIPA*, 2016.

[56] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2007.

[57] E. Ideli, *Audio-visual speech processing using deep learning techniques*. MSc thesis, Applied Sciences: School of Engineering Science, Simon Fraser University, 2019.

[58] ITU-R, "BS.562: Subjective assessment of sound quality," 1990.

[59] ——, "BS.1284-2: General methods for the subjective assessment of sound quality," 2019.

[60] ITU-T, "Recommendation P.830: Subjective performance assessment of telephone-band and wideband digital codecs," 1996.

[61] ——, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[62] ——, "Recommendation P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO," 2003.

[63] ——, "Recommendation P.862.2: Wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs," 2005.

[64] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[65] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[66] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[67] E. Keller, "The analysis of voice quality in speech processing," in *International School on Neural Networks, Initiated by IIASS and EMFCSC*. Springer, 2004.

[68] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952.

[69] J. Kim, A. Sironic, and C. Davis, "Hearing speech in noise: Seeing a loud talker is better," *Perception*, vol. 40, no. 7, pp. 853–862, 2011.

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, 2015.

[71] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *the Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[72] M. Kolbæk, *Single-Microphone Speech Enhancement and Separation Using Deep Learning*. PhD thesis, Aalborg University Press, 2018.

[73] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. of NIPS*, 2012.

[74] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[75] T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. of Interspeech*, 2015.

[76] ——, "Generating intelligible audio speech from visual speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, 2017.

[77] Y. LeCun, "Generalization and network design strategies," *Connectionism in Perspective*, vol. 19, pp. 143–155, 1989.

[78] F.-F. Li *et al.*, "Notes from the Stanford CS class CS231n: Convolutional neural networks for visual recognition," http://cs231n.github.io, Accessed: 01-08-2020.

[79] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2013.

[80] E. Lombard, "Le signe de l'elevation de la voix," *Annales des Maladies de L'Oreille et du Larynx*, vol. 37, no. 2, pp. 101–119, 1911.

[81] P. Ma, S. Petridis, and M. Pantic, "Investigating the Lombard effect influence on end-to-end audio-visual speech recognition," *Proc. of Interspeech*, 2019.

[82] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, "The impact of the Lombard effect on audio and visual speech recognition systems," *Speech Communication*, vol. 100, pp. 58–68, 2018.

[83] D. W. Massaro and J. A. Simpson, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Psychology Press, 2014.

[84] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[85] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[86] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *Proc. of Interspeech*, 2017.

[87] T. B. Moeslund, *Introduction to Video and Image Processing: Building Real Systems and Applications*. Springer Science & Business Media, 2012.

[88] B. C. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.

[89] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[90] M. Morise and Y. Watanabe, "Sound quality comparison among high-quality vocoders by using re-synthesized speech," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 263–265, 2018.

[91] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[92] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. of ICML*, 2011.

[93] C. Olah, "Understanding LSTM networks," http://colah.github.io/posts/2015-08-Understanding-LSTMs, 2015, Accessed: 01-08-2020.

[94] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. of ECCV*, 2018.

[95] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *Proc. of Interspeech*, 2017.

[96] R. Patel and K. W. Schell, "The influence of linguistic content on the Lombard effect," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 1, pp. 209–220, 2008.

[97] J. Pickles, *An introduction to the physiology of hearing*. Brill, 2013.

[98] D. Pisoni, R. Bernacki, H. Nusbaum, and M. Yuchtman, "Some acoustic-phonetic correlates of speech produced in noise," in *Proc. of ICASSP*, 1985.

[99] A. L. Pittman and T. L. Wiley, "Recognition of speech produced in noise," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 3, pp. 487–496, 2001.

[100] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. of ICASSP*, 2001.

[101] S. Puria, W. T. Peake, and J. J. Rosowski, "Sound-pressure measurements in the cochlear vestibule of human-cadaver ears," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2754–2770, 1997.

[102] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, and S. M. Williams, *Neuroscience*. Sunderland, MA, Sinauer Associates, 2001.

[103] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[104] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP*, 2001.

[105] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[106] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, p. 386, 1958.

[107] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[108] M. Sadeghi and X. Alameda-Pineda, "Mixture of inference networks for VAE-based audio-visual speech enhancement," *arXiv preprint arXiv:1912.10647*, 2019.

[109] ——, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *Proc. of ICASSP*, 2020.

[110] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.

[111] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proceedings of the IEEE*, vol. 54, no. 5, pp. 720–734, 1966.

[112] B. G. Shinn-Cunningham and V. Best, "Selective attention in normal and impaired hearing," *Trends in Amplification*, vol. 12, no. 4, pp. 283–299, 2008.

[113] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[114] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *Proc. of ICASSP*, 1999.

[115] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[116] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[117] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[118] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[119] T. M. F. Taha and A. Hussain, "A survey on techniques for enhancing speech," *International Journal of Computer Applications*, vol. 179, no. 17, pp. 1–14, 2018.

[120] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. of CVPR*, 2014.

[121] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542–553, 2020.

[122] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 1902–1915, 2008.

[123] E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan, and H. C. Yehia, "Audiovisual lombard speech: reconciling production and perception," in *Proc. of AVSP*, 2007.

[124] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," in *Proc. of Interspeech*, 2019.

[125] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[126] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[127] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. John Wiley & Sons, 1949.

[128] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. of ASRU*, 2019.

[129] Z. Wu, S. Sivadas, Y. K. Tan, M. Bin, and R. S. M. Goh, "Multi-modal hybrid deep neural network for speech enhancement," *arXiv preprint arXiv:1606.04750*, 2016.

[130] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin," in *Proc. of ICASSP*, 2012.

[131] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural spatio-temporal beamformer for target speech separation," *Proc. of Interspeech (to appear)*, 2020.

[132] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," *arXiv preprint arXiv:2001.04758*, 2020.

[133] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer Science & Business Media, 2013, vol. 22.

This page intentionally left blank.

# Part II

# Papers

This page intentionally left blank.

# Paper A

## On Training Targets and Objective Functions for Deep-Learning-Based Audio-Visual Speech Enhancement

Daniel Michelsanti, Zheng-Hua Tan, Sigurdur Sigurdsson, Jesper Jensen

The paper has been published in
*Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 8077–8081, 2019.

# Abstract

*Audio-visual speech enhancement (AV-SE) is the task of improving speech quality and intelligibility in a noisy environment using audio and visual information from a talker. Recently, deep learning techniques have been adopted to solve the AV-SE task in a supervised manner. In this context, the choice of the target, i.e. the quantity to be estimated, and the objective function, which quantifies the quality of this estimate, to be used for training is critical for the performance. This work is the first that presents an experimental study of a range of different targets and objective functions used to train a deep-learning-based AV-SE system. The results show that the approaches that directly estimate a mask perform the best overall in terms of estimated speech quality and intelligibility, although the model that directly estimates the log magnitude spectrum performs as good in terms of estimated speech quality.*

# 1 Introduction

Human-human and human-machine interaction that involves speech as a communication form can be affected by acoustical background noise, which may have a strong impact on speech quality and speech intelligibility. The improvement of one or both of these two speech aspects is known as speech enhancement (SE). Traditionally, this problem has been tackled by adopting audio-only SE (AO-SE) techniques [1, 2]. However, speech communication is generally not a unimodal process: visual cues play an important role in speech perception, since they can improve or even alter how phonemes are perceived [3]. This suggests that integrating auditory and visual information can lead to a general improvement in the performance of SE systems. This intuition has lead to the proposal of several audio-visual SE (AV-SE) techniques, e.g. [4], including deep-learning-based approaches [5–7].

When supervised learning-based methods are used either for AV-SE or for AO-SE, the choice of the target and the objective function used to train the model has a crucial impact on the performance of the system. In this paper, *training target* denotes the desired output of a supervised learning algorithm, e.g. a neural network (NN), while *objective function*, or *cost function*, is the function that quantifies how close the algorithm output is to the target. The effect that targets and objective functions have on AO-SE has been investigated in several works [8–10]. The estimation of a *mask*, which is used to reconstruct the target speech signal by an element-wise multiplication with a time-frequency (TF) representation of the noisy signal, is usually preferred to a direct estimation of a TF representation of the clean speech signal [11]. The reason is that a mask is easier to estimate [11], because it is generally smoother than a spectrogram, its values have a narrow dynamic range [8],

**Table A.1:** Objective functions of the approaches used in this study organised according to our taxonomy. Here, $a = \frac{1}{TF}$ and $b = \frac{1}{TQ}$.

| | Direct Mapping (DM) | Indirect Mapping (IM) | Mask Approximation (MA) |
|---|---|---|---|
| STSA | $J = a\sum_{k,l}\left(A_{k,l} - \widehat{A}_{k,l}\right)^2$ (1) | $J = a\sum_{k,l}\left(A_{k,l} - \widehat{M}_{k,l}R_{k,l}\right)^2$ (6) | $J = a\sum_{k,l}\left(M^{\mathrm{IAM}}_{k,l} - \widehat{M}_{k,l}\right)^2$ (11) |
| LSA | $J = a\sum_{k,l}\left(\log(A_{k,l}) - \log(\widehat{A}_{k,l})\right)^2$ (2) | $J = a\sum_{k,l}\left(\log(A_{k,l}) - \log(\widehat{M}_{k,l}R_{k,l})\right)^2$ (7) | - |
| MSA | $J = b\sum_{q,l}\left(\overline{A}_{q,l} - \widehat{\overline{A}}_{q,l}\right)^2$ (3) | $J = b\sum_{q,l}\left(\overline{A}_{q,l} - \widehat{\overline{M}}_{q,l}\overline{R}_{q,l}\right)^2$ (8) | - |
| LMSA | $J = b\sum_{q,l}\left(\log(\overline{A}_{q,l}) - \log(\widehat{\overline{A}}_{q,l})\right)^2$ (4) | $J = b\sum_{q,l}\left(\log(\overline{A}_{q,l}) - \log(\widehat{\overline{M}}_{q,l}\overline{R}_{q,l})\right)^2$ (9) | - |
| PSSA | $J = a\sum_{k,l}\left(A_{k,l}\cos(\theta_{k,l}) - \widehat{A}_{k,l}\right)^2$ (5) | $J = a\sum_{k,l}\left(A_{k,l}\cos(\theta_{k,l}) - \widehat{M}_{k,l}R_{k,l}\right)^2$ (10) | $J = a\sum_{k,l}\left(M^{\mathrm{PSM}}_{k,l} - \widehat{M}_{k,l}\right)^2$ (12) |

and also because a filtering approach is considered less challenging than the synthesis of a clean spectrogram [7]. Since no studies on this matter have been performed in the AV domain, design choices of AV frameworks [6, 7] and their performance [6] are often motivated by the findings in the AO related works. However, these findings may be inappropriate in the AV domain because, especially at very low signal to noise ratios (SNRs), the estimation of the target is mostly driven by the visual component of the speech. Hence, there is a need for a comprehensive study of the role of training targets and cost functions in AV-SE.

The contribution of this paper is two-fold. First, we propose a new taxonomy that unifies the different terminologies used in the literature, from classical statistical model-based schemes to more recent deep-learning-based ones. Furthermore, we present a comparison of several targets and objective functions to understand if a particular training target that performs universally good (across various acoustic situations) exists, and if training targets that are good in the AO domain remain good in the AV domain.

## 2 Training Targets and Objective Functions

Recent works on AO-SE [8, 10, 12] make use of different terminologies for the same approaches. Sometimes, this lack of uniformity can be confusing. In this section, we review cost functions and training targets from the AO domain and introduce a new taxonomy for SE, unifying the terminology used for the classical SE optimisation criteria [13, 14] and for the objective functions adopted in the recent deep-learning-based techniques [8, 10] (cf. Table A.1).

The problem of SE is often formulated as the task of estimating the clean speech signal $x(n)$ given the mixture $y(n) = x(n) + d(n)$, where $d(n)$ is an additive noise signal, and $n$ denotes a discrete-time index. We can formulate the signal model also in the TF domain, as: $Y(k,l) = X(k,l) + D(k,l)$, where $k$

indicates the frequency bin index, $l$ denotes the time frame index, and $Y(k,l)$, $X(k,l)$, and $D(k,l)$ are the short-time Fourier transform (STFT) coefficients of the mixture, the clean signal, and the noise, respectively. Since the STFTs' phases do not have a clear structure, their estimation is hard to perform with a NN [15]. Hence, generally, only the magnitude of the clean STFT is estimated, and the clean signal is reconstructed using the phase of $Y(k,l)$ [8, 10].

## 2.1 Direct Mapping

Let $A_{k,l} = |X(k,l)|$ and $R_{k,l} = |Y(k,l)|$ denote the magnitude of the clean and the noisy STFT coefficients, respectively. A straightforward way to estimate the short-time spectral amplitude (STSA) of the clean signal is a direct mapping (DM) approach [12], in which a NN is trained to output an estimate $\widehat{A}_{k,l}$ that minimises a cost function, e.g. Eq. (1) [13, 16], with $k = 1, \dots, F$ and $l = 1, \dots, T$, where $F$ is the number of frequency bins of the spectrum estimated by the NN, and $T$ is the number of time frames.

Since a logarithmic law reflects better the human loudness perception [17], a cost function that operates in the log spectral amplitude (LSA) domain may be formulated as in Eq. (2) [14, 18].

To incorporate the fact that the human auditory system is more discriminative at low than at high frequencies [19], a Mel-scaled spectrum may be defined as $\overline{A}_l = BA_l$, where $A_l$ denotes an $F$-dimensional vector of STFT coefficient magnitudes for time frame $l$, and $B \in \mathbb{R}^{Q \times F}$ is a matrix, implementing a Mel-spaced filter bank, with $Q$ being the number of the Mel-frequency bins. We denote the $q$-th coefficient of the Mel-scaled spectrum at frame $l$ of the clean signal as $\overline{A}_{q,l}$, and its estimate as $\widehat{\overline{A}}_{q,l}$. Then, a cost function in the Mel-scaled spectral amplitude (MSA) domain can be defined as in Eq. (3) [20].

We can combine the considerations leading to Eqs. (2) and (3) to find an estimate that minimises a cost function in the log Mel-scaled spectral amplitude (LMSA) domain, as in Eq. (4) [5, 21].

Considering only the STSA of the clean signal for the estimation can lead to an inaccurate complex STFT estimation, since the phase of $X(k,l)$ is, generally, different from the phase of $Y(k,l)$ [11]. For this reason, in [10], a factor to compensate for the phase mismatch[1] is proposed. The cost function that makes use of a phase sensitive spectral amplitude (PSSA) is defined in Eq. (5), where $\theta_{k,l}$ denotes the phase difference between the noisy and the clean signals.

[1] In [10] a phase compensation factor is used to learn a mask, cf. Eq. (10).

## 2.2 Indirect Mapping

An alternative approach is to have a different training target, and perform an indirect mapping (IM) [9, 10, 12], where a NN is trained to estimate a mask, which is easier to estimate [11], using an objective function which is defined based on reconstructed spectral amplitudes. The cost functions analogous to Eqs. (1)–(5) are defined in Eqs. (6)–(10), where $\widehat{M}_{k,l}$ is the estimate of the magnitude mask, $\widehat{\overline{M}}_{q,l}$ is the estimate of the Mel-scaled mask, and $\overline{R}_{q,l}$ is the Mel-spectrum in frequency subband $q$ and frame $l$ of the noisy signal.

## 2.3 Mask Approximation

Since in the IM approach a NN learns a mask, one can also define an objective function directly in the mask domain and perform a mask approximation (MA). In the literature, many different masks have been defined, but in this work we only consider the ideal amplitude mask (IAM), $M_{k,l}^{\text{IAM}} = \frac{A_{k,l}}{R_{k,l}}$, and the phase sensitive mask (PSM), $M_{k,l}^{\text{PSM}} = \frac{A_{k,l}}{R_{k,l}} \cos(\theta_{k,l})$, because they appear to be the best-performing and allow us to directly compare with the respective IM versions, cf. Eqs. (6) and (10). The cost functions are defined in Eqs. (11) and (12) [8, 11], respectively.

While Eqs. (11) and (12) have led to good performance in the AO-SE domain [8, 15], the cost functions have been proposed on a heuristic basis. To get insights into their operation, we can rewrite Eq. (11) as $J = \frac{1}{TF} \sum_{k,l} \frac{\left(A_{k,l} - \widehat{M}_{k,l} R_{k,l}\right)^2}{{R_{k,l}}^2}$, which differs from Eq. (6) only due to the $\frac{1}{{R_{k,l}}^2}$ factor. Hence, Eq. (11) is nothing more than a spectrally weighted version of Eq. (6) [22], which reduces the cost of estimation errors at high-energy spectral regions of the noisy signal relative to low-energy spectral regions, and is related to a perceptually motivated cost function proposed in [23]. Similar considerations can be done for Eqs. (10) and (12), leading to the conclusion that Eq. (12) is a spectrally weighted version of Eq. (10). For simplicity, we refer to the approaches that estimate the IAM and the PSM as STSA-MA and PSSA-MA, respectively.

# 3 Experiments

## 3.1 Audio-Visual Corpus and Noise Data

We conducted experiments on the GRID corpus [24], consisting of audio and video recordings of 1000 six-word utterances spoken by each of 34 talkers (s1−34). Each video consists of 75 frames recorded at 25 frames per second with a resolution of 720×576 pixels. The audio tracks have a sample frequency of 44.1 kHz. To train our models, we divided the data as follows: 600 utterances of 25 speakers for training; 600 utterances of 2 speakers (s14 and s15) not in the training set for validation; 25 utterances of each of the speakers in the training set for testing the models in a seen speaker setting; 100 utterances of 6 speakers (s1$-$4, s7, and s11, 3 males and 3 females) not in the training set for testing the models in an unseen speaker setting. The utterances have been randomly chosen among the ones for which the mouth was successfully detected with the approach described in Sec. 3.2.

Six kinds of additive noise have been used in the experiments: bus (BUS), cafeteria (CAF), street (STR) pedestrian (PED), babble (BBL), and speech shaped noise (SSN) as in [25]. For the training and the validation sets, we mixed the first five noise types with the clean speech signals at 9 different SNRs, in uniform steps between $-20$ dB and 20 dB. We included SSN in the test set, for the evaluation of the generalisation performance to unseen noise, and evaluated the models between $-15$ dB and 15 dB SNRs (the performance at $-20$ dB and 20 dB can be found in [26], omitted here due to space limitations). The noise signals used to generate the mixtures in the training, the validation, and the test sets are disjoint over the 3 sets.

## 3.2 Audio and Video Preprocessing

Each audio signal was downsampled to 16 kHz and peak-normalised to 1. A TF representation was obtained by applying a 640-point STFT to the waveform signal, using a 640-sample Hamming window and a hop size of 160 samples. The magnitude spectrum was then split into 20-frame-long parts, corresponding to 200 ms, the duration of 5 video frames. Due to spectral symmetry, only the 321 frequency bins that cover the positive frequencies were taken into account.

For each video signal, we first determined a bounding box containing the mouth with the Viola-Jones detection algorithm [27], and, inside that, we extracted feature points as in [28] and tracked them across all the video frames using the Kanade-Lucas-Tomasi (KLT) algorithm [29, 30]. Then, we cropped a mouth-centred region of size 128×128 pixels based on the tracked feature points, and we concatenated 5 consecutive grayscale frames, corresponding to 200 ms.

## 3.3 Architecture and Training Procedure

Inspired by [5], we used a NN architecture that operates in the STFT domain. The NN consists of a video encoder, an audio encoder, a feature fusion subnetwork, and an audio decoder.

The video encoder takes as input 5 frames of size 128×128 pixels obtained as described before, and processes them with 6 convolutional layers, each of them followed by: leaky-ReLU activation, batch normalisation, 2×2 strided max-pooling with kernel of size 2×2, and dropout with a probability of 25%. Also for the audio encoder, 6 convolutional layers are adopted, followed by leaky-ReLU activation and batch normalisation. The details of the convolutional layers used for the two encoders can be found in [26]. The input of the audio encoder is a 321×20 spectrogram of the noisy speech signal. Both the audio and video inputs were normalised to have zero mean and unit variance based on the statistics of the full training set.

The two feature vectors obtained as output of the video and the audio encoders are concatenated and used as input to 3 fully-connected layers, the first two having 1312 elements, and the last one 3840 elements. A leaky-ReLU is used as activation function for all the layers. The obtained vector is reshaped to the size of the audio encoder output, and fed into the audio decoder, which has 6 transposed convolutional layers that mirror the layers of the audio encoder. To avoid that the information flow is blocked by the network bottleneck, three skip connections [31] between the layers 1, 3, and 5 of the audio encoder and the corresponding mirrored layers of the decoder are added to the architecture. A ReLU output layer is applied when the target can assume only positive values (i.e. for all the IM and MA approaches except PSSA-IM and PSSA-MA), otherwise, a linear activation function is used. We clipped the target values between 0 and 10 for the IAM [8], and between -10 and 10 for the PSM. The NN outputs a 321×20 spectrogram or a mask.

The networks' weights were initialised with the Xavier approach. For training, we used the Adam optimiser with the objectives previously described. The batch size has been set to 64 and the initial learning rate to $4 \cdot 10^{-4}$. The NN was evaluated on the validation set every 2 epochs: if the validation loss increased, then the learning rate was decreased to 50% of its current value. An early stopping technique was adopted: if the validation error did not decrease for 10 epochs, the training was stopped and the model that performed the best on the validation set was used for testing.

## 3.4 Audio-Visual Enhancement and Waveform Reconstruction

To perform the enhancement of a noisy speech signal, we first applied the preprocessing described in Sec. 3.2 and forward propagated the non-overlapping audio and video segments through the NN. The outputs were concatenated to obtain the enhanced spectrogram of the full speech signal. If the output of the NN was a mask, then the enhanced spectrogram was obtained as the point-wise product between the mask and the spectrogram of the mixture. Finally, the inverse STFT was applied to reconstruct the time-domain signal using the noisy phase.

### 3.5 Evaluation and Experimental Setup

The performance of the models was evaluated in terms of perceptual evaluation of speech quality (PESQ) [32], as implemented in [1], and extended short-time objective intelligibility (ESTOI) [33]. These metrics have proven to be good estimators of speech quality and intelligibility, respectively, for the noise types considered here.

We designed our experiments to evaluate the approaches listed in Table A.1 in a range of different situations: seen and unseen speaker settings; seen and unseen noise types; different SNRs.

To have a fair comparison for the objective functions, we used the same NN architecture, cf. Sec. 3.3, and the same input, i.e. a 20-frame-long amplitude spectrum sequence, for all the approaches. The output of the NN always has the same size and can be a magnitude spectrum or a mask to be applied to the noisy spectral amplitudes in the linear domain. When the objective function required the computation of the Mel-scaled spectrum, 80 Mel-spaced frequency bins from 0 to 8 kHz are used [5].

For the DM approaches, an exponential function, which can be interpreted as a particular activation function, is applied to the NN output to impose a logarithmic compression of the output values. This makes the dynamic range narrower improving convergence behaviour during training [8]. No logarithmic compression is applied to PSSA-DM, because PSSA can assume negative values.

## 4 Results and Discussion

Table A.2 shows the results of the experiments. For the seen speaker case (left half of the table), all SE methods clearly improve the noisy signals in terms of both estimated quality and intelligibility. Regarding PESQ, LSA-DM achieves the best results overall, closely followed by the MA approaches. Among the IM techniques, the ones that operate in the log domain are the best at high SNRs, but at low SNRs the phase-aware target appears to be beneficial. There is no big difference in terms of ESTOI among the various methods, however at very low SNRs, the phase sensitive approaches do not perform as well as the other methods. This is surprising, since it was not observed in the AO setting [10, 26], and should be investigated further. Even though the approaches that operate in the Mel domain seem to have no advantages in terms of PESQ, they allow to achieve slightly higher ESTOI for both DM and IM.

For the unseen speaker case, the behaviour is similar, with small differences among the methods in terms of ESTOI. Regarding PESQ, LSA-DM is the approach showing the largest improvements among the DM ones, and it is slightly worse than PSSA-MA.

**Table A.2:** Results in terms of PESQ and ESTOI. The values are averaged across all the six noise types. The *Unproc.* rows refer to the unprocessed signals, and the *AO* columns show the average scores for models without the video encoder, trained only on the audio signals.

| PESQ | Seen Speakers | | | | | | | | | Unseen Speakers | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNR (dB) | -15 | -10 | -5 | 0 | 5 | 10 | 15 | Avg. | AO | -15 | -10 | -5 | 0 | 5 | 10 | 15 | Avg. | AO |
| Unproc. | 1.09 | 1.08 | 1.08 | 1.11 | 1.20 | 1.39 | 1.71 | 1.24 | 1.24 | 1.10 | 1.09 | 1.08 | 1.11 | 1.20 | 1.39 | 1.70 | 1.24 | 1.24 |
| STSA-DM | **1.27** | 1.35 | 1.48 | 1.65 | 1.86 | 2.08 | 2.31 | 1.71 | 1.59 | 1.13 | 1.19 | 1.30 | 1.48 | 1.73 | 1.99 | 2.24 | 1.58 | 1.57 |
| LSA-DM | 1.24 | 1.37 | **1.57** | **1.84** | **2.14** | **2.45** | **2.74** | **1.91** | **1.74** | **1.15** | **1.23** | **1.37** | **1.59** | **1.91** | **2.25** | **2.57** | **1.72** | **1.70** |
| MSA-DM | **1.27** | 1.36 | 1.49 | 1.67 | 1.87 | 2.07 | 2.28 | 1.72 | 1.58 | 1.14 | 1.20 | 1.32 | 1.51 | 1.75 | 1.99 | 2.21 | 1.59 | 1.56 |
| LMSA-DM | **1.27** | **1.39** | 1.56 | 1.78 | 2.01 | 2.18 | 2.31 | 1.79 | 1.62 | **1.15** | 1.22 | 1.34 | 1.53 | 1.77 | 1.98 | 2.14 | 1.59 | 1.59 |
| PSSA-DM | 1.24 | 1.32 | 1.44 | 1.61 | 1.82 | 2.04 | 2.25 | 1.67 | 1.62 | 1.13 | 1.18 | 1.28 | 1.45 | 1.70 | 1.94 | 2.17 | 1.55 | 1.58 |
| STSA-IM | 1.24 | 1.33 | 1.45 | 1.61 | 1.77 | 1.95 | 2.19 | 1.65 | 1.58 | 1.13 | 1.18 | 1.28 | 1.44 | 1.65 | 1.87 | 2.11 | 1.52 | 1.56 |
| LSA-IM | 1.17 | 1.25 | 1.39 | 1.60 | 1.89 | 2.19 | 2.49 | 1.71 | 1.57 | 1.13 | 1.17 | 1.28 | 1.46 | 1.72 | 2.02 | 2.34 | 1.59 | 1.57 |
| MSA-IM | 1.26 | 1.34 | 1.47 | 1.64 | 1.85 | 2.07 | 2.30 | 1.70 | **1.65** | 1.13 | 1.19 | 1.29 | 1.47 | 1.71 | 1.98 | 2.24 | 1.57 | **1.63** |
| LMSA-IM | 1.21 | 1.32 | 1.48 | **1.72** | **1.99** | **2.26** | **2.53** | **1.79** | 1.56 | 1.13 | 1.19 | 1.30 | 1.49 | **1.76** | **2.06** | **2.35** | **1.61** | 1.55 |
| PSSA-IM | **1.29** | **1.37** | **1.50** | 1.68 | 1.87 | 2.05 | 2.22 | 1.71 | **1.65** | **1.16** | **1.22** | **1.33** | **1.51** | 1.74 | 1.96 | 2.15 | 1.58 | 1.62 |
| STSA-MA | **1.31** | **1.42** | **1.57** | **1.78** | 2.02 | 2.29 | 2.58 | 1.85 | 1.62 | 1.15 | 1.21 | 1.32 | 1.52 | 1.81 | 2.15 | 2.48 | 1.66 | 1.62 |
| PSSA-MA | 1.28 | 1.38 | 1.54 | **1.78** | **2.08** | **2.40** | **2.71** | **1.88** | **1.77** | **1.18** | **1.25** | **1.38** | **1.61** | **1.95** | **2.31** | **2.63** | **1.76** | **1.76** |

| ESTOI | Seen Speakers | | | | | | | | | Unseen Speakers | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNR (dB) | -15 | -10 | -5 | 0 | 5 | 10 | 15 | Avg. | AO | -15 | -10 | -5 | 0 | 5 | 10 | 15 | Avg. | AO |
| Unproc. | 0.08 | 0.15 | 0.24 | 0.35 | 0.47 | 0.58 | 0.67 | 0.36 | 0.36 | 0.08 | 0.14 | 0.23 | 0.34 | 0.46 | 0.57 | 0.66 | 0.35 | 0.35 |
| STSA-DM | 0.35 | 0.41 | 0.49 | 0.57 | 0.64 | 0.70 | 0.74 | 0.56 | 0.48 | 0.23 | 0.29 | 0.39 | 0.49 | 0.59 | 0.67 | 0.72 | 0.48 | **0.47** |
| LSA-DM | 0.35 | 0.41 | 0.49 | 0.58 | 0.65 | **0.71** | **0.76** | 0.56 | 0.48 | 0.24 | 0.30 | 0.39 | 0.49 | 0.60 | **0.68** | **0.73** | 0.49 | **0.47** |
| MSA-DM | 0.36 | 0.42 | 0.49 | 0.57 | 0.64 | 0.70 | 0.74 | 0.56 | **0.49** | 0.24 | **0.31** | **0.40** | **0.51** | **0.61** | **0.68** | **0.73** | **0.50** | **0.47** |
| LMSA-DM | **0.37** | **0.44** | **0.51** | **0.60** | **0.66** | **0.71** | 0.75 | **0.58** | 0.48 | **0.25** | **0.31** | **0.40** | **0.51** | **0.61** | **0.68** | 0.72 | **0.50** | 0.46 |
| PSSA-DM | 0.29 | 0.36 | 0.46 | 0.56 | 0.64 | 0.70 | 0.74 | 0.53 | **0.49** | 0.19 | 0.27 | 0.37 | 0.49 | 0.60 | **0.68** | 0.72 | 0.48 | **0.47** |
| STSA-IM | 0.33 | 0.40 | 0.48 | 0.56 | 0.64 | 0.69 | 0.74 | 0.55 | 0.49 | 0.23 | 0.29 | 0.39 | **0.50** | 0.60 | 0.67 | 0.72 | 0.48 | 0.47 |
| LSA-IM | 0.33 | 0.38 | 0.46 | 0.55 | 0.63 | 0.70 | 0.75 | 0.54 | 0.46 | 0.22 | 0.28 | 0.36 | 0.46 | 0.57 | 0.66 | **0.73** | 0.47 | 0.45 |
| MSA-IM | **0.36** | **0.42** | **0.50** | 0.58 | 0.65 | 0.70 | 0.75 | **0.57** | **0.50** | **0.25** | **0.31** | **0.40** | **0.50** | 0.60 | **0.68** | **0.73** | **0.50** | **0.48** |
| LMSA-IM | **0.36** | **0.42** | **0.50** | **0.59** | **0.66** | **0.72** | **0.76** | **0.57** | 0.47 | 0.24 | 0.30 | 0.38 | 0.49 | 0.60 | **0.68** | **0.73** | 0.49 | 0.46 |
| PSSA-IM | 0.29 | 0.37 | 0.46 | 0.56 | 0.64 | 0.70 | 0.75 | 0.54 | 0.49 | 0.21 | 0.28 | 0.38 | **0.50** | **0.61** | **0.68** | **0.73** | 0.48 | 0.47 |
| STSA-MA | **0.39** | **0.45** | **0.52** | **0.60** | **0.67** | **0.72** | **0.77** | **0.59** | 0.49 | **0.26** | **0.32** | **0.41** | 0.51 | 0.62 | **0.70** | **0.75** | **0.51** | 0.48 |
| PSSA-MA | 0.29 | 0.36 | 0.46 | 0.57 | 0.66 | **0.72** | **0.77** | 0.55 | **0.50** | 0.22 | 0.29 | 0.40 | **0.52** | **0.63** | **0.70** | **0.75** | 0.50 | **0.49** |

A comparison between the seen and the unseen speakers conditions makes it clear that, at very low SNRs, knowledge of the speaker is an advantage: for example, ESTOI values at $-15$ dB SNR for the seen speakers are higher than the ones for the unseen speakers at $-10$ dB. This can be explained by the fact that the speech characteristics of an unseen speaker are harder to reconstruct by the NN, because some information of the voice attributes, e.g. pitch and timbre, cannot be easily derived from the mouth movements only.

From the results of the AO models, we observe that, generally, visual information helps in improving systems performance. The widest gap between the AV-SE systems and the respective AO-SE ones is reported for the seen speakers case. However, for unseen speakers, we see no significant improvements in terms of estimated speech quality, but for estimated speech intelligibility, the AV models are, on average, slightly better than the respective AO models. The performance difference between AO and AV models is mostly notable at low SNRs, with a gain of about 5 dB (cf. [26]).

The results for the unseen noise type (SSN) in isolation have not been reported due to space limitations, but can be found in [26]. All the systems show reasonable generalisation performance to this noise type with an improvement over the noisy signals similar to the one observed for the seen BBL noise type in terms of ESTOI.

Overall, the three best approaches among the ones investigated are LSA-DM, STSA-MA, and PSSA-MA.

# 5 Conclusion

In this study, we proposed a new taxonomy to have a uniform terminology that links classical speech enhancement methods with more recent techniques, and investigated several training targets and objective functions for audio-visual speech enhancement. We used a deep-learning-based framework to directly and indirectly learn the short time spectral amplitude of the target speech in different domains. The mask approximation approaches and the direct estimation of the log magnitude spectrum are the methods that perform the best. In contrast to the results for audio-only speech enhancement, the use of a phase-aware mask is not as effective in improving estimated intelligibility especially at low SNRs.

# References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice.* CRC Press, 2013.

[2] D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[3] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[4] I. Almajai and B. Milner, “Visually derived Wiener filters for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[5] A. Gabbay, A. Shamir, and S. Peleg, “Visual speech enhancement,” in *Proc. of Interspeech*, 2018.

[6] A. Ephrat *et al.*, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[7] T. Afouras, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” *Proc. of Interspeech*, 2018.

[8] Y. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[9] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *Proc. of GlobalSIP*, 2014.

[10] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. of ICASSP*, 2015.

[11] ——, "Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio," in *New Era for Robust Speech Recognition*. Springer, 2017, pp. 165–186.

[12] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. of HSCMA*, 2017.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[14] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[15] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[16] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. of Interspeech*, 2017.

[17] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[19] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[20] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of Interspeech*, 2013.

[21] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 133–143, 2004.

[22] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, 2008.

[23] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[25] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. of SLT*, 2016.

[26] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "On training targets and objective functions for deep-learning-based audio-visual speech enhancement - supplementary material," http://kom.aau.dk/~zt/online/icassp2019_sup_mat.pdf, 2019.

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of CVPR*, 2001.

[28] J. Shi and C. Tomasi, "Good features to track," in *Proc. of CVPR*, 1994.

[29] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of IJCAI*, 1981.

[30] C. Tomasi and T. Kanade, "Detection and tracking of point features," *Technical Report CMU-CS-91-132*, 1991.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, 2015.

[32] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP*, 2001.

[33] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

# A Supplementary Material

This appendix shows extended experiments, but it is not formally part of the publication.

**Table A.3:** Convolutional layers of the audio and video encoders.

| | Video Encoder | | | Audio Encoder | | |
|---|---|---|---|---|---|---|
| Layer | Filters | Kernel | Stride | Filters | Kernel | Stride |
| 1 | 128 | 5×5 | 1×1 | 64 | 5×5 | 2×2 |
| 2 | 128 | 5×5 | 1×1 | 64 | 4×4 | 2×1 |
| 3 | 256 | 3×3 | 1×1 | 128 | 4×4 | 2×2 |
| 4 | 256 | 3×3 | 1×1 | 128 | 2×2 | 2×1 |
| 5 | 512 | 3×3 | 1×1 | 128 | 2×2 | 2×1 |
| 6 | 512 | 3×3 | 1×1 | 128 | 2×2 | 2×1 |

**Table A.4:** PESQ and ESTOI results in the audio-visual setting where the scores are averaged across all the SNRs.

| PESQ | Seen Speakers | | | | | | Unseen Speakers | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Noise Type | BBL | BUS | CAF | PED | STR | SSN | BBL | BUS | CAF | PED | STR | SSN |
| Unproc. | 1.32 | 1.44 | 1.30 | 1.30 | 1.30 | 1.29 | 1.31 | 1.44 | 1.31 | 1.30 | 1.30 | 1.28 |
| STSA-DM | 1.67 | 1.97 | 1.77 | 1.72 | 1.79 | 1.58 | 1.53 | 1.84 | 1.64 | 1.59 | 1.67 | 1.50 |
| LSA-DM | **1.85** | **2.22** | **1.98** | **1.90** | **2.03** | **1.71** | **1.65** | **2.05** | **1.81** | **1.75** | **1.88** | **1.57** |
| MSA-DM | 1.65 | 1.94 | 1.78 | 1.72 | 1.79 | 1.58 | 1.52 | 1.82 | 1.64 | 1.60 | 1.68 | 1.50 |
| LMSA-DM | 1.72 | 1.95 | 1.83 | 1.78 | 1.86 | 1.60 | 1.53 | 1.77 | 1.63 | 1.60 | 1.67 | 1.45 |
| PSSA-DM | 1.61 | 1.91 | 1.71 | 1.66 | 1.76 | 1.60 | 1.49 | 1.78 | 1.58 | 1.55 | 1.65 | 1.51 |
| STSA-IM | 1.60 | 1.90 | 1.72 | 1.66 | 1.73 | 1.54 | 1.47 | 1.78 | 1.58 | 1.54 | 1.62 | 1.46 |
| LSA-IM | 1.71 | 2.00 | 1.76 | 1.72 | 1.83 | 1.57 | 1.56 | **1.89** | 1.66 | 1.62 | 1.73 | 1.46 |
| MSA-IM | 1.67 | 1.96 | 1.78 | 1.72 | 1.77 | 1.56 | 1.52 | 1.85 | 1.64 | 1.60 | 1.66 | 1.48 |
| LMSA-IM | **1.78** | **2.04** | **1.85** | **1.79** | **1.89** | **1.61** | **1.58** | 1.88 | **1.69** | **1.64** | **1.74** | 1.48 |
| PSSA-IM | 1.64 | 1.96 | 1.74 | 1.70 | 1.79 | 1.58 | 1.51 | 1.83 | 1.61 | 1.58 | 1.67 | **1.51** |
| STSA-MA | 1.84 | 2.09 | **1.94** | **1.90** | 1.98 | 1.64 | 1.64 | 1.93 | 1.76 | 1.72 | 1.81 | 1.51 |
| PSSA-MA | **1.85** | **2.19** | 1.93 | **1.90** | **2.02** | **1.68** | **1.71** | **2.07** | **1.82** | **1.79** | **1.92** | **1.59** |
| ESTOI | Seen Speakers | | | | | | Unseen Speakers | | | | | |
| Noise Type | BBL | BUS | CAF | PED | STR | SSN | BBL | BUS | CAF | PED | STR | SSN |
| Unproc. | 0.33 | 0.46 | 0.38 | 0.33 | 0.37 | 0.33 | 0.33 | 0.45 | 0.37 | 0.33 | 0.37 | 0.32 |
| STSA-DM | 0.51 | 0.62 | 0.57 | 0.54 | 0.57 | 0.49 | 0.42 | 0.57 | 0.50 | 0.47 | 0.51 | 0.42 |
| LSA-DM | 0.52 | 0.63 | 0.58 | 0.55 | 0.58 | 0.49 | 0.42 | **0.58** | 0.51 | **0.48** | **0.52** | 0.42 |
| MSA-DM | 0.52 | 0.63 | 0.58 | 0.55 | 0.58 | **0.50** | **0.44** | **0.58** | 0.51 | **0.48** | **0.52** | **0.44** |
| LMSA-DM | **0.54** | **0.64** | **0.59** | **0.57** | **0.59** | 0.49 | **0.44** | **0.58** | **0.52** | **0.48** | **0.52** | 0.42 |
| PSSA-DM | 0.49 | 0.61 | 0.54 | 0.51 | 0.54 | 0.47 | 0.41 | 0.56 | 0.49 | 0.45 | 0.49 | 0.42 |
| STSA-IM | 0.51 | 0.62 | 0.56 | 0.53 | 0.56 | 0.48 | 0.42 | 0.57 | 0.50 | 0.47 | 0.51 | **0.43** |
| LSA-IM | 0.51 | 0.61 | 0.56 | 0.53 | 0.56 | 0.48 | 0.42 | 0.56 | 0.49 | 0.46 | 0.50 | 0.40 |
| MSA-IM | 0.53 | 0.63 | 0.58 | 0.55 | 0.58 | **0.50** | **0.44** | **0.58** | **0.52** | **0.48** | **0.52** | **0.43** |
| LMSA-IM | **0.54** | **0.64** | **0.59** | **0.56** | **0.59** | **0.50** | 0.43 | **0.58** | 0.51 | **0.48** | **0.52** | 0.41 |
| PSSA-IM | 0.49 | 0.61 | 0.55 | 0.51 | 0.55 | 0.48 | 0.42 | 0.57 | 0.50 | 0.46 | 0.50 | **0.43** |
| STSA-MA | **0.55** | **0.65** | **0.60** | **0.57** | **0.60** | **0.52** | **0.45** | **0.60** | **0.53** | **0.49** | **0.54** | **0.44** |
| PSSA-MA | 0.50 | 0.62 | 0.56 | 0.52 | 0.56 | 0.48 | 0.44 | 0.59 | 0.52 | 0.48 | 0.52 | **0.44** |

**Table A.5:** PESQ and ESTOI results in the audio-visual setting where the scores are averaged across all the noise types.

| PESQ | Seen Speakers | | | | | | | | | Unseen Speakers | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNR (dB) | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 |
| Unproc. | 1.10 | 1.09 | 1.08 | 1.08 | 1.11 | 1.20 | 1.39 | 1.71 | 2.14 | 1.10 | 1.10 | 1.09 | 1.08 | 1.11 | 1.20 | 1.39 | 1.70 | 2.13 |
| STSA-DM | **1.21** | **1.27** | 1.35 | 1.48 | 1.65 | 1.86 | 2.08 | 2.31 | 2.53 | 1.11 | 1.13 | 1.19 | 1.30 | 1.48 | 1.73 | 1.99 | 2.24 | 2.47 |
| LSA-DM | 1.17 | 1.24 | 1.37 | **1.57** | **1.84** | **2.14** | **2.45** | **2.74** | **3.01** | **1.12** | **1.15** | **1.23** | **1.37** | **1.59** | **1.91** | **2.25** | **2.57** | **2.86** |
| MSA-DM | **1.21** | **1.27** | 1.36 | 1.49 | 1.67 | 1.87 | 2.07 | 2.28 | 2.47 | 1.11 | 1.14 | 1.20 | 1.32 | 1.51 | 1.75 | 1.99 | 2.21 | 2.41 |
| LMSA-DM | 1.20 | **1.27** | **1.39** | 1.56 | 1.78 | 2.01 | 2.18 | 2.31 | 2.40 | **1.12** | **1.15** | 1.22 | 1.34 | 1.53 | 1.77 | 1.98 | 2.14 | 2.25 |
| PSSA-DM | 1.19 | 1.24 | 1.32 | 1.44 | 1.61 | 1.82 | 2.04 | 2.25 | 2.46 | 1.11 | 1.13 | 1.18 | 1.28 | 1.45 | 1.70 | 1.94 | 2.17 | 2.38 |
| STSA-IM | 1.19 | 1.24 | 1.33 | 1.45 | 1.61 | 1.77 | 1.95 | 2.19 | 2.49 | 1.10 | 1.13 | 1.18 | 1.28 | 1.44 | 1.65 | 1.87 | 2.11 | 2.43 |
| LSA-IM | 1.14 | 1.17 | 1.25 | 1.39 | 1.60 | 1.89 | 2.19 | 2.49 | **2.77** | 1.11 | 1.13 | 1.17 | 1.28 | 1.46 | 1.72 | 2.02 | 2.34 | **2.64** |
| MSA-IM | 1.20 | 1.26 | 1.34 | 1.47 | 1.64 | 1.85 | 2.07 | 2.30 | 2.56 | 1.11 | 1.13 | 1.19 | 1.29 | 1.47 | 1.71 | 1.98 | 2.24 | 2.51 |
| LMSA-IM | 1.16 | 1.21 | 1.32 | 1.48 | **1.72** | **1.99** | **2.26** | **2.53** | **2.77** | 1.11 | 1.13 | 1.19 | 1.30 | 1.49 | **1.76** | **2.06** | **2.35** | 2.61 |
| PSSA-IM | **1.23** | **1.29** | **1.37** | **1.50** | 1.68 | 1.87 | 2.05 | 2.22 | 2.41 | **1.13** | **1.16** | **1.22** | **1.33** | **1.51** | 1.74 | 1.96 | 2.15 | 2.36 |
| STSA-MA | **1.24** | **1.31** | **1.42** | **1.57** | **1.78** | 2.02 | 2.29 | 2.58 | 2.88 | 1.12 | 1.15 | 1.21 | 1.32 | 1.52 | 1.81 | 2.15 | 2.48 | 2.80 |
| PSSA-MA | 1.21 | 1.28 | 1.38 | 1.54 | **1.78** | **2.08** | **2.40** | **2.71** | **2.98** | **1.14** | **1.18** | **1.25** | **1.38** | **1.61** | **1.95** | **2.31** | **2.63** | **2.90** |

| ESTOI | Seen Speakers | | | | | | | | | Unseen Speakers | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNR (dB) | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 |
| Unproc. | 0.04 | 0.08 | 0.15 | 0.24 | 0.35 | 0.47 | 0.58 | 0.67 | 0.74 | 0.04 | 0.08 | 0.14 | 0.23 | 0.34 | 0.46 | 0.57 | 0.66 | 0.73 |
| STSA-DM | 0.30 | 0.35 | 0.41 | 0.49 | 0.57 | 0.64 | 0.70 | 0.74 | 0.77 | 0.19 | 0.23 | 0.29 | 0.39 | 0.49 | 0.59 | 0.67 | 0.72 | **0.76** |
| LSA-DM | 0.30 | 0.35 | 0.41 | 0.49 | 0.58 | 0.65 | **0.71** | **0.76** | **0.79** | 0.20 | 0.24 | 0.30 | 0.39 | 0.49 | 0.60 | **0.68** | **0.73** | **0.76** |
| MSA-DM | 0.31 | 0.36 | 0.42 | 0.49 | 0.57 | 0.64 | 0.70 | 0.74 | 0.77 | 0.20 | 0.24 | **0.31** | **0.40** | **0.51** | **0.61** | **0.68** | **0.73** | **0.76** |
| LMSA-DM | **0.33** | **0.37** | **0.44** | **0.51** | **0.60** | **0.66** | **0.71** | 0.75 | 0.77 | **0.21** | **0.25** | **0.31** | **0.40** | **0.51** | **0.61** | **0.68** | 0.72 | 0.75 |
| PSSA-DM | 0.23 | 0.29 | 0.36 | 0.46 | 0.56 | 0.64 | 0.70 | 0.74 | 0.77 | 0.14 | 0.19 | 0.27 | 0.37 | 0.49 | 0.60 | **0.68** | 0.72 | **0.76** |
| STSA-IM | 0.28 | 0.33 | 0.40 | 0.48 | 0.56 | 0.64 | 0.69 | 0.74 | 0.78 | 0.19 | 0.23 | 0.29 | 0.39 | **0.50** | 0.60 | 0.67 | 0.72 | 0.76 |
| LSA-IM | 0.29 | 0.33 | 0.38 | 0.46 | 0.55 | 0.63 | 0.70 | 0.75 | **0.79** | 0.19 | 0.22 | 0.28 | 0.36 | 0.46 | 0.57 | 0.66 | **0.73** | **0.77** |
| MSA-IM | **0.32** | **0.36** | **0.42** | **0.50** | 0.58 | 0.65 | 0.70 | 0.75 | 0.78 | **0.21** | **0.25** | **0.31** | **0.40** | **0.50** | 0.60 | **0.68** | **0.73** | **0.77** |
| LMSA-IM | **0.32** | **0.36** | **0.42** | **0.50** | **0.59** | **0.66** | **0.72** | **0.76** | **0.79** | 0.20 | 0.24 | 0.30 | 0.38 | 0.49 | 0.60 | **0.68** | **0.73** | **0.77** |
| PSSA-IM | 0.24 | 0.29 | 0.37 | 0.46 | 0.56 | 0.64 | 0.70 | 0.75 | 0.78 | 0.16 | 0.21 | 0.28 | 0.38 | **0.50** | **0.61** | **0.68** | **0.73** | **0.77** |
| STSA-MA | **0.35** | **0.39** | **0.45** | **0.52** | **0.60** | **0.67** | **0.72** | **0.77** | **0.80** | **0.22** | **0.26** | **0.32** | **0.41** | 0.51 | 0.62 | **0.70** | **0.75** | 0.78 |
| PSSA-MA | 0.24 | 0.29 | 0.36 | 0.46 | 0.57 | 0.66 | **0.72** | **0.77** | **0.80** | 0.17 | 0.22 | 0.29 | 0.40 | **0.52** | **0.63** | **0.70** | **0.75** | **0.79** |

**Table A.6:** PESQ and ESTOI results in the audio-only setting where the scores are averaged across all the noise types.

| PESQ | Seen Speakers | | | | | | | | | Unseen Speakers | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNR (dB) | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 |
| Unproc. | 1.10 | 1.09 | 1.08 | 1.08 | 1.11 | 1.20 | 1.39 | 1.71 | 2.14 | 1.10 | 1.10 | 1.09 | 1.08 | 1.11 | 1.20 | 1.39 | 1.70 | 2.13 |
| STSA-DM | 1.11 | 1.14 | 1.20 | 1.32 | 1.51 | 1.74 | 1.97 | 2.21 | 2.44 | **1.11** | 1.13 | 1.19 | 1.30 | 1.49 | 1.72 | 1.95 | 2.18 | 2.40 |
| LSA-DM | 1.10 | 1.14 | 1.22 | **1.37** | **1.61** | **1.94** | **2.28** | **2.60** | **2.87** | 1.10 | **1.14** | **1.21** | **1.34** | **1.58** | **1.89** | **2.22** | **2.52** | **2.77** |
| MSA-DM | 1.10 | 1.14 | 1.20 | 1.33 | 1.51 | 1.74 | 1.96 | 2.17 | 2.37 | **1.11** | 1.13 | 1.19 | 1.31 | 1.49 | 1.72 | 1.94 | 2.15 | 2.34 |
| LMSA-DM | 1.09 | 1.12 | 1.18 | 1.32 | 1.54 | 1.82 | 2.08 | 2.29 | 2.42 | 1.10 | 1.12 | 1.18 | 1.30 | 1.52 | 1.78 | 2.01 | 2.19 | 2.31 |
| PSSA-DM | **1.12** | **1.16** | **1.23** | 1.35 | 1.55 | 1.77 | 2.01 | 2.24 | 2.45 | **1.11** | **1.14** | **1.21** | 1.32 | 1.50 | 1.74 | 1.98 | 2.20 | 2.41 |
| STSA-IM | 1.10 | 1.14 | 1.20 | 1.32 | 1.50 | 1.71 | 1.95 | 2.23 | 2.57 | 1.10 | 1.13 | 1.18 | 1.29 | 1.47 | 1.69 | 1.92 | 2.20 | 2.54 |
| LSA-IM | 1.09 | 1.11 | 1.15 | 1.25 | 1.43 | 1.70 | 2.02 | **2.36** | **2.70** | 1.09 | 1.11 | 1.15 | 1.25 | 1.42 | 1.69 | 2.01 | **2.34** | **2.67** |
| MSA-IM | 1.11 | 1.14 | 1.21 | 1.34 | 1.56 | **1.83** | **2.10** | **2.36** | 2.63 | 1.11 | 1.14 | 1.20 | 1.32 | 1.53 | **1.80** | **2.08** | **2.34** | 2.61 |
| LMSA-IM | 1.09 | 1.11 | 1.16 | 1.28 | 1.47 | 1.71 | 1.97 | 2.21 | 2.43 | 1.09 | 1.11 | 1.17 | 1.27 | 1.46 | 1.70 | 1.95 | 2.17 | 2.38 |
| PSSA-IM | **1.13** | **1.17** | **1.25** | **1.39** | **1.60** | **1.83** | 2.05 | 2.27 | 2.51 | **1.13** | **1.16** | **1.23** | **1.36** | **1.56** | **1.80** | 2.02 | 2.23 | 2.47 |
| STSA-MA | 1.09 | 1.12 | 1.17 | 1.27 | 1.47 | 1.76 | 2.11 | 2.47 | 2.82 | 1.10 | 1.12 | 1.17 | 1.26 | 1.45 | 1.75 | 2.10 | 2.45 | 2.80 |
| PSSA-MA | **1.13** | **1.17** | **1.24** | **1.38** | **1.62** | **1.96** | **2.34** | **2.68** | **2.98** | **1.12** | **1.16** | **1.23** | **1.36** | **1.61** | **1.96** | **2.33** | **2.65** | **2.94** |

| ESTOI | Seen Speakers | | | | | | | | | Unseen Speakers | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNR (dB) | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 | -20 | -15 | -10 | -5 | 0 | 5 | 10 | 15 | 20 |
| Unproc. | 0.04 | 0.08 | 0.15 | 0.24 | 0.35 | 0.47 | 0.58 | 0.67 | 0.74 | 0.04 | 0.08 | 0.14 | 0.23 | 0.34 | 0.46 | 0.57 | 0.66 | 0.73 |
| STSA-DM | 0.10 | 0.17 | 0.27 | 0.39 | 0.51 | 0.61 | 0.68 | 0.73 | 0.77 | 0.10 | 0.16 | 0.25 | 0.37 | 0.50 | 0.60 | **0.68** | 0.72 | **0.76** |
| LSA-DM | **0.11** | 0.17 | 0.26 | 0.38 | 0.51 | **0.62** | **0.69** | **0.74** | **0.78** | **0.11** | **0.17** | 0.25 | 0.36 | 0.49 | 0.60 | **0.68** | **0.73** | **0.76** |
| MSA-DM | **0.11** | **0.18** | **0.28** | **0.40** | 0.52 | 0.61 | 0.68 | 0.73 | 0.77 | **0.11** | **0.17** | **0.26** | **0.38** | 0.50 | 0.60 | **0.68** | 0.72 | **0.76** |
| LMSA-DM | **0.11** | 0.17 | 0.26 | 0.38 | 0.51 | 0.61 | 0.68 | 0.73 | 0.76 | **0.11** | 0.16 | 0.25 | 0.36 | 0.49 | 0.60 | 0.67 | 0.72 | 0.74 |
| PSSA-DM | 0.10 | 0.17 | 0.27 | **0.40** | **0.53** | **0.62** | **0.69** | **0.74** | 0.77 | 0.10 | 0.16 | 0.25 | **0.38** | **0.51** | **0.61** | **0.68** | **0.73** | **0.76** |
| STSA-IM | **0.11** | **0.18** | **0.28** | 0.40 | 0.52 | 0.62 | 0.69 | 0.74 | **0.78** | 0.10 | 0.17 | 0.26 | 0.38 | 0.50 | 0.60 | 0.68 | **0.73** | **0.77** |
| LSA-IM | 0.09 | 0.15 | 0.24 | 0.35 | 0.47 | 0.58 | 0.67 | 0.73 | **0.78** | 0.09 | 0.15 | 0.23 | 0.34 | 0.46 | 0.57 | 0.66 | 0.72 | **0.77** |
| MSA-IM | **0.11** | **0.18** | **0.28** | **0.41** | **0.53** | **0.63** | 0.69 | 0.74 | **0.78** | **0.11** | **0.18** | **0.27** | **0.39** | **0.51** | **0.61** | **0.69** | **0.73** | **0.77** |
| LMSA-IM | 0.10 | 0.16 | 0.25 | 0.37 | 0.49 | 0.60 | 0.68 | 0.73 | 0.77 | 0.10 | 0.16 | 0.24 | 0.35 | 0.47 | 0.58 | 0.67 | 0.72 | 0.76 |
| PSSA-IM | 0.10 | 0.17 | 0.27 | 0.40 | 0.52 | 0.62 | **0.70** | **0.75** | **0.78** | 0.10 | 0.16 | 0.25 | 0.38 | 0.50 | **0.61** | **0.69** | **0.73** | **0.77** |
| STSA-MA | **0.11** | **0.18** | **0.27** | 0.39 | 0.51 | 0.62 | 0.70 | 0.75 | 0.79 | **0.11** | **0.17** | **0.26** | 0.37 | 0.50 | 0.61 | 0.69 | 0.74 | 0.78 |
| PSSA-MA | **0.11** | **0.18** | **0.27** | **0.40** | **0.53** | **0.64** | **0.71** | **0.76** | **0.80** | 0.10 | **0.17** | **0.26** | **0.38** | **0.51** | **0.63** | **0.70** | **0.75** | **0.79** |

# Paper B

## Effects of Lombard Reflex on the Performance of Deep-Learning-Based Audio-Visual Speech Enhancement Systems

Daniel Michelsanti, Zheng-Hua Tan, Sigurdur Sigurdsson,
Jesper Jensen

The paper has been published in
*Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 6615–6619, 2019.

# Abstract

*Humans tend to change their way of speaking when they are immersed in a noisy environment, a reflex known as Lombard effect. Current speech enhancement systems based on deep learning do not usually take into account this change in the speaking style, because they are trained with neutral (non-Lombard) speech utterances recorded under quiet conditions to which noise is artificially added. In this paper, we investigate the effects that the Lombard reflex has on the performance of audio-visual speech enhancement systems based on deep learning. The results show that a gap in the performance of as much as approximately 5 dB between the systems trained on neutral speech and the ones trained on Lombard speech exists. This indicates the benefit of taking into account the mismatch between neutral and Lombard speech in the design of audio-visual speech enhancement systems.*

# 1 Introduction

Background noise can make vocal communication hard, because it degrades the speech of interest. However, speakers instinctively react to the presence of background noise and change their speaking style to maintain their speech intelligible. This reflex is known as *Lombard effect* [1], and it is characterised, acoustically, by an increase in speech sound level [2], a longer word duration [3], and modifications of the speech spectrum [2], and, visually, by a speech hyper-articulation [4].

In particularly challenging situations, e.g. when the listener is hearing impaired and the noise level is high, this natural change of speaking style might not suffice to guarantee an effective communication. Hence, there is a need to reduce the negative effects of background noise on speech quality and intelligibility with speech enhancement (SE) techniques. SE is important in several applications, and proposed approaches range from classical statistical model-based methods [5], to deep-learning-based ones [6]. These techniques use only audio signals to perform enhancement, and we refer to them as audio-only SE (AO-SE) systems.

During speech production, the movements of some articulatory organs, e.g. lips and jaw, might be visible to the listener, enhancing speech perception [7, 8]. Exploiting the information conveyed by these visual cues, which are independent of the acoustical environment where SE systems operate, leads to systems that are more robust than the AO-SE ones to background noise. This has already been shown in early work on audio-visual SE (AV-SE) [9]. More complex frameworks have been proposed later, e.g. [10], in which a voice activity detector and phoneme-specific methods are used to estimate noise and clean speech statistics for a visually derived Wiener filter. Very recently, deep-learning-based techniques have also been adopted to solve the

AV-SE task [11–14].

AV-SE systems are likely to be deployed in acoustic situations where AO-SE systems underperform or fail, e.g. in situations where the background noise level is high and the Lombard effect is clearly present. In other words, the typical input to AV-SE systems is Lombard speech in noise. However, existing SE systems usually ignore this effect, being trained with clean speech signals recorded in quiet to which noise is artificially added. The mismatch between the neutral and the Lombard speaking styles can lead to sub-optimal performance of audio-only-based speaker [15] and speech recognition [2] systems. Only a few works investigate the impact of the Lombard effect on visual [16, 17] and audio-visual [16] automatic speech recognition, but, to the best knowledge of the authors, no studies have been conducted for AV-SE systems.

The aim of the current paper is to examine to which extent a deep-learning-based AV-SE system trained on neutral speech can effectively enhance Lombard speech. This is important to understand, because if such a system can model well Lombard speech, then there is no need to include in the training procedure speech recorded in Lombard conditions, which is usually hard to obtain. Specifically, we are interested in answering the following research questions:

1. Is an AV-SE system trained on neutral speech able to improve Lombard speech?

2. Does a performance gap exist between a system trained on Lombard speech and a system trained on neutral speech when tested on Lombard speech from speakers that have been observed during training (seen speakers)?

3. Is a performance gap still present for speakers that have not been observed during training (unseen speakers)?

The last two questions are relevant to understand the impact of inter-speaker differences of Lombard speech on SE. We expect that the system trained on Lombard speech enhances the Lombard speech of a seen speaker better than the system trained on neutral speech, because it should model well the Lombard speaking style of that speaker. However, the system trained on Lombard speech may have difficulties in generalising to unseen speakers, because the characteristics of the Lombard speech of one person might significantly differ from the characteristics of the Lombard speech of another person [2, 17].

# 2 Audio-Visual Corpus and Noise Data

The dataset used in this study is the English language Lombard GRID corpus [18], consisting of audio and visual (frontal and profile view[1]) recordings from 54 subjects (24 males and 30 females). The audio and video channels are temporally aligned. Each speaker is recorded, while pronouncing 50 unique six-word sentences, whose syntax is identical to the one in GRID [19], in each of two conditions: non-Lombard (NL) and Lombard (L). In the condition NL, the speakers are recorded with a microphone placed at 30 cm in front of their mouth and two cameras mounted on a helmet worn by them. The condition L replicates the same setup, but it simulates the Lombard effect by presenting speech shaped noise (SSN) at a level of 80 dB sound pressure level (SPL) through headphones. In addition, the speakers are provided with a carefully adjusted self-monitoring feedback, while reading aloud some sentences to a listener, who asks to repeat the utterances from time to time in order to simulate possible miscomprehensions of speech in noise. This scenario allows to take into account the two factors responsible for the Lombard adaptation: first, speakers tend to regulate their vocal effort based on the auditory feedback, i.e. they involuntarily react to the perceived level of their own speech [17]; secondly, they change their speaking style to communicate better with others [20, 21].

The impact of the noise type on the Lombard effect is currently unclear. While some studies have found no evidence to support a systematic active response of speakers to the spectral characteristics of the noise [22, 23], Hansen and Varadarajanare [15] indicate the presence of differences across noise types in the way that Lombard effect occurs. Following this finding, we use SSN, since this is the noise type that was presented to the speakers of the Lombard GRID corpus. The noise was generated as reported in [24].

The audio-visual corpus and the noise data are used to build training, validation and test sets as explained in Sec. 4.

# 3 Methods

The goal of many SE systems is to estimate the clean signal $x(n)$, given a mixture $y(n) = x(n) + d(n)$, where $d(n)$ is an additive noise signal and $n$ denotes the discrete-time index. Usually, the SE problem is tackled in the time-frequency (TF) domain, where the additive noise model is expressed as $Y(k,l) = X(k,l) + D(k,l)$, with $k$ indicating the frequency bin index, $l$ denoting the time frame index, and $Y(k,l)$, $X(k,l)$, and $D(k,l)$ being the short-time Fourier transform (STFT) coefficients of $y(n)$, $x(n)$, and $d(n)$, respectively.

[1] In this study, the audio and the frontal view video recordings are used.

Since the estimation of the phase of the clean STFT coefficients, $X(k,l)$, with a neural network is hard [25], enhancement can be performed by estimating $A_{k,l} = |X(k,l)|$ from $R_{k,l} = |Y(k,l)|$. The time-domain signal is obtained using the estimated clean magnitude spectrum and the noisy phase in an inverse STFT procedure.

In this study, we use a mask approximation (MA) approach, where a neural network is trained to learn the ideal amplitude mask (IAM), defined as $M_{k,l}^{\mathrm{IAM}} = \frac{A_{k,l}}{R_{k,l}}$, with the following objective function:

$$J = \frac{1}{TF} \sum_{k,l} \left( M_{k,l}^{\mathrm{IAM}} - \widehat{M}_{k,l} \right)^2, \tag{B.1}$$

where $\widehat{M}_{k,l}$ is the output of the network, $k \in \{1, \ldots, F\}$, $l \in \{1, \ldots, T\}$, and $TF$ is the size of the training target. This objective function showed best performance in several conditions in a comparison study [26], where a range of targets and cost functions, used to train a deep-learning-based AV-SE system, are investigated.

## 3.1 Preprocessing

The audio signals, which have a sample rate of 16 kHz, are peak-normalised to 1 per signal. Then, a 640-point STFT is applied, using a 640-sample-long Hamming window and a hop size of 160 samples. Due to spectral symmetry, we consider only the 321 bins that cover the positive frequencies.

To preprocess the video signals, resampled at a frame rate of 25 fps, we make use of the detection and alignment algorithms implemented in the dlib toolkit [27]. In particular, for each frame we detect the face with a linear support vector machine (SVM) on histogram of oriented gradients (HOG) features, and track it across the frames with a Kalman filter. Then, the detected face is aligned using 5 landmark points that identify the corners of the eyes and the bottom of the nose and scaled to 256 × 256 pixels. Finally, the 128 × 128-pixel region around the mouth is extracted.

## 3.2 Architecture and Training Procedure

The neural network architecture, inspired by [12] and identical to [26], consists of four blocks: a video encoder, an audio encoder, a fusion subnetwork, and an audio decoder.

The video encoder takes as input 5 consecutive grayscale frames of the mouth region, corresponding to 200 ms. Six convolutional layers are applied, and each of them is followed by: leaky-ReLU activation, batch normalisation, 2×2 strided max-pooling with a 2×2 kernel, and dropout with a probability of 25%. The audio encoder is fed with a 20-frame-long magnitude spectrogram, corresponding to 200 ms, and consists of 6 convolutional layers, followed by leaky-ReLU activation and batch normalisation. Further details regarding the convolutional layers of the encoders are shown in Table B.1. The inputs of the encoders are normalised to have zero mean and unit variance based on the training set statistics.

**Table B.1:** Convolutional layers of the audio and video encoders. For the video encoder, a 1×1 stride is always applied.

| | Video Encoder | | Audio Encoder | | |
|---|---|---|---|---|---|
| Layer | # Filters | Kernel | # Filters | Kernel | Stride |
| 1 | 128 | 5×5 | 64 | 5×5 | 2×2 |
| 2 | 128 | 5×5 | 64 | 4×4 | 2×1 |
| 3 | 256 | 3×3 | 128 | 4×4 | 2×2 |
| 4 | 256 | 3×3 | 128 | 2×2 | 2×1 |
| 5 | 512 | 3×3 | 128 | 2×2 | 2×1 |
| 6 | 512 | 3×3 | 128 | 2×2 | 2×1 |

The outputs of the encoders are concatenated and fed into the fusion subnetwork, consisting of 3 fully connected layers, the first 2 with 1312 neurons and the last one with 3840. A leaky-ReLU activation is used for all the layers.

The audio decoder takes as input the result vector of the fusion subnetwork and processes it with 6 transposed convolutional layers that mirror the audio encoder ones. The architecture has 3 skip connections between layers 1, 3, and 5 of the audio encoder and the corresponding layers of the decoder. The output layer uses ReLU activation. In the end, a 321×20 mask matrix, which estimates the IAM, is obtained. The target values are clipped between 0 and 10 [28].

After the initialisation of the weights with the Xavier approach, the network is trained for 50 epochs adopting the Adam optimiser, with the objective function in Eq. (B.1), a batch size of 64 and an initial learning rate of $4 \cdot 10^{-4}$. The network is evaluated on the validation set every epoch, and the learning rate is halved, if the validation loss increases. For testing, we use the network that performs the best on the validation set across the 50 epochs to avoid overfitting issues.

Besides this AV-SE system, we also train an AO-SE and a video-only SE (VO-SE) architectures, obtained by removing the video encoder or the audio encoder, respectively, from the AV-SE system.

## 3.3 Audio-Visual Speech Enhancement

The enhancement of the noisy signals is performed in three steps. First, the preprocessed non-overlapping audio and video (or only one of the two modalities when AO-SE and VO-SE architectures are used) sequences are forward propagated through the network. Then, the resulting masks $\widehat{M}_{k,l}$ are concatenated and point-wise multiplied with the complex-valued STFT spectrogram of the mixture. Finally, the enhanced signals are reconstructed with the inverse STFT.

# 4 Experiments

This section describes the experimental setup and the evaluation measures used in this study. The training, validation and test data have been allocated differently between the seen and the unseen speaker cases, since the amount of speech material available to train deep-learning-based systems is relatively small.

## 4.1 Seen Speaker Case

Each person may exhibit a different Lombard speaking style [2, 17]. Modelling these differences could be performed by training several speaker-dependent SE systems. However, this choice requires a large amount of speech data for each speaker. Instead, we adopt a multi-speaker setup and train one AV-SE system with 54 speakers for each of the two conditions of the database, L and NL, obtaining two models, AV-L and AV-NL, respectively.

For AV-L, the utterances of the database recorded in condition L are randomly shuffled and organised into: a test set with 10 utterances for each speaker; a validation set consisting of 5 utterances per speaker; and a training set with the remaining material.

For AV-NL, the training and validation sets are arranged by picking the neutral utterances corresponding to the Lombard utterances used for the training and validation of AV-L. The test set is the same as the one used for AV-L, because we are interested in investigating the enhancement potential of AV-NL in condition L and compare it with the AV-L performance. We also train two VO-SE systems, for the conditions L and NL, and two AO-SE systems, for the conditions L and NL, obtaining four additional models: VO-L, VO-NL, AO-L, and AO-NL, respectively. This should be considered as an additional aspect to the research questions introduced in Sec. 1, which allows us to understand the contribution of each modality to the enhancement of Lombard speech.

## 4.2 Unseen Speaker Case

Generalisation of SE systems to unseen speakers is important, especially in applications where it is hard to collect speech data to train a speaker-dependent system. For this reason, we want to examine whether a system trained with Lombard speech can generalise well, i.e. better than a system trained on neutral speech, to the Lombard speech of an unseen speaker. We perform a 6-fold cross-validation by training 6 AV-SE systems on utterances in condition L (5 per speaker for validation and the rest for training) of 45 speakers, and testing them on utterances in condition L of 9 unseen speakers (4 males and 5 females). We refer to each of these models as AV-L*. The same procedure is applied for training and validation of 6 AV-SE systems using the corresponding utterances recorded in condition NL. The obtained models are denoted as AV-NL*.

All the models used in this study are summarised in Table B.2.

**Table B.2:** Models used in this study for the seen and the unseen (indicated with a *) speaker cases.

| | Training Material | |
|---|---|---|
| Modality | Non-Lombard Speech | Lombard Speech |
| Vision | VO-NL | VO-L |
| Audio | AO-NL | AO-L |
| Audio-visual | AV-NL / AV-NL* | AV-L / AV-L* |

## 4.3 Additive Noise Levels

To construct the training, the validation, and the test sets for all the models, the speech signals from the Lombard GRID database are mixed with additive SSN at 6 signal to noise ratios (SNRs), in uniform steps between -20 dB and 5 dB. The SNR range has been chosen due to the following considerations:

1. Current SE systems are trained on noisy signals in which noise is added to the clean signals at several SNRs to ensure robustness to different noise levels. For this reason, we do not train SNR-specific systems.

2. In the Lombard GRID database, the energy difference between Lombard and neutral utterances is between 3 dB and 13 dB [17]. If we assume that the listener is immersed in SSN at 80 dB SPL, like in the recording conditions of the Lombard GRID database, and that the conversational speech level is between 60 and 70 dB SPL [29, 30], the SNR is between -17 dB and 3 dB. The slightly wider SNR range used in the experiments (between -20 dB and 5 dB) is chosen to take into account the possible speech level variations due to the distance of the listener from the speaker.

## 4.4 Evaluation Metrics

The performance of all the models are evaluated in terms of two objective measures, namely perceptual evaluation of speech quality (PESQ) [31] as implemented in [5], and extended short-time objective intelligibility (ESTOI)

[32], because they are good estimators of speech quality and intelligibility, respectively. PESQ ranges from -0.5 to 4.5, where high values correspond to high speech quality. For ESTOI, whose range is practically between 0 and 1, higher scores correspond to higher speech intelligibility.

# 5 Results and Discussion

Figs. B.1 and B.2 show the PESQ and the ESTOI scores, respectively, for the seen speaker case. It can be seen that all the models improve the mixtures in terms of both estimated speech quality and intelligibility at all SNRs.

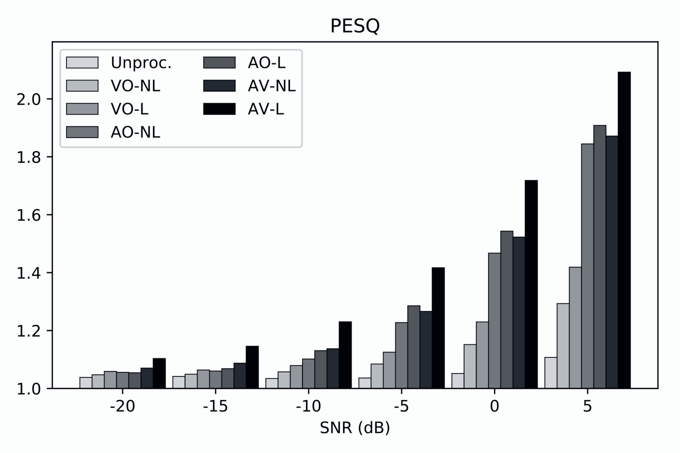


**Fig. B.1:** PESQ scores for the seen speaker case. *Unproc.* refers to the unprocessed signals.

Regarding PESQ (Fig. B.1), AV-L performs slightly better than AV-NL at -20 dB SNR, but the gap between the two models become larger when the SNR increases. The PESQ performance of the AV-L system at a particular SNR is almost as high as that of the AV-NL system at an SNR of 5 dB higher. This makes it clear that training a model with speech recorded in condition L is beneficial. The VO-L and the AO-L systems also outperform their NL counterparts, with a smaller performance gap.

The ESTOI performance (Fig. B.2) shows a similar trend, with L models outperforming the corresponding NL ones. In this case, the performance difference is substantial even at very low SNRs, where the SNR gain as defined above is slightly less than 5 dB. As expected, the contribution of vision to intelligibility is higher at low SNRs. Interestingly, the gap between VO-NL and VO-L is larger than the one between AO-NL and AO-L, suggesting that visual differences between the two speaking styles have a higher impact on intelligibility enhancement than acoustic differences.

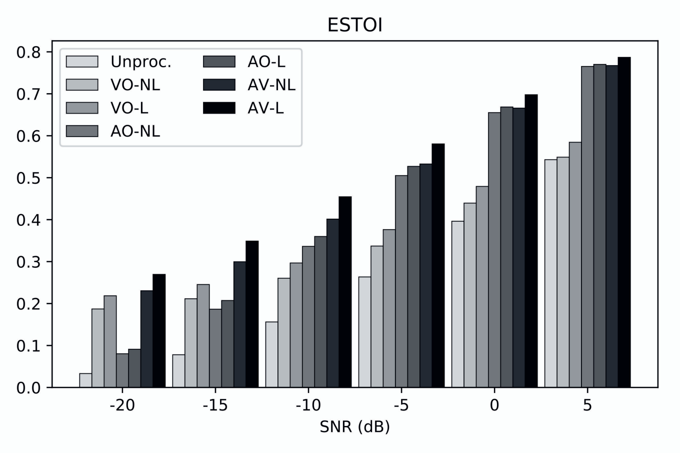


**Fig. B.2:** ESTOI scores for the seen speaker case. *Unproc.* refers to the unprocessed signals.

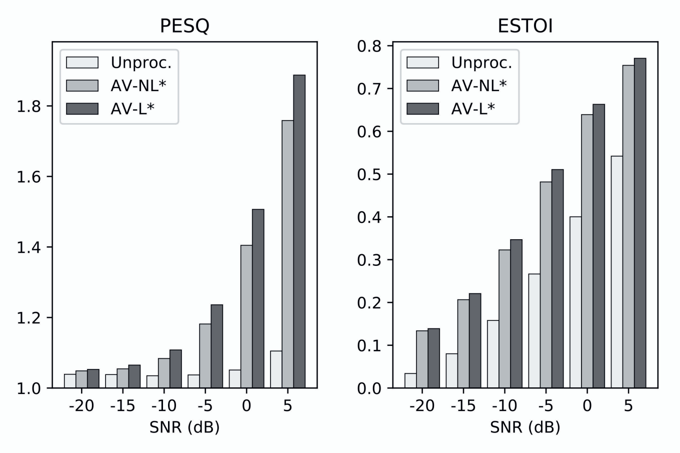


**Fig. B.3:** PESQ and ESTOI scores for the unseen speaker case. *Unproc.* refers to the unprocessed signals.

The results for the unseen speaker case are shown in Fig. B.3. The performance of AV-L* is always better than AV-NL* at all the SNRs in terms of both PESQ and ESTOI. As observed in the seen speaker case, the major PESQ improvements are reported at high SNRs, where the performance gap between AV-NL* and AV-L* is substantial. Regarding ESTOI, the difference between AV-NL* and AV-L* is smaller than the one observed in the seen speaker case. This can be explained by potential difficulties in modelling the inter-speaker variations of the Lombard speaking style. However, the performance gap between the two models is still evident, especially between -10 dB and 0 dB

SNRs, indicating the benefit of using Lombard speech for training.

# 6 Conclusion

This paper investigated the impact of Lombard effect on audio-visual speech enhancement. For this purpose, the Lombard GRID database containing the recordings of 54 different speakers in both Lombard and non-Lombard conditions has been used to train and test deep-learning-based speech enhancement systems. From the results of the experiments, the following conclusions can be drawn:

1. A network trained on neutral speech is able to improve noisy Lombard speech in terms of both estimated speech quality and intelligibility.
2. When the models are evaluated on seen speakers, the gap in the performance between the systems trained on neutral speech and the ones trained on Lombard speech indicates a benefit of as much as 5 dB if Lombard speech is used during training.
3. When the models are evaluated on unseen speakers, the performance difference between the systems trained on neutral speech and the systems trained on Lombard speech is smaller than the one observed in the seen speakers scenario, but it still suggests the advantage of training speech enhancement systems with Lombard speech.

This study showed that the Lombard effect has an impact on the performance of audio-visual speech enhancement systems and that the mismatch between neutral and Lombard speech should be taken into account in the design of these systems. Future works include listening tests to confirm the findings obtained with objective measures of speech quality and speech intelligibility.

# References

[1] H. Brumm and S. A. Zollinger, "The evolution of the Lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, no. 11-13, pp. 1173–1198, 2011.

[2] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[3] A. L. Pittman and T. L. Wiley, "Recognition of speech produced in noise," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 3, pp. 487–496, 2001.

[4] M. Garnier, L. Ménard, and B. Alexandre, "Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues?" *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 1059–1074, 2018.

[5] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[6] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[7] N. P. Erber, "Auditory-visual perception of speech," *Journal of Speech and Hearing Disorders*, vol. 40, no. 4, 1975.

[8] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, 1954.

[9] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[10] I. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[11] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[12] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. of Interspeech*, 2018.

[13] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[14] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *Proc. of Interspeech*, 2018.

[15] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, 2009.

[16] P. Heracleous, C. T. Ishi, M. Sato, H. Ishiguro, and N. Hagita, "Analysis of the visual Lombard effect and automatic recognition experiments," *Computer Speech & Language*, vol. 27, no. 1, pp. 288–300, 2013.

[17] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, "The impact of the Lombard effect on audio and visual speech recognition systems," *Speech Communication*, vol. 100, pp. 58–68, 2018.

[18] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[20] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech, Language, and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[21] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[22] ——, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1495–1499, 2009.

[23] M. Garnier and N. Henrich, "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?" *Computer Speech & Language*, vol. 28, no. 2, pp. 580–597, 2014.

[24] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. of SLT*, 2016.

[25] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[26] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "On training targets and objective functions for deep-learning-based audio-visual speech enhancement," *ICASSP (Accepted)*, 2019.

[27] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[28] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[29] L. J. Raphael, G. J. Borden, and K. S. Harris, *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Lippincott Williams & Wilkins, 2007.

[30] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP*, 2001.

[32] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

# Paper C

## Deep-Learning-Based Audio-Visual Speech Enhancement in Presence of Lombard Effect

Daniel Michelsanti, Zheng-Hua Tan, Sigurdur Sigurdsson,
Jesper Jensen

The paper has been published in
*Speech Communication*, vol. 115, pp. 38–50, 2019.

https://doi.org/10.1016/j.specom.2019.10.006

*The layout has been revised.*

# Abstract

*When speaking in presence of background noise, humans reflexively change their way of speaking in order to improve the intelligibility of their speech. This reflex is known as Lombard effect. Collecting speech in Lombard conditions is usually hard and costly. For this reason, speech enhancement systems are generally trained and evaluated on speech recorded in quiet to which noise is artificially added. Since these systems are often used in situations where Lombard speech occurs, in this work we perform an analysis of the impact that Lombard effect has on audio, visual and audio-visual speech enhancement, focusing on deep-learning-based systems, since they represent the current state of the art in the field.*

*We conduct several experiments using an audio-visual Lombard speech corpus consisting of utterances spoken by* 54 *different talkers. The results show that training deep-learning-based models with Lombard speech is beneficial in terms of both estimated speech quality and estimated speech intelligibility at low signal to noise ratios, where the visual modality can play an important role in acoustically challenging situations. We also find that a performance difference between genders exists due to the distinct Lombard speech exhibited by males and females, and we analyse it in relation with acoustic and visual features. Furthermore, listening tests conducted with audio-visual stimuli show that the speech quality of the signals processed with systems trained using Lombard speech is statistically significantly better than the one obtained using systems trained with non-Lombard speech at a signal to noise ratio of* $-5$ *dB. Regarding speech intelligibility, we find a general tendency of the benefit in training the systems with Lombard speech.*

# 1 Introduction

*Speech* is perhaps the most common way that people use to communicate with each other. Often, this kind of communication is harmed by several sources of disturbance that may have different nature, such as the presence of competing speakers, the loud music during a party, and the noise inside a car cabin. We refer to the sounds other than the speech of interest as *background noise*.

Background noise is known to affect two attributes of speech: *intelligibility* and *quality* [1]. Both of these aspects are important in a conversation, since poor intelligibility makes it hard to comprehend what a speaker is saying and poor quality may affect speech naturalness and listening effort [1]. Humans tend to tackle the negative effects of background noise by instinctively changing the way of speaking, their *speaking style*, in a process known as *Lombard effect* [2, 3]. The changes that can be observed vary widely across individuals [4, 5] and affect multiple dimensions: acoustically, the average fundamental frequency (F0) and the sound energy increase, the spectral tilt flattens due to an energy increment at high frequencies and the centre frequency of the first and second formant (F1 and F2) shifts [4, 6]; visually, head and face motion are more pronounced and the movements of the lips and jaw are amplified [7–9]; temporally, the speech rate changes due to an increase of the vowel duration [4, 10].

Although Lombard effect improves the intelligibility of speech in noise [11, 12], effective communication might still be challenged by some particular conditions, e.g. the hearing impairment of the listener. In these situations, *speech enhancement* (SE) algorithms may be applied to the noisy signal aiming at improving speech quality and speech intelligibility. In the literature, several SE techniques have been proposed. Some approaches consider SE as a *statistical estimation* problem [1], and include some well-known methods, like the Wiener filtering [13] and the minimum mean square error estimator of the short-time magnitude spectrum [14]. Many improved methods have been proposed, which primarily distinguish themselves by refined statistical speech models [15–17] or noise models [1, 18]. These techniques, which make statistical assumptions on the distributions of the signals, have been reported to be largely unable to provide speech intelligibility improvements [19, 20]. As an alternative, *data-driven techniques*, especially deep learning, make less strict assumptions on the distribution of the speech, of the noise or on the way they are mixed: a learning algorithm is used to find a function that best maps features from degraded speech to features from clean speech. Over the years, the speech processing community has put a considerable effort into designing training targets and objective functions [21–24] for different neural network models, including deep neural networks [25, 26], denoising autoencoders [27], recurrent neural networks [28], fully convolutional neural networks [29], and generative adversarial networks [30]. These methods represent the current state of the art in the field [31], and since they use only audio signals, we refer to them as audio-only SE (AO-SE) systems.

Previous studies show that observing the speaker's facial and lip movements contributes to speech perception [32–34]. This finding suggests that a SE system could tolerate higher levels of background noise, if visual cues could be used in the enhancement process. This intuition is confirmed by a pioneering study on audio-visual SE (AV-SE) by Girin et al. [35], where simple geometric features extracted from the video of the speaker's mouth are used. Later, more complex frameworks based on classical statistical approaches have been proposed [36–38], and very recently deep learning methods have been used for AV-SE [39–44].

It is reasonable to think that visual features are mostly helpful for SE when the speech is so degraded that AO-SE systems achieve poor performance, i.e. when background noise heavily dominates over the speech of interest. Since in such acoustical environment spoken communication is particularly hard, we can assume that the speakers are under the influence of Lombard effect.

In other words, the input to SE systems in this situation is Lombard speech. Despite this consideration, state-of-the-art SE systems do not take Lombard effect into account, because collecting Lombard speech is usually expensive. The training and the evaluation of the systems are usually performed with speech recorded in quiet and afterwards degraded with additive noise. Previous works show that speaker [45] and speech recognition [4] systems that ignore Lombard effect achieve sub-optimal performance, also in visual [5, 46] and audio-visual settings [46]. It is therefore of interest to conduct a similar study also in a SE context.

With the objective of providing a more extensive analysis of the impact of Lombard effect on deep-learning-based SE systems, the present work extends a preliminary study [47], providing the following novel contributions. First, new experiments are conducted, where deep-learning-based SE systems trained with Lombard or non-Lombard speech are evaluated on Lombard speech using a cross-validation setting to avoid that a potential intra-speaker variability of the adopted dataset leads to biased conclusions. Then, an investigation of the effect that the inter-speaker variability has on the systems is carried out, both in relation to acoustic as well as visual features. Next, as an example application, a system trained with both Lombard and non-Lombard data using a wide signal-to-noise-ratio (SNR) range is compared with a system trained only on non-Lombard speech, as it is currently done for the state-of-the-art models. Finally, especially since existing objective measures are limited to predict speech quality and intelligibility from the audio signals in isolation, listening tests using audio-visual stimuli have been performed. This test setup, which is generally not employed to evaluate SE systems, is closer to a real-world scenario, where a listener is usually able to look at the face of the talker.

## 2 Materials: Audio-Visual Speech Corpus and Noise Data

The speech material used in this study is the Lombard GRID corpus [48], which is an extension of the popular audio-visual GRID dataset [49]. It consists of 55 native speakers of British English (25 males and 30 females) that are between 18 and 30 years old. The sentences pronounced by the talkers adhere to the syntax from the GRID corpus, six-word sentences with the following structure: <command> <color*> <preposition> <letter*> <digit*> <adverb> (Table C.1). The words marked with a * are keywords, whereas the others are fillers [49].

Each speaker was recorded while reading a unique set of 50 sentences in non-Lombard (NL) and Lombard (L) conditions (in total, 100 utterances per speaker). In both cases, the audio signals were recorded with a microphone

**Table C.1:** Sentence structure for the Lombard GRID corpus [48]. The '*' indicates a keyword. Adapted from [49].

| Command | Colour* | Preposition | Letter* | Digit* | Adverb |
|---|---|---|---|---|---|
| bin | blue | at | A–Z (no W) | 0–9 | again |
| lay | green | by | | | now |
| place | red | in | | | please |
| set | white | with | | | soon |

placed in front of the speakers, while the video recordings were collected with two cameras mounted on a helmet to have a frontal and a profile views of the talkers.

In order to induce the Lombard effect, speech shaped noise (SSN) at 80 dB sound pressure level (SPL) was presented to the speakers, while they were reading the sentences to a listener. The presence of a listener, who assured a natural communication environment by asking the participants to repeat the utterances from time to time, was needed, because talkers usually adjust their speech to communicate better with the people they are talking to [6, 50], a process known as *external* or *public loop* [50]. Since talkers tend to regulate their speaking style also based on the level of their own speech, in what is generally called *internal* or *private loop* [50], the speech signal was mixed with the SSN at a carefully adjusted level, providing a self-monitoring feedback to the speakers.

In our study, the audio and the video signals from the frontal camera were arranged as explained in Section 4 to build training, validation, and test sets. The audio signals have a sampling rate of 16 kHz. The resolution of the frontal video stream is 720 × 480 pixels with a variable frame rate of around 24 frames per second (FPS). Audio and video signals are temporally aligned.

To generate speech in noise, SSN was added to the audio signals of the Lombard GRID database. SSN was chosen to match the kind of noise used in the database, since, as reported by Hansen and Varadarajan [45], Lombard effect occurs differently across noise types, although other studies [51, 52] failed to find such an evidence. The SSN we used was generated as in [53], by filtering white noise with a low-order linear predictor, whose coefficients were found using 100 random sentences from the Akustiske Databaser for Dansk (ADFD)[1] speech database.

# 3 Methodology

In this study, we train and evaluate systems that perform spectral SE using deep learning, as illustrated in Figure C.1. The processing pipeline is inspired by Gabbay et al. [40] and the same as the one used in [47]. To have

[1] `https://www.nb.no/sbfil/dok/nst_taledat_dk.pdf`

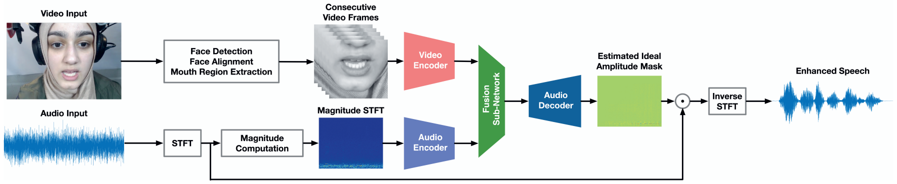


**Fig. C.1:** Pipeline of the audio-visual speech enhancement framework used in this study, adapted from [40], and identical to [47]. The deep-learning-based system estimates an ideal amplitude mask from the video of the speaker's mouth and the magnitude spectrogram of the noisy speech. The estimated mask is used to enhance the speech in time-frequency domain. *STFT* indicates the short-time Fourier transform.

a self-contained exposition, we report the main details of it in this section. We did not explore the effect of changing the network topology because we are interested in the performance gap between Lombard and non-Lombard systems, and, for this, it is essential that the systems which are compared use exactly the same architecture.

## 3.1 Audio-Visual Speech Enhancement

We assume to have access to two streams of information: the video of the talker's face, and an audio signal, $y(n) = x(n) + d(n)$, where $x(n)$ is the clean signal of interest, $d(n)$ is an additive noise signal, and $n$ indicates the discrete-time index. The additive noise model presented in time domain can also be expressed in the time-frequency (TF) domain as $Y(k,l) = X(k,l) + D(k,l)$, where $Y(k,l)$, $X(k,l)$, and $D(k,l)$ are the short-time Fourier transform (STFT) coefficients at frequency bin $k$ and at time frame $l$ of $y(n)$, $x(n)$, and $d(n)$, respectively. Our models adopt a mask approximation approach [24], producing an estimate $\widehat{M}(k,l)$ of the ideal amplitude mask, defined as $M(k,l) = |X(k,l)|/|Y(k,l)|$, with the following objective function:

$$J = \frac{1}{TF} \sum_{k,l} \left( M(k,l) - \widehat{M}(k,l) \right)^2, \tag{C.1}$$

with $k \in \{1, \ldots, F\}$, $l \in \{1, \ldots, T\}$, and $T \times F$ being the dimension of the training target. Recent preliminary experiments have shown that using this objective function leads to better performance for AV-SE than competing methods [24].

## 3.2 Preprocessing

In this work, each audio signal was peak-normalised. We used a sample rate of 16 kHz and a 640-point STFT, with a Hamming window of 640 samples (40 ms) and a hop size of 160 samples (10 ms). Only the 321 bins that cover the positive frequencies were used, because of the conjugate symmetry of the STFT.

Each video signal was resampled at a frame rate of 25 FPS using motion interpolation as implemented in FFMPEG[2]. The face of the speaker was detected in every frame using the frontal face detector implemented in the dlib toolkit [54], consisting of 5 histogram of oriented gradients (HOG) filters and a linear support vector machine (SVM). The bounding box of the single-frame detections was tracked using a Kalman filter. The face was aligned based on 5 landmarks using a model that estimated the position of the corners of the eyes and of the bottom of the nose [54] and was scaled to $256 \times 256$ pixels. The mouth was extracted by cropping the central lower face region of size $128 \times 128$ pixels.

Each segment of 5 consecutive grayscale video frames spanning a total of 200 ms was paired with the respective 20 consecutive audio frames.

### 3.3 Neural Network Architecture and Training

The preprocessed audio and video signals, standardised using the mean and the variance from the training set, were used as inputs to a video and an audio encoders, respectively. Both encoders consisted of 6 convolutional layers (Table C.2), each of them followed by leaky-ReLU activation functions [55] and batch normalisation [56]. For the video encoder, also max-pooling and 0.25 dropout [57] were adopted. The fusion of the two modalities was accomplished using a sub-network consisting of 2 fully connected layers with 1312 neurons each and another one with 3840 neurons, followed by leaky-ReLU activations. The $321 \times 20$ estimated mask was obtained with an audio decoder having 6 transposed convolutional layers, mirroring the convolutional layers of the audio encoder, followed by leaky-ReLU activations and a ReLU activation as output layer. Skip connections between the layers 1, 3, and 5 of the audio encoder and the corresponding decoder layers were used to avoid that the bottleneck hindered the information flow [58]. Following the approach in [21], we limited the values of the training target, $M(k,l)$, in the $[0, 10]$ interval.

The weights of the network were initialised with the Xavier approach [59]. The training was performed using the Adam optimiser [60] with the objective function in Equation (C.1) and a batch size of 64. The learning rate, initially set to $4 \cdot 10^{-4}$, was scaled by a factor of 0.5 when the loss increased on the validation set. An early stopping technique was used, by selecting the network that performed the best on the validation set across the 50 epochs used for training.

[2] http://ffmpeg.org

**Table C.2:** Convolutional layers of the audio and video encoders. Adapted from [47].

| | Video Encoder | | | Audio Encoder | | |
|---|---|---|---|---|---|---|
| Layer | Filters | Kernel | Stride | Filters | Kernel | Stride |
| 1 | 128 | 5×5 | 1×1 | 64 | 5×5 | 2×2 |
| 2 | 128 | 5×5 | 1×1 | 64 | 4×4 | 2×1 |
| 3 | 256 | 3×3 | 1×1 | 128 | 4×4 | 2×2 |
| 4 | 256 | 3×3 | 1×1 | 128 | 2×2 | 2×1 |
| 5 | 512 | 3×3 | 1×1 | 128 | 2×2 | 2×1 |
| 6 | 512 | 3×3 | 1×1 | 128 | 2×2 | 2×1 |

## 3.4 Postprocessing

The estimated ideal amplitude mask of an utterance was obtained by concatenating the network's outputs from the processed non-overlapping consecutive audio-visual paired segments. The estimated mask was point-wise multiplied with the magnitude STFT spectrogram of the noisy signal, the noisy STFT phase was appended, and the result was inverted to get the time-domain signal with an overlap-add procedure [61, 62].

## 3.5 Mono-Modal Speech Enhancement

Until now, we only presented AV-SE systems. In order to understand the relative contribution of the audio and the visual modalities, we also trained networks to perform mono-modal SE, by removing one of the two encoders from the neural network architecture, without changing the other explained settings and procedures. Both AO-SE and video-only SE (VO-SE) systems estimate a mask and apply it to the noisy speech, but they differ in the signals used as input.

# 4 Experiments

The experiments conducted in this study compare the performance of AO-SE, VO-SE, and AV-SE systems in terms of two widely adopted objective measures: perceptual evaluation of speech quality (PESQ) [63], specifically the wideband extension [64] as implemented by Loizou [1], and extended short-time objective intelligibility (ESTOI) [65]. PESQ scores, used to estimate speech quality, lie between $-0.5$ and 4.5, where high values correspond to high speech quality. However, the wideband extension that we use maps these scores to mean opinion score (MOS) values, on a scale from approximately 1 to 4.64. ESTOI scores, which estimate speech intelligibility, practically range from 0 to 1, where high values correspond to high speech intelligibility.

As mentioned before (Section 2), clean speech signals were mixed with

**Table C.3:** Models used in this study. The '$^{(w)}$' is used to distinguish the systems trained with a wide SNR range from the ones trained with a narrow SNR range.

| | Training Material | | | |
|---|---|---|---|---|
| | Non-Lombard Speech | | Lombard Speech | |
| System Input | Narrow SNR Range | Wide SNR Range | Narrow SNR Range | Wide SNR Range |
| Vision | VO-NL | VO-NL$^{(w)}$ | VO-L | VO-L$^{(w)}$ |
| Audio | AO-NL | AO-NL$^{(w)}$ | AO-L | AO-L$^{(w)}$ |
| Audio-Visual | AV-NL | AV-NL$^{(w)}$ | AV-L | AV-L$^{(w)}$ |

SSN to match the noise type used in the Lombard GRID corpus. Current state-of-the-art SE systems are trained with signals at several SNRs to make them robust to various noise levels. We followed a similar methodology and trained our models with two different SNR ranges, narrow (between $-20$ dB and 5 dB) and wide (between $-20$ dB and 30 dB). We used these two ranges because on the one hand we would like to assess the performance of SE systems when Lombard speech occurs, and on the other hand we would like to have SNR-independent systems, i.e. systems that also work well at higher SNRs. Such a setup allows us to better understand whether Lombard speech, which is usually not available because it is hard to collect, should be used to train SE systems and which are the advantages and the disadvantages of various training configurations. The models used in this work are shown in Table C.3.

Similarly to the work by by Marxer et al. [5], the experiments were conducted adopting a multi-speaker setup, where some utterances of several speakers were used for training and other utterances of the same speakers (mixed with different realisations of the noise) were used for testing. A multi-speaker setup was preferred to single-speaker training and speaker-independent training for the following reasons:

- People may exhibit speech characteristics that differ considerably from each other when they speak in presence of noise [4, 5]. It is possible to model these differences by training speaker-dependent systems, but this requires a large set of Lombard speech for every speaker. Unfortunately, the audio-visual speech corpus that we use, despite being one of the largest existing audio-visual databases for Lombard speech, only contains 50 utterances per speaker, which are not enough to train a deep-learning-based model.

- Our systems were evaluated using speakers observed at training time, because we are interested in studying the impact of Lombard effect in isolation from other factors that might pollute the results, such as test conditions that could be very different from the training data due to the limited amount of speakers in the dataset.

The experiments were performed according to a stratified five-fold cross-validation procedure [66]. Specifically, the data was divided into five folds of approximately the same size, four of them used for training and validation, and one for testing. This process was repeated five times for different test sets in order to evaluate the systems on the whole dataset. Before the split, the signals were rearranged to have about the same amount of data for each speaker across the training ($\sim$ 35 utterances), the validation ($\sim$ 5 utterances), and the test ($\sim$ 10 utterances) sets. This ensured that each fold was a good representative of the inter-speaker variations of the whole dataset. For some speakers, some data was missing or corrupted, so we used fewer utterances. Among the 55 speakers, the recordings from speaker s1 were discarded by the database collectors due to technical issues, and the data from speaker s51 was used only in the training set, because only 40 of the utterances could be used. Effectively, 53 speakers were used to evaluate our systems.

## 4.1 Systems Trained on a Narrow SNR Range

Since we would like to assess the performance of SE systems when Lombard speech occurs, SSN was added to the speech signals from the Lombard GRID corpus at 6 different SNRs, in uniform steps between $-20$ dB and 5 dB. This choice was driven by the following considerations [47]. Since Lombard and non-Lombard utterances from the Lombard GRID corpus have an energy difference between 3 and 13 dB [5], the actual SNR can be computed assuming that the conversational speech level is between 60 and 70 dB sound pressure level (SPL) [67, 68] and the noise level at 80 dB SPL, like in the recording conditions of the database. The SNR range obtained in this way is between $-17$ and 3 dB. In the experiments, we used a slightly wider range because of the possible speech level variations caused by the distance between the listener and the speaker.

For all the systems, Lombard speech was used to build the test set, while for training and validation we used Lombard speech for VO-L, AO-L, and AV-L, and non-Lombard speech for VO-NL, AO-NL, and AV-NL (Table C.3).

### 4.1.1 Results and Discussion

Figure C.2 shows the cross-validation results in terms of PESQ and ESTOI for all the different systems. On average, every model improves the estimated speech quality and the estimated speech intelligibility of the unprocessed signals, with the exception of VO-NL at 5 dB SNR, which shows an ESTOI score comparable with the one of noisy speech. Another general trend that can be observed is that AV systems outperform the respective AO and VO systems, an expected result since the information that can be exploited using

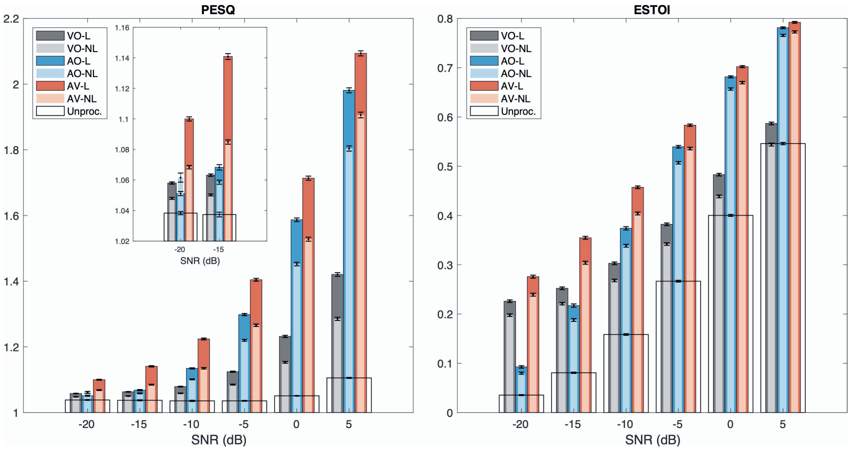


**Fig. C.2:** Cross-validation results in terms of PESQ and ESTOI for the systems trained on a narrow SNR range. At every SNR, there are three pairs of coloured bars with error bars, each of them referring to VO, AO, and AV systems (from left to right). The wide bars in dark colours represent L systems, while the narrow ones in light colours represent NL systems. The heights of each bar and the error bars indicate the average scores and the 95% confidence intervals computed on the pooled data, respectively. The transparent boxes with black edges, overlaying the bars of the other systems, and the error bars indicate the average scores of the unprocessed signals (*Unproc.*) and their 95% confidence intervals, respectively.

two modalities is no less than the information of the single modalities taken separately.

It is worth noting that VO systems' performance changes across SNR, although they do not use the audio signal to estimate the ideal amplitude mask. This is because the estimated mask is applied to the noisy input signal, so the performance depends on the noise level of the input audio signal.

PESQ scores show that the performance that can be obtained with AO systems is comparable with VO systems performance at very low SNRs. Only for SNR $\geq -10$ dB, AO models start to perform substantially better than VO models. This difference increases with higher SNRs, where gains of even more than 5 dB in SNR can be observed. Also for ESTOI, this pattern can be observed when SNR $\geq -10$ dB, but for SNR $\leq -15$ dB VO systems perform better than the respective AO systems, especially at $-20$ dB SNR, where the gain is approximately 5 dB in SNR. This can be explained by the fact that the noise level is so high that recovering the clean speech only using the noisy audio input is very challenging, and that the visual modality provides a richer information source at this noise level.

For all the modalities, L systems tend to be better than the respective NL systems. The only exception is AO-NL, which have a higher PESQ score than AO-L at $-20$ dB SNR, but this difference is very modest (0.011). AV-L always outperforms AV-NL in terms of PESQ by a large margin, with more than 5

**Table C.4:** Average scores for the systems trained on a narrow SNR range.

| PESQ | VO-L | VO-NL | AO-L | AO-NL | AV-L | AV-NL |
|---|---|---|---|---|---|---|
| −20 - 5 dB | 1.163 | 1.113 | 1.353 | 1.283 | 1.446 | 1.331 |
| ESTOI | VO-L | VO-NL | AO-L | AO-NL | AV-L | AV-NL |
| −20 - 5 dB | 0.372 | 0.335 | 0.448 | 0.423 | 0.528 | 0.488 |

dB SNR gain, if we consider the performance between $-20$ dB and $-10$ dB SNR. For higher SNRs, the gain is about 2.5 dB in SNR. On average (Table C.4), the performance gap in terms of PESQ between L and NL systems, is greater for the audio-visual case (0.115) than for the audio-only (0.070) and the video-only (0.050) cases, meaning that the speaking style mismatch is more detrimental when both the modalities are used. Regarding ESTOI, the gap between AV-L and AV-NL (0.040) is still the largest, but the one between VO-L and VO-NL (0.037) is greater than the gap between AO-L and AO-NL (0.025): this suggests that the impact of visual differences between Lombard and non-Lombard speech on estimated speech intelligibility is higher than the impact of acoustic differences. In general, the gain of training AV systems with Lombard speech is equivalent to an ESTOI increase over AV systems trained on non-Lombard speech between 1.5 dB and 2.8 dB in SNR.

These results suggest that training systems with Lombard speech is beneficial in terms of both estimated speech quality and estimated speech intelligibility. This is in line with and extends our preliminary study [47], where only a subset of the whole database was used to evaluate the models.

### 4.1.2 Effects of Inter-Speaker Variability

Previous work found a large inter-speaker variability for Lombard speech, especially between male and female speakers [4]. Here, we investigate whether this variability affects the performance of SE systems.

Figure C.3 shows the average PESQ and ESTOI scores by gender. Since the scores are computed on different speech material, it may be hard to make a direct comparison between males and females by looking at the absolute performance. Instead, we focus on the gap between L and NL systems averaged across SNRs for same gender. At a first glance, the trends of the different conditions are as expected: L systems are better than the respective NL ones, and AV systems outperform AO systems trained with speech of the same speaking style, in terms of both estimated speech quality and estimated speech intelligibility. We also notice that the scores of VO systems are worse than the AO ones, also for ESTOI. This is because we average across all the SNRs and VO is better than AO only at very low SNRs, but considerably worse for SNR $\geq -5$ dB (Figure C.2).

The difference between L and NL systems is larger for females than it is

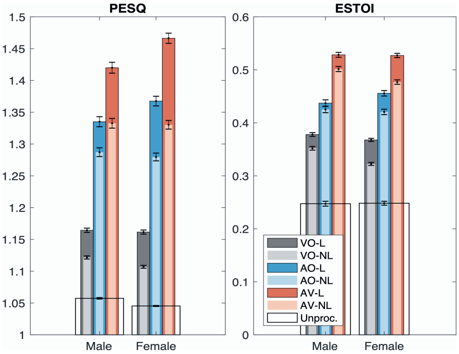


**Fig. C.3:** Cross-validation results for male and female speakers in terms of PESQ and ESTOI.

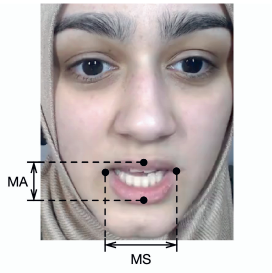


**Fig. C.4:** Mouth aperture (MA) and mouth spreading (MS) from the 4 facial landmarks (black dots) estimated with the algorithm [73] implemented in the dlib toolkit [54].

for males. This can be observed for all the modalities and it is more noticeable for AV systems, most likely because they account for both audio and visual differences. In order to better understand this behaviour, we provide a more in-depth analysis, investigating the impact that some acoustic and geometric articulatory features have on estimated speech quality and estimated speech intelligibility.

We consider three different features that have already been used to study Lombard speech in previous work [9, 69–71]: F0, mouth aperture (MA) and mouth spreading (MS). The average F0 for each speaker was estimated with Praat [72], using the default settings for pitch estimation. The average MA and MS per speaker were computed from 4 facial landmarks (Figure C.4) obtained with the pose estimation algorithm [73], trained on the iBUG 300-W database [74], implemented in the dlib toolkit [54]. Since all the videos from the database show frontal faces, with no drastic changing in pose and illumination, a good landmarks' estimation can be obtained with this algorithm (see Figure C.4 for an example).

Let $\Delta$F0, $\Delta$MA, and $\Delta$MS denote the average difference in audio and visual features, respectively, between Lombard and non-Lombard speech. Similarly,

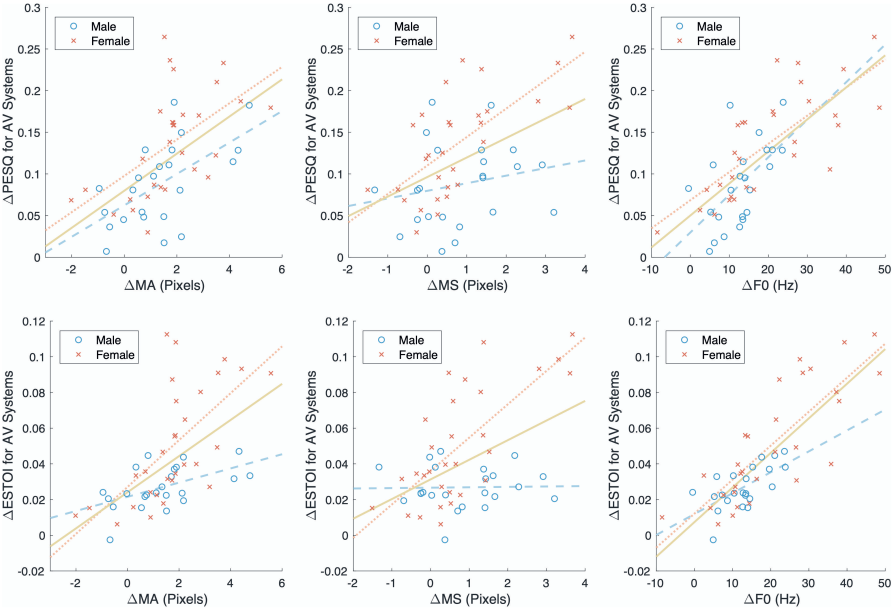


**Fig. C.5:** Scatter plots showing the relationship between the audio/visual features and PESQ/ESTOI. For each blue circle (for males) and red cross (for females), which refer to a particular speaker, the $y$-coordinate indicates the average performance increment of AV-L with respect to AV-NL in terms of PESQ or ESTOI, while the $x$-coordinate indicates the average increment of audio (fundamental frequency) or visual (mouth aperture and mouth spreading) features in Lombard condition with respect to the respective feature in non-Lombard condition. The lines show the least-squares fitted lines for male speakers (dashed blue), female speakers (dotted red), and all the speakers (solid yellow). MA, MS, and F0 indicate mouth aperture, mouth spreading, and fundamental frequency, respectively.

let ΔPESQ and ΔESTOI denote the increment in PESQ and ESTOI, respectively, of AV-L with respect to AV-NL. Figure C.5 illustrates the relationship between ΔF0/ΔMA/ΔMS and ΔPESQ/ΔESTOI. We notice that, on average, for each speaker ΔPESQ and ΔESTOI are both positive, with only one exception represented by a male speaker, whose ΔESTOI is slightly less than 0. This indicates that no matter how different the speaking style of a person is in presence of noise, there is a benefit in training a system with Lombard speech. Focusing on the range of the features' variations, most of the speakers have positive ΔMA, ΔMS, and ΔF0. This is in accordance with previous research, which suggests that in Lombard condition there is a tendency to amplify lips' movements and rise the pitch [4, 8, 9]. ΔMA and ΔMS values lie between $-2$ and 6 pixels, and between $-2$ and 4 pixels, respectively, for both male and female speakers. Regarding the ΔF0 range, it is wider for females, up to 50 Hz, against the 25 Hz reached by males.

Among the three features considered, ΔF0 is the one that seems to be related the most with ΔPESQ and ΔESTOI. This can be seen by comparing the distributions of the circles with the least-squares lines in the plots of Figure C.5 or by analysing the correlation between PESQ/ESTOI increments and audio/visual feature increments, using Pearson's and Spearman's correlation coefficients.

Given $n$ pairs of $(x_i, y_i)$ observations, with $i \in \{1, \dots, n\}$, from two variables $x$ and $y$, whose sample means are denoted as $\bar{x}$ and $\bar{y}$, respectively, we refer to the Pearson's correlation coefficient as $\rho_P(x, y)$. We have that $-1 \leq \rho_P(x, y) \leq 1$, where 0 denotes the absence of a linear relationship between the two variables, and $-1$ and 1 a perfect positive linear relationship and a perfect negative linear relationship, respectively. To complement the Pearson's correlation coefficient, we also consider the Spearman's correlation coefficient, $\rho_S(x, y)$, defined as [75]:

$$\rho_S(x, y) = \rho_P(r_x, r_y), \tag{C.2}$$

where $r_x$ and $r_y$ indicate rank variables. The advantage of using ranks is that $\rho_S$ allows to assess whether the relationship between $x$ and $y$ is monotonic (not limited to linear).

As shown in Table C.5, for AV systems, ΔF0 has a higher correlation with ΔPESQ ($\rho_P = 0.73$, $\rho_S = 0.73$) and ΔESTOI ($\rho_P = 0.81$, $\rho_S = 0.77$) than ΔMA and ΔMS. We observe that for female speakers, the correlation between the features' increments and the performance measures' increments is usually higher, especially when considering ΔMS, suggesting that some inter-gender difference should be present not only for ΔF0 (whose range is way wider for females as previously stated), but also for visual features.

In Table C.5 we also report the correlation coefficients for the single modalities. The correlation of visual features' increments with ΔPESQ or ΔESTOI is sometimes higher for AO systems than it is for VO systems. This might seem counter-intuitive, because AO systems do not use visual information. However, correlation does not imply causation [76]: since visual and acoustic features are correlated [77], it is possible that other acoustic features, which are not considered in this study even though they might be correlated with ΔMA and ΔMS, play a role in the enhancement. Similar considerations can be done for ΔF0, which has a correlation with ΔESTOI for VO systems ($\rho_P = 0.77$, $\rho_S = 0.77$) higher than the one for AO systems ($\rho_P = 0.64$, $\rho_S = 0.60$). By looking at the inter-gender differences, we find that, in general, the correlation coefficients computed for female speakers are higher than the ones computed for male speakers, especially when considering ΔMS.

In general, a performance difference between genders exists when L systems are compared with NL ones, with a gap that is larger for females. This is unlikely to be caused by the small gender imbalance in the training set (23 males and 30 females). Instead, it is reasonable to assume that this result is due to the characteristics of the Lombard speech of female speakers, which

**Table C.5:** Pearson's ($\rho_P$) and Spearman's ($\rho_S$) correlation coefficients between PESQ/ESTOI increments and audio/visual feature increments for male speakers (m), female speakers (f), and all the speakers. MA, MS, and F0 indicate mouth aperture, mouth spreading, and fundamental frequency, respectively.

| | | $\rho_P$ | | | $\rho_S$ | |
|---|---|---|---|---|---|---|
| | all | m | f | all | m | f |
| ΔPESQ (VO) - ΔMA | .29 | .32 | .24 | .35 | .30 | .29 |
| ΔPESQ (AO) - ΔMA | .43 | .49 | .40 | .55 | .49 | .51 |
| ΔPESQ (AV) - ΔMA | .57 | .59 | .56 | .65 | .52 | .66 |
| ΔESTOI (VO) - ΔMA | .46 | .19 | .57 | .52 | .16 | .69 |
| ΔESTOI (AO) - ΔMA | .43 | .47 | .46 | .52 | .52 | .50 |
| ΔESTOI (AV) - ΔMA | .57 | .53 | .65 | .67 | .47 | .72 |
| ΔPESQ (VO) - ΔMS | .19 | −.08 | .35 | .12 | −.03 | .31 |
| ΔPESQ (AO) - ΔMS | .31 | .20 | .45 | .33 | .19 | .54 |
| ΔPESQ (AV) - ΔMS | .45 | .21 | .68 | .44 | .28 | .71 |
| ΔESTOI (VO) - ΔMS | .45 | −.12 | .73 | .22 | −.21 | .62 |
| ΔESTOI (AO) - ΔMS | .30 | .05 | .47 | .22 | .07 | .48 |
| ΔESTOI (AV) - ΔMS | .47 | .02 | .72 | .34 | −.02 | .66 |
| ΔPESQ (VO) - ΔF0 | .34 | .26 | .31 | .36 | .23 | .35 |
| ΔPESQ (AO) - ΔF0 | .62 | .53 | .58 | .61 | .52 | .61 |
| ΔPESQ (AV) - ΔF0 | .73 | .58 | .75 | .73 | .59 | .80 |
| ΔESTOI (VO) - ΔF0 | .77 | .57 | .77 | .77 | .58 | .82 |
| ΔESTOI (AO) - ΔF0 | .64 | .55 | .60 | .60 | .56 | .61 |
| ΔESTOI (AV) - ΔF0 | .81 | .64 | .81 | .77 | .61 | .84 |

shows a large increment of F0, the feature that correlates the most with the estimated speech quality and the estimated speech intelligibility increases, among the ones considered.

## 4.2 Systems Trained on a Wide SNR Range

The models presented in Section 4.1 have been trained to enhance signals when Lombard effect occurs, i.e. at SNRs between $-20$ and 5 dB. However, from a practical perspective, SNR-independent systems, capable of enhancing both Lombard and non-Lombard speech, are preferred. There are several ways to achieve this goal. For example, it is possible to train a system (with Lombard speech) that works at low SNRs, and another one (with non-Lombard speech) that works at high SNRs. This approach requires switching between the two systems, which can be problematic, because it involves an online estimation of the SNR. An alternative approach is to train general systems with Lombard speech at low SNRs and non-Lombard speech at high SNRs. We followed this alternative approach, building such systems and studying their strengths and limitations. We also compared them with systems trained only with non-Lombard speech for the whole SNR range, because this is what current state-of-the-art systems do.

The test set was built by mixing additive SSN with Lombard speech at 6 SNRs between $-20$ and 5 dB, and with non-Lombard speech at 5 SNRs between 10 and 30 dB. For VO-NL$^{(w)}$, AO-NL$^{(w)}$, and AV-NL$^{(w)}$, only non-

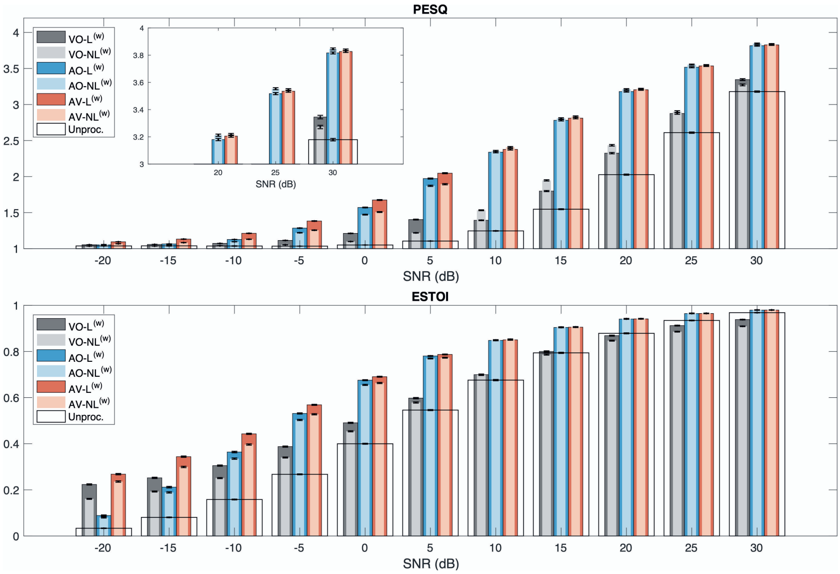


**Fig. C.6:** As Figure C.2, but for the systems trained on a wide SNR range.

Lombard speech was used during training, while for VO-L$^{(w)}$, AO-L$^{(w)}$, and AV-L$^{(w)}$, Lombard speech was used with SNR $\leq$ 5 dB and non-Lombard speech with SNR $\geq$ 10 dB, to match the speaking style of the test set (Table C.3). The results in terms of PESQ and ESTOI are shown in Figure C.6.

The relative performance of the systems at SNR $\leq$ 5 dB is similar to the one observed for the systems trained on a narrow SNR range (Section 4.1): L$^{(w)}$ systems outperform the respective NL$^{(w)}$ systems, AV performance is higher than AO and VO performance, and VO is considerably better than AO only in terms of ESTOI at very low SNRs.

When SNR $\geq$ 10 dB, NL$^{(w)}$ systems perform better than L$^{(w)}$ systems in terms of PESQ. The difference is on average (Table C.6) larger for VO (0.070) than it is for AO (0.028) and AV (0.018). This can be explained by the fact that it is harder for VO-L$^{(w)}$ to recognise when non-Lombard speech occurs using only the video of the speaker. However, these performance gaps are smaller than the ones between L$^{(w)}$ and NL$^{(w)}$ systems at SNR $\leq$ 5 dB (0.073 for VO, 0.051 for AO, and 0.101 for AV).

Regarding ESTOI at SNR $\geq$ 10 dB, the difference between AO and AV becomes negligible, with VO systems that perform considerably worse. This is because audio features are more informative than visual ones at high SNRs, making AO-SE systems already good to recover speech intelligibility. In addition, the average gaps between NL$^{(w)}$ and L$^{(w)}$ are quite small: 0.002 for AO and AV, while for VO it is actually $-0.019$.

In general, at SNR $\leq$ 5 dB, the systems that use both Lombard and non-

**Table C.6:** Average scores for the systems trained on a wide SNR range.

| PESQ | VO-L$^{(w)}$ | VO-NL$^{(w)}$ | AO-L$^{(w)}$ | AO-NL$^{(w)}$ | AV-L$^{(w)}$ | AV-NL$^{(w)}$ |
|---|---|---|---|---|---|---|
| −20 - 5 dB | 1.153 | 1.080 | 1.346 | 1.295 | 1.424 | 1.323 |
| 10 - 30 dB | 2.348 | 2.418 | 3.127 | 3.155 | 3.151 | 3.169 |
| ESTOI | VO-L$^{(w)}$ | VO-NL$^{(w)}$ | AO-L$^{(w)}$ | AO-NL$^{(w)}$ | AV-L$^{(w)}$ | AV-NL$^{(w)}$ |
| −20 - 5 dB | 0.376 | 0.330 | 0.442 | 0.422 | 0.517 | 0.483 |
| 10 - 30 dB | 0.844 | 0.825 | 0.927 | 0.929 | 0.928 | 0.930 |

Lombard speech for training perform better than the ones that only use non-Lombard speech. At higher SNRs, their PESQ and ESTOI scores are slightly worse than the ones of the systems trained only with non-Lombard speech. However, this performance gap is small, and seems to be larger for the estimated speech quality than for the estimated speech intelligibility. The way we combined non-Lombard and Lombard speech for training seems to be the best solution for an SNR-independent system, although a small performance loss may occur at high SNRs.

# 5 Listening Tests

Although it has been shown that visual cues have an impact on speech perception [32, 34], the currently available objective measures used to estimate speech quality and speech intelligibility, e.g. PESQ and ESTOI, only take into account the audio signals. Even when listening tests are performed to evaluate the performance of a SE system, visual stimuli are usually ignored and not presented to the participants [78], despite the fact that visual inputs are typically available during practical deployment of SE systems.

For these reasons, and in an attempt to evaluate the proposed enhancement systems in a setting as realistic as possible, we performed two listening tests, one to assess the speech quality and the other to assess the speech intelligibility, where all the processed and the unprocessed audio signals from the Lombard GRID corpus were accompanied by their corresponding visual stimuli. Both tests were conducted in a silent room, where a MacBookPro11,4 equipped with an external monitor, a sound card (Focusrite Scarlett 2i2) and a set of closed headphones (Beyerdynamic DT770) was used for audio and video playback. The multimedia player (VLC media player 3.0.4) was controlled by the subjects with a graphical user interface (GUI) modified from MUSHRAM [79]. The processed signals used in this test were from the systems trained on the narrow SNR range previously described (Section 4.1). All the audio stimuli were normalised according to the two-pass EBU R128 loudness normalisation procedure [80], as implemented in ffmpeg-normalize[3], to guarantee that signals of different conditions were perceived as having the

[3] `https://github.com/slhck/ffmpeg-normalize`

same volume. The subjects were allowed to adjust the general loudness to a comfortable level during the training session of each test.

## 5.1 Speech Quality Test

The quality test was carried out by 13 experienced listeners, who volunteered to be part of the study. The participants were between 26 and 44 years old, and had self-reported normal hearing and normal (or corrected to normal) vision. On average, each participant spent approximately 30 minutes to complete the test.

### 5.1.1 Procedure

The test used the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [81] paradigm to assess the speech quality on a scale from 0 to 100, divided into 5 equal intervals labelled as *bad*, *poor*, *fair*, *good*, and *excellent*. No definition of *speech quality* was provided to the participants. Each subject was presented with 2 sequences of 8 trials each, 4 to evaluate the systems at $-5$ dB SNR, and 4 to evaluate the systems at 5 dB SNR. Lower SNRs were not considered to ensure that the perceptual quality assessment was not influenced too much by the decrease in intelligibility. One trial consisted of one reference (clean speech signal) and seven other signals to be rated with respect to the reference: 1 hidden reference, 4 systems under test (AO-L, AO-NL, AV-L, AV-NL), 1 unprocessed signal, and 1 hidden anchor (unprocessed signal at $-10$ dB SNR). The participants were allowed to switch at will between any of the signals inside the same trial. The order of presentation of both the trials and the conditions was randomised, and signals from 4 different randomly chosen speakers were used for each sequence of trials.

Before the actual test, the participants were trained in a special separate session, with the purpose of exposing them to the nature of the impairments and making them familiar with the equipment and the grading system.

### 5.1.2 Results and Discussion

The average scores assigned by the subjects for each condition are shown in Figure C.7 in the form of box plots.

Non-parametric approaches are used to analyse the data [82, 83], since the assumption of normal distribution of the data is invalid, given the number of participants and their different interpretation of the MUSHRA scale. Specifically, the paired two-sided Wilcoxon signed-rank test [84] is adopted to determine whether there exists a median difference between the MUSHRA scores obtained for two different conditions. Differences in median are considered significant for $p < \alpha/m = 0.0083$ ($\alpha = 0.05$, $m = 6$), where the significance

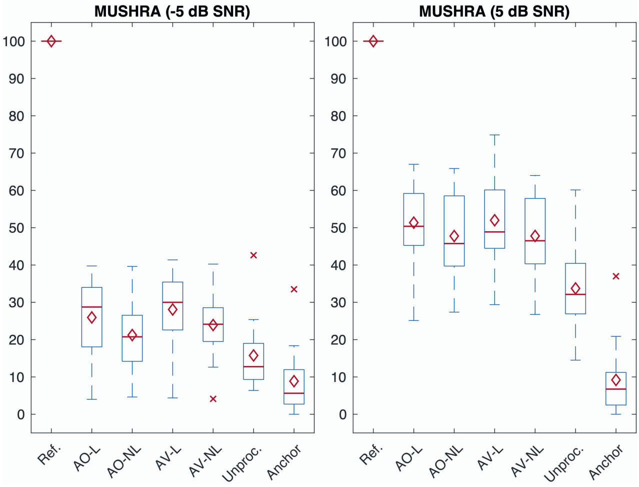


**Fig. C.7:** Box plots showing the results of the MUSHRA experiments for the signals at $-5$ dB SNR (left) and at 5 dB SNR (right). The red horizontal lines and the diamond markers indicate the median and the mean values, respectively. Outliers (identified according to the 1.5 interquartile range rule) are displayed as red crosses. *Ref.* indicates the reference signals.

**Table C.7:** Interpretation of the effect size (Cliff's delta, $d_C$). Adapted from [88].

| Small Effect Size | Medium Effect Size | Large Effect Size |
|---|---|---|
| $0.11 \leq \lvert d_C \rvert < 0.28$ | $0.28 \leq \lvert d_C \rvert < 0.43$ | $0.43 \leq \lvert d_C \rvert \leq 1$ |

level is corrected with the Bonferroni method to compensate for multiple hypotheses tests [76]. The use of $p$-values as the only analysis strategy has been heavily criticised [85] because statistical significance can be obtained with a big sample size [68, 86] even if the magnitude of the effect is negligible [85]. For this reason, we complement $p$-values with a non-parametric measure of the effect size, the Cliff's delta [87]:

$$d_C = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}[x_i > y_j] - \sum_{i=1}^{m}\sum_{j=1}^{n}[x_i < y_j]}{mn}, \tag{C.3}$$

where $x_i$ and $y_j$ are the observations of the samples of sizes $m$ and $n$ to be compared and $[P]$ indicates the Iverson bracket, which is 1 if $P$ is true and 0 otherwise. As reported in Table C.7, we consider the effect size to be small if $0.11 \leq |d_C| < 0.28$, medium if $0.28 \leq |d_C| < 0.43$, and large if $|d_C| \geq 0.43$, according to the indication by Vargha and Delaney [88]. The $p$-values and the effect sizes for the comparisons considered in this study are shown in Table C.8.

At SNR $= -5$ dB, a significant ($p < 0.0083$) medium ($0.28 < |d_C| < 0.43$) difference exists between Lombard and non-Lombard systems for both the

**Table C.8:** $p$-values ($p$) and effect sizes (Cliff's delta, $d_C$) for the MUSHRA experiments. The significant level (0.0083) for the $p$-values is corrected with the Bonferroni method.

| | −5 dB SNR | | 5 dB SNR | |
|---|---|---|---|---|
| Comparison | $p$ | $d_C$ | $p$ | $d_C$ |
| AO-L - AO-NL | $< .0083$ | .30 | .0134 | .22 |
| AV-L - AV-NL | $< .0083$ | .32 | $< .0083$ | .23 |
| AO-L - AV-L | .0498 | −.14 | .7476 | .02 |
| AO-NL - AV-NL | $< .0083$ | −.21 | .8262 | −.02 |
| AO-L - Unproc. | .0479 | .57 | $< .0083$ | .74 |
| AV-L - Unproc. | .0134 | .59 | $< .0083$ | .79 |

audio-only and the audio-visual cases. The increment in quality when using vision with respect to audio-only systems is perceived by the subjects ($|d_C| > 0.11$), but it has only a relatively small effect ($|d_C| < 0.28$). This was expected, since visual cues affect more the intelligibility at low SNRs than quality, as also shown by objective measures (Figure C.2). More specifically, for non-Lombard systems, this difference is significant and greater than the one found for Lombard systems, meaning that vision contributes more when the enhancement of Lombard speech is performed with systems that were not trained with it. We can notice that there is a large ($|d_C| > 0.43$) difference between the unprocessed signals and the version enhanced with Lombard systems. However, this difference is not significant, probably due to the heterogeneous interpretation of the MUSHRA scale by the subjects and their preference of the different natures of the impairment (presence of noise or artefacts caused by the enhancement).

At an SNR of 5 dB a small difference between Lombard and non-Lombard systems is observed, despite being not significant in the audio-only case ($p = 0.0134$). At this noise level, audio-visual systems appear to be indistinguishable ($|d_C| < 0.11$) from the respective audio-only systems. This confirms the intuition that vision does not help in improving the speech quality at high SNRs. Finally, the difference between the unprocessed signals and the respective enhanced versions using Lombard systems is both large ($|d_C| > 0.43$) and significant ($p < 0.0083$), which makes it clear that both AO-L and AV-L improve the speech quality.

## 5.2 Speech Intelligibility Test

The intelligibility test was carried out by 11 listeners, who volunteered to be part of the study. The participants were between 24 and 65 years old, and had self-reported normal hearing and normal (or corrected to normal) vision. On average, each participant spent approximately 45 minutes to complete the test.

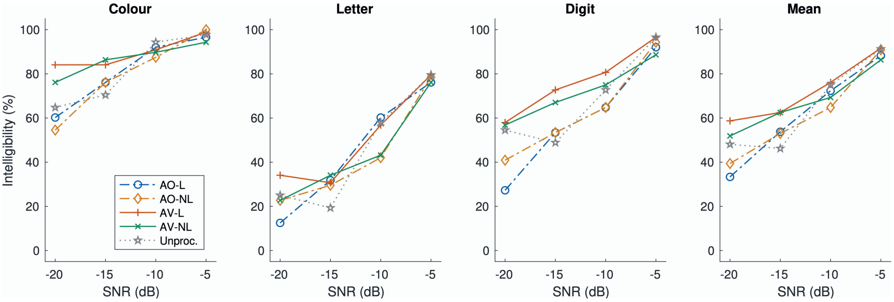


**Fig. C.8:** Percentage of correctly identified words obtained in the listening tests for the colour, the letter, and the digit fields, averaged across 11 subjects. The mean intelligibility scores for all the fields are also reported.

### 5.2.1 Procedure

Each subject was presented with 2 sequences of 80 audio-visual stimuli from the Lombard GRID corpus: 8 speakers × 4 SNRs ($-20$, $-15$, $-10$, and $-5$ dB) × 5 processing conditions (unprocessed, AO-L, AO-NL, AV-L, AV-NL). The participants were asked to listen to each stimulus only once and, based on what they heard, they had to select the colour and the digit from a list of options and to write the letter (Table C.1). The order of presentation of the stimuli was randomised.

Before the actual test, the participants were trained in a special separate session consisting of a sequence of 40 audio-visual stimuli.

### 5.2.2 Results and Discussion

The mean percentage of correctly identified keywords as a function of the SNR is shown in Figure C.8. We can see that among the three fields, the colour is the easiest word to be identified by the participants. In general, the following trends can be observed. At low SNRs the intelligibility of the signals enhanced with AV systems is higher than the intelligibility obtained with AO systems. This difference substantially diminishes when the SNR increases. There is no big performance difference between L and NL systems, but in general AV-L tends to have higher percentage scores than the other systems. AV-L is also the only system that does not decrease the mean intelligibility scores for all the fields if compared to the unprocessed signals.

Table C.9 shows Cliff's deltas and $p$-values, computed with the paired two-sided Wilcoxon signed-rank test, as in the MUSHRA experiments.

The effect sizes support the observations made from Figure C.8. Medium and large differences ($|d_C| > 0.28$) exist between AO and AV systems, especially at low SNRs. While AO-L and AO-NL are indistinguishable ($|d_C| < 0.11$) for SNR $< -10$ dB, there is a medium ($0.28 \leq |d_C| < 0.43$)

**Table C.9:** $p$-values ($p$) and effect sizes (Cliff's delta, $d_C$) for the mean intelligibility scores for all the keywords obtained in the listening tests.

| $p$ | SNR | | | |
|---|---|---|---|---|
| Comparison | -20 dB | -15 dB | -10 dB | -5 dB |
| AO-L - AO-NL | .3066 | .4688 | .0430 | .2539 |
| AV-L - AV-NL | .0625 | .8633 | .0742 | .1055 |
| AO-L - AV-L | .0010 | .0117 | .5625 | .2344 |
| AO-NL - AV-NL | .0527 | .0430 | .3359 | .2070 |
| AO-L - Unproc. | .0332 | .0547 | .9004 | .1250 |
| AV-L - Unproc. | .1270 | .0078 | .8828 | .8828 |
| $d_C$ | SNR | | | |
| Comparison | -20 dB | -15 dB | -10 dB | -5 dB |
| AO-L - AO-NL | $-.08$ | .06 | .31 | $-.31$ |
| AV-L - AV-NL | .32 | .01 | .39 | .28 |
| AO-L - AV-L | $-.91$ | $-.35$ | $-.17$ | $-.34$ |
| AO-NL - AV-NL | $-.32$ | $-.37$ | $-.31$ | .21 |
| AO-L - Unproc. | $-.31$ | .17 | $-.09$ | $-.26$ |
| AV-L - Unproc. | .18 | .46 | 0 | .08 |

difference between AV-L and AV-NL, except for $-15$ dB SNR ($d_C = 0.01$). Moreover, the intelligibility increase of AV-L over the unprocessed signals is perceived by the subjects at SNR $\leq -15$ dB ($|d_C| > 0.11$).

Regarding the $p$-values, if we focus on each SNR separately, the difference between two approaches can be considered significant for $p < 0.0083$ (cf. Section 5.1.2). This condition is met only when we compare AO-L with AV-L at $-20$ dB SNR and AV-L with the noisy speech at $-15$ dB SNR.

There are three main sources of variability that most likely prevent the differences to be significant. First, the variation in lipreading ability among individuals is large and does not directly reflect the variation found in auditory speech perception skills [89]. Secondly, individuals have very different fusion responses to discrepancy in the auditory and visual syllables [90], which in our case might occur due to the artefacts produced in the enhancement process. Finally, the participants were not exposed to the same utterances processed with the different approaches like in MUSHRA. Since the vocabulary set of the Lombard GRID corpus is small and some words are easier to understand because they contain unambiguous visemes, the intelligibility scores are affected not only by the various processing conditions, but also by the different sentences used.

# 6 Conclusion

In this paper, we presented an extensive analysis of the impact of Lombard effect on audio, visual and audio-visual speech enhancement systems based on deep learning. We conducted several experiments using a database consisting of 54 speakers and showed the general benefit of training a system with Lombard speech.

In more detail, we first trained systems with Lombard or non-Lombard speech and evaluated them on Lombard speech adopting a cross-validation setup. The results showed that systems trained with Lombard speech outperform the respective systems trained with non-Lombard speech in terms of both estimated speech quality and estimated speech intelligibility. We also observed a performance difference across speakers, with an evident gap between genders: the performance difference between the systems trained with Lombard speech and the ones trained with non-Lombard speech is larger for females than it is for males. The analysis that we performed suggests that this difference might be primarily due to the large increment in the fundamental frequency that female speakers exhibit from non-Lombard to Lombard conditions.

With the objective of building more general systems able to deal with a wider SNR range, we then trained systems using Lombard and non-Lombard speech and compared them with systems trained only on non-Lombard speech. As in the narrow SNR case, systems that include Lombard speech perform considerably better than the others at low SNRs. At high SNRs, the estimated speech quality and the estimated speech intelligibility obtained with systems trained only with non-Lombard speech are higher, even though the performance gap is very small for the audio and the audio-visual cases. Combining non-Lombard and Lombard speech for training in the way we did guarantees a good compromise for the enhancement performance across all the SNRs.

We also performed subjective listening tests with audio-visual stimuli, in order to evaluate the systems in a situation closer to the real-world scenario, where the listener can see the face of the talker. For the speech quality test, we found significant differences between Lombard and non-Lombard systems at all the used SNRs for the audio-visual case and only at $-5$ dB SNR for the audio-only case. Regarding the speech intelligibility test, we observed that on average the scores obtained with the audio-visual system trained with Lombard speech are higher than the other processing conditions.

This work can be considered as a first extensive study of the impact of Lombard effect on speech enhancement systems based on deep learning. Although we tried to answer several research questions, there are still other areas that fall outside the scope of the current paper, but are still worth exploring in future works:

- This study shows that training systems with Lombard speech is beneficial. However, collecting large-scale Lombard speech data is not practical. A possible path for future research would be to generate synthetic Lombard speech signals: we know the acoustic characteristics of Lombard speech, so one might artificially generate it by modifying non-Lombard speech data and use it for training the systems. Another possibility is to reduce the gap between Lombard and non-Lombard systems with transfer learning: if a system is pre-trained with a large dataset, even in non-Lombard condition, the network will be more robust to different speaking styles. One can expect that the general performance of such a network, as well as its generalisation to Lombard speech, would improve.

- All the experiments of this paper used speakers observed by the networks at training time. As we already mentioned in Section 4, this was a choice made to isolate the impact of Lombard effect from other factors that could make it hard to interpret the results. However, it is interesting to explore the behaviour of the systems in a scenario where speakers are not observed during training. We conducted some initial experiments in this direction in our previous work [47]. There, we showed that there is still a benefit in training a system with Lombard speech, despite being of different speakers. It would be of interest to expand this early study with more extensive experiments. Currently, such experiment is limited by the availability of sufficient training data.
- In our evaluation through audio-visual listening tests, we were unable to find significant differences in intelligibility for most of the comparisons. This suggests a benefit of designing new paradigms for speech intelligibility tests to control the several sources of variability caused by the combination of auditory and visual stimuli.

## 7 Acknowledgements

This work was supported, in part, by the Oticon Foundation.

## References

[1] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC press, 2007.

[2] E. Lombard, "Le signe de l'elevation de la voix," *Annales des Maladies de L'Oreille et du Larynx*, vol. 37, no. 2, pp. 101–119, 1911.

[3] S. A. Zollinger and H. Brumm, "The evolution of the Lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, no. 11-13, pp. 1173–1198, 2011.

[4] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[5] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, “The impact of the Lombard effect on audio and visual speech recognition systems,” *Speech Communication*, vol. 100, pp. 58–68, 2018.

[6] Y. Lu and M. Cooke, “Speech production modifications produced by competing talkers, babble, and stationary noise,” *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[7] E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan, and H. C. Yehia, “Audiovisual Lombard speech: Reconciling production and perception,” in *Proceedings of AVSP*, 2007, p. 41.

[8] M. Garnier, N. Henrich, and D. Dubois, “Influence of sound immersion and communicative interaction on the Lombard effect,” *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 3, pp. 588–608, 2010.

[9] M. Garnier, L. Ménard, and G. Richard, “Effect of being seen on the production of visible speech cues. A pilot study on Lombard speech,” in *Proceedings of Interspeech/ICSLP*, 2012, pp. 611–614.

[10] M. Cooke, S. King, M. Garnier, and V. Aubanel, “The listening talker: A review of human and algorithmic context-induced modifications of speech,” *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, 2014.

[11] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, “Effects of noise on speech production: Acoustic and perceptual analyses,” *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[12] A. L. Pittman and T. L. Wiley, “Recognition of speech produced in noise,” *Journal of Speech, Language, and Hearing Research*, 2001.

[13] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[14] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[15] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[16] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[17] T. Gerkmann and R. Martin, “On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling,” *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4165–4174, 2009.

[18] R. Martin and C. Breithaupt, “Speech enhancement in the DFT domain using Laplacian speech priors,” in *Proceedings of IWAENC*, vol. 3, 2003, pp. 87–90.

[19] Y. Hu and P. C. Loizou, “A comparative intelligibility study of single-microphone noise reduction algorithms,” *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[20] J. Jensen and R. C. Hendriks, “Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 92–102, 2012.

[21] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[22] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *Proceedings of ICASSP*. IEEE, 2015, pp. 708–712.

[23] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[24] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, “On training targets and objective functions for deep-learning-based audio-visual speech enhancement,” in *Proceedings of ICASSP*, 2019, pp. 8077–8081.

[25] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[26] M. Kolbæk, Z.-H. Tan, and J. Jensen, “Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[27] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proceedings of Interspeech*, 2013, pp. 436–440.

[28] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *Proceedings of GlobalSIP*. IEEE, 2014, pp. 577–581.

[29] S. R. Park and J. Lee, “A fully convolutional neural network for speech enhancement,” in *Proceedings of Interspeech*, 2017, pp. 1993–1997.

[30] D. Michelsanti and Z.-H. Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” in *Proceedings of Interspeech*, 2017, pp. 2008–2012.

[31] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[32] W. H. Sumby and I. Pollack, “Visual contribution to speech intelligibility in noise,” *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[33] N. P. Erber, “Auditory-visual perception of speech,” *Journal of Speech and Hearing Disorders*, vol. 40, no. 4, pp. 481–492, 1975.

[34] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[35] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[36] I. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[37] A. Abel and A. Hussain, "Novel two-stage audiovisual speech filtering in noisy environments," *Cognitive Computation*, vol. 6, no. 2, pp. 200–217, 2014.

[38] A. Abel, A. Hussain, and B. Luo, "Cognitively inspired speech processing for multimodal hearing technology," in *Proceedings of CICARE*. IEEE, 2014, pp. 56–63.

[39] J.-C. Hou *et al.*, "Audio-visual speech enhancement based on multimodal deep convolutional neural network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[40] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proceedings of Interspeech*, 2018, pp. 1170–1174.

[41] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[42] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proceedings of Interspeech*, 2018, pp. 3244–3248.

[43] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of ECCV*, 2018, pp. 631–648.

[44] G. Morrone, L. Pasa, V. Tikhanoff, S. Bergamaschi, L. Fadiga, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *Proceedings of ICASSP*, 2019, pp. 6900–6904.

[45] J. H. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, 2009.

[46] P. Heracleous, C. T. Ishi, M. Sato, H. Ishiguro, and N. Hagita, "Analysis of the visual Lombard effect and automatic recognition experiments," *Computer Speech & Language*, vol. 27, no. 1, pp. 288–300, 2013.

[47] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Effects of Lombard reflex on the performance of deep-learning-based audio-visual speech enhancement systems," in *Proceedings of ICASSP*, 2019, pp. 6615–6619.

[48] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[49] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[50] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[51] Y. Lu and M. Cooke, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1495–1499, 2009.

[52] M. Garnier and N. Henrich, "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?" *Computer Speech & Language*, vol. 28, no. 2, pp. 580–597, 2014.

[53] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proceedings of SLT*. IEEE, 2016, pp. 305–311.

[54] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[55] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of ICML*, 2015, pp. 448–456.

[57] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[58] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of CVPR*, 2017, pp. 1125–1134.

[59] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of AISTATS*, 2010, pp. 249–256.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of ICLR*, 2015.

[61] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[62] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[63] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.

[64] ITU, "Recommendation P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2005.

[65] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[66] L. Liu and M. T. Özsu, *Encyclopedia of database systems*. Springer New York, NY, USA, 2009, vol. 6.

[67] L. J. Raphael, G. J. Borden, and K. S. Harris, *Speech science primer: Physiology, acoustics, and perception of speech*. Lippincott Williams & Wilkins, 2007.

[68] D. S. Moore, G. P. McCabe, and B. A. Craig, *Introduction to the Practice of Statistics*. WH Freeman New York, 2012.

[69] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Lœvenbruck, "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," in *Proceedings of Interspeech/ICSLP*, 2006, pp. 2246–2249.

[70] L. Y. Tang, B. Hannah, A. Jongman, J. Sereno, Y. Wang, and G. Hamarneh, "Examining visible articulatory features in clear and plain speech," *Speech Communication*, vol. 75, pp. 1–13, 2015.

[71] N. Alghamdi, "Visual speech enhancement and its application in speech perception training," Ph.D. dissertation, University of Sheffield, 2017.

[72] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.fon.hum.uva.nl/praat/, 2001, accessed: March 20, 2019.

[73] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of CVPR*, 2014, pp. 1867–1874.

[74] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[75] A. Sharma, *Text book of correlations and regression*. Discovery Publishing House, 2005.

[76] A. Field, *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[77] I. Almajai, B. Milner, and J. Darch, "Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise," in *Proceedings of Interspeech/ICSLP*, 2006, p. 1634.

[78] A. Hussain *et al.*, "Towards multi-modal hearing aid design and evaluation in realistic audio-visual settings: Challenges and opportunities," in *Proceedings of CHAT*, 2017, pp. 29–34.

[79] E. Vincent, "MUSHRAM: A MATLAB interface for MUSHRA listening tests," http://c4dm.eecs.qmul.ac.uk/downloads/#mushram, 2005, accessed: March 20, 2019.

[80] EBU, "EBU recommendation R128 - Loudness normalisation and permitted maximum level of audio signals," 2014.

[81] ITU, "Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," 2003.

[82] C. Mendonça and S. Delikaris-Manias, "Statistical tests with mushra data," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[83] F. Winter, H. Wierstorf, C. Hold, F. Krüger, A. Raake, and S. Spors, "Colouration in local wave field synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1913–1924, 2018.

[84] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[85] H. Hentschke and M. C. Stüttgen, "Computation of measures of effect size for neuroscience data sets," *European Journal of Neuroscience*, vol. 34, no. 12, pp. 1887–1894, 2011.

[86] G. M. Sullivan and R. Feinn, "Using effect size - or why the P value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–282, 2012.

[87] N. Cliff, "Dominance statistics: Ordinal analyses to answer ordinal questions," *Psychological bulletin*, vol. 114, no. 3, p. 494, 1993.

[88] A. Vargha and H. D. Delaney, "A critique and improvement of the CL common language effect size statistics of McGraw and Wong," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 2, pp. 101–132, 2000.

[89] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[90] D. B. Mallick, J. F. Magnotti, and M. S. Beauchamp, "Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type," *Psychonomic Bulletin & Review*, vol. 22, no. 5, pp. 1299–1307, 2015.

# Paper D

## Vocoder-Based Speech Synthesis from Silent Videos

Daniel Michelsanti, Olga Slizovskaia, Gloria Haro,
Emilia Gómez, Zheng-Hua Tan, Jesper Jensen

The paper has been accepted for publication in
*Proceedings of Interspeech*, 2020.

*The layout has been revised.*

# Abstract

*Both acoustic and visual information influence human perception of speech. For this reason, the lack of audio in a video sequence determines an extremely low speech intelligibility for untrained lip readers. In this paper, we present a way to synthesise speech from the silent video of a talker using deep learning. The system learns a mapping function from raw video frames to acoustic features and reconstructs the speech with a vocoder synthesis algorithm. To improve speech reconstruction performance, our model is also trained to predict text information in a multi-task learning fashion and it is able to simultaneously reconstruct and recognise speech in real time. The results in terms of estimated speech quality and intelligibility show the effectiveness of our method, which exhibits an improvement over existing video-to-speech approaches.*

# 1 Introduction

Most of the events that we experience in our life consist of visual and acoustic stimuli. Recordings of such events may lack the acoustic component, for example due to limitations of the recording equipment or technical issues in the transmission of the information. Since acoustic and visual modalities are often correlated, methods to reconstruct audio signals using videos have been proposed [1–3].

In this paper, we focus on one particular case of the aforementioned problem: *speech reconstruction (or synthesis) from a silent video*. Solving this task might be useful to automatically generate speech for surveillance videos and for extremely challenging speech enhancement applications, e.g. hearing assistive devices, where noise completely dominates the target speech, making the acoustic signal worth less than its video counterpart.

A possible way to tackle the problem is to decompose it into two steps: first, a *visual speech recognition* (VSR) system [4–6] predicts the spoken sentences from the video; then, a *text-to-speech* (TTS) model [7–9] synthesises speech based on the output of the VSR system. However, at least two drawbacks can be identified when such an approach is used. In order to generate speech from text, each word should be spoken in its entirety to be processed by the VSR and the TTS systems, imposing great limitations for real-time applications. Furthermore, when the TTS method is applied, useful information that should be captured by the system, such as emotion and prosody, gets lost, making the synthesised speech unnatural. For these reasons, approaches that estimate speech from a video, without using text as an intermediate step, have been proposed.

Le Cornu and Miller [10, 11] developed a video-to-speech method with a focus on speech intelligibility rather than quality. This is achieved by estimating spectral envelope (SP) audio features from visual features and then reconstructing the time-domain signal with the STRAIGHT vocoder [12]. Since the vocoder also requires other audio features, i.e. the fundamental frequency (F0) and the aperiodic parameter (AP), these are artificially created independently of the visual features.

Ephrat and Peleg [13] treated speech reconstruction as a regression problem using a neural network which takes as input raw visual data and predicts a line spectrum pairs (LSP) representation of linear predictive coding (LPC) coefficients computed from the audio signal. The waveform is reconstructed from the estimated audio features using Gaussian white noise as excitation, producing unnatural speech. This issue is tackled in a subsequent work [14], where a neural network estimates the mel-scale spectrogram of the audio from video frames and optical flow information derived from the visual input. The time-domain speech signal is reconstructed using either example-based synthesis, in which estimated audio features are replaced with their closest match in the training set, or speech synthesis from predicted linear-scale spectrograms.

Akbari et. al. [15] tried to reconstruct natural sounding speech using a neural network that takes as input the face region of the talker and estimates bottleneck features extracted from the auditory spectrogram by a pre-trained autoencoder. The time-domain signal is obtained with the algorithm in [16]. This approach shows its effectiveness when compared to [13].

All the methods reported until now have a major limitation: they estimate either a magnitude spectrogram, SPs or LSPs, which do not contain all the information of a speech signal. Vougioukas et al. [17] addressed this issue and proposed an end-to-end model that can directly synthesise audio waveforms from videos using a generative adversarial network (GAN). However, their direct estimation of a time-domain signal causes artefacts in the reconstructed speech.

In this work, we propose an approach, *vid2voc*, to estimate WORLD vocoder [18] features from the silent video of a speaker[1]. We trained the systems using either the whole face or the mouth region only, since previous work [13] shows a benefit in using the entire face. Our method differs from the work in [10, 11], because we predict all the vocoder features (not only SP) directly from raw video frames. The estimation of F0 and AP, alongside with SP, allows to have a framework with a focus on speech intelligibility (as in [10, 11]) and speech quality, able to outperform even the recently proposed GAN-based approach in [17] in several conditions. In addition, we train a system that can simultaneously perform speech reconstruction (our main goal) and VSR, in a multi-task learning fashion. This can be useful in

[1] Although this paper aims at synthesising speech from frontal-view silent videos, it is worth mentioning that some methods using multi-view video feeds have also been developed [19–22].

all the applications that require video captioning without adding considerable extra complexity to the system. Although Kumar et al. [21] incorporate a text-prediction model in their multi-view speech reconstruction pipeline, this model is trained separately from the main system and it is quite simple: it classifies encoded audio features estimated with a pre-trained network into 10 text classes. This makes the method dependent on the number of different sentences of the specific database used for training and not suitable for real-time applications. Instead, we make use of the more flexible connectionist temporal classification (CTC) [23] sequence modelling which has already shown its success in VSR [4].

Additional material, including samples of reconstructed speech that the reader is encouraged to listen to for a better understanding of the effectiveness of our approach, can be found in `https://danmic.github.io/vid2voc/`.

# 2 Methodology and Experimental Setup

## 2.1 Audio-Visual Speech Corpus

Experiments are conducted on the GRID corpus [24], which consists of audio and video recordings from 34 speakers (s1−34), 18 males and 16 females, each of them uttering 1000 six-word sentences with the following structure: <command> <color> <preposition> <letter> <digit> <adverb>. Each video has a resolution of 720×576 pixels, a duration of 3 s and a frame rate of 25 frames per second. The audio tracks have the same duration as the videos and a sample frequency of 50 kHz. In addition, text transcription for every utterance is provided.

As in [17], we evaluate our systems in speaker dependent and speaker independent settings. Regarding the speaker dependent scenario, the data from 4 speakers (s1, s2, s4, s29) is pooled together, then 90% of the data is used for training, 5% for validation and 5% for testing. Regarding the speaker independent scenario, the data from 15 speakers (s1, s3, s5−8, s10, s12, s14, s16, s17, s22, s26, s28, s32) is used for training, the data from 7 speakers (s9, s20, s23, s27, s29, s30, s34) for validation and the data from 10 speakers (s2, s4, s11, s13, s15, s18, s19, s25, s31, s33) for testing.

## 2.2 Audio and Video Preprocessing

The acoustic model used in this work is based on the WORLD vocoder [18], with a sample frequency of 50 kHz and a hop size of 250 samples[2]. WORLD consists of three analysis algorithms to determine SP, F0 and AP features,

[2] The window length is automatically determined by the WORLD algorithm.

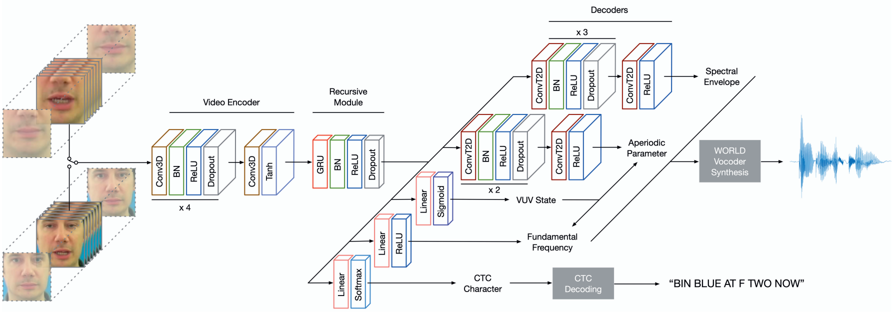


**Fig. D.1:** Pipeline of our system. Conv3D: 3-D convolution. BN: Batch normalisation. GRU: Gated recurrent unit. ConvT2D: 2-D transposed convolution. VUV: Voiced-unvoiced. CTC: Connectionist temporal classification.

and a synthesis algorithm which incorporates these three features. Here, we use SWIPE [25] and D4C [26] to estimate F0 and AP, respectively. As done in [27], a dimensionality reduction of the features is applied: SP is reduced to 60 log mel-frequency spectral coefficients (MFSCs) and AP is reduced to 5 coefficients according to the D4C band-aperiodicity estimation. In addition, a voiced-unvoiced (VUV) state is obtained by thresholding the F0 obtained with SWIPE. All the acoustic features are min-max normalised using the statistics of the training set as in [28].

As in [17], videos are preprocessed as follows: first, the faces are aligned to the canonical face[3]; then, the video frames are normalised in the range $[-1, 1]$, resized to $128 \times 96$ pixels and, for the models that use only the mouth region as input, cropped preserving the bottom half; finally, the videos are mirrored with a probability of 0.5 during training.

## 2.3 Architecture and Training Procedure

As shown in Figure D.1, our network maps video frames of a speaker to vocoder features and consists of a *video encoder*, a *recursive module* and five decoders: *SP decoder*, *AP decoder*, *VUV decoder*, *F0 decoder* and *VSR decoder*. We also tried not to use the VSR decoder, to see whether it has any impact on the performance.

The video encoder is inspired by [17]: it takes as input one video frame concatenated with the three previous and the three next frames and applies five 3-D convolutions (conv3D). Each of the first four convolutional layers is followed by batch normalisation (BN) [30], ReLU activation and dropout [31], while the last one is followed by Tanh activation.

[3] We use the face processor library in `https://github.com/DinoMan/face-processor`, which makes use of [29].

To model the sequential nature of video data, a recursive module is used: it consists of a single-layer gated recurrent unit (GRU) [32], BN, ReLU activation and dropout.

Each decoder takes the GRU features as input. For every video frame the SP decoder produces an eight-frame-long estimate $\widehat{W}_{se} \in \mathbb{R}^{60\times 8}$ of the normalised dimensionality-reduced SP, $W_{se}$, through three 2-D transposed convolutions (convT2D), each followed by BN, ReLU activation and dropout, and another convT2D followed by ReLU activation.

The VUV decoder consists of a linear layer followed by ReLU activation. A threshold of 0.2 is applied to the output obtaining $\widehat{W}_{vuv} \in \mathbb{R}^{8}$, an estimate of the VUV state, $W_{vuv}$.

The AP decoder has a structure similar to the SP decoder, with a total of three convT2D in this case. Its output, $O_{nap} \in \mathbb{R}^{5\times 8}$, together with $\widehat{W}_{vuv}$ is used to get $\widehat{W}_{nap}$, an estimate of $W_{nap} = I_{5,8} - W_{ap}$, where $I_{5,8}$ indicates an all-ones matrix with 5 rows and 8 columns, and $W_{ap}$ is the normalised dimensionality-reduced AP:

$$(\widehat{W}_{nap})_i = (O_{nap})_i \odot \widehat{W}_{vuv} \quad \text{for } i \in \{1, \ldots, 5\} \tag{D.1}$$

where $(A)_i$ indicates the $i$-th row of $A$ and $\odot$ denotes the element-wise product.

The F0 decoder has a linear layer followed by a sigmoid activation function. Its output, $O_{f0} \in \mathbb{R}^{8}$, is point-wise multiplied with $\widehat{W}_{vuv}$ to obtain $\widehat{W}_{f0}$, an estimate of the normalised F0, $W_{f0}$:

$$\widehat{W}_{f0} = O_{f0} \odot \widehat{W}_{vuv}. \tag{D.2}$$

Finally, the VSR decoder, consisting of a linear and a softmax layers, outputs a CTC character that will be used to predict the text transcription of the utterance.

The system is trained to minimise the following loss:

$$J = \frac{\lambda_1}{\lambda} J_{se} + \frac{\lambda_2}{\lambda} J_{nap} + \frac{\lambda_3}{\lambda} J_{f0} + \frac{\lambda_4}{\lambda} J_{vuv} + \frac{\lambda_5}{\lambda} J_{vsr} \tag{D.3}$$

where $\lambda_1 = 600$, $\lambda_2 = 50$, $\lambda_3 = 10$, $\lambda_4 = 10$, $\lambda_5 = 1$, $\lambda = \sum_{i=1}^{5} \lambda_i$ and:

- $J_{se}$: mean squared error (MSE) between $W_{se}$ and $\widehat{W}_{se}$.
- $J_{nap}$: MSE between $W_{nap}$ and $\widehat{W}_{nap}$.
- $J_{f0}$: MSE between $W_{f0}$ and $\widehat{W}_{f0}$.
- $J_{vuv}$: MSE between $W_{vuv}$ and $\widehat{W}_{vuv}$.
- $J_{vsr}$: CTC loss [23] between the target text transcription and the estimated one.

Details regarding architecture and training hyperparameters can be found in Table D.1.

### 2.4 Waveform Reconstruction and Lipreading

The network outputs are used to reconstruct the speech waveform with the WORLD synthesis algorithm [18] and to get a text transcription adopting the best path CTC decoding scheme [23].

### 2.5 Evaluation Metrics

The system is evaluated in terms of perceptual evaluation of speech quality (PESQ) [35] and extended short-time objective intelligibility (ESTOI) [36], two of the most used measures that provide estimates of speech quality and speech intelligibility, respectively. PESQ scores are in the range from $-0.5$ to 4.5 and ESTOI scores practically lie between 0 and 1. In both cases, higher values correspond to better performance.

For the systems having the VSR decoder, we also provide the word error rate (WER), a standard metric for automatic speech recognition systems. In this case, lower values correspond to better performance.

## 3 Results and Discussion

As shown in Table D.2, four systems are trained based on the input (mouth or full face) and the presence of the VSR decoder (only speech synthesis or speech synthesis and VSR).

The systems are compared with the recently proposed GAN-based approach in [17]. As an additional baseline, we also report the PESQ score for [15], since this method, which makes use of bottleneck features extracted from auditory spectrograms, outperforms [17] in terms of estimated speech quality for the speaker dependent case.

### 3.1 Speaker Dependent Case

Table D.3 (left part) shows the speaker dependent results. We observe that our models outperform the approach in [17] in terms of both PESQ and ESTOI by a considerable margin. Vougioukas et al. [17] mention that their system produces low-power hum artefacts that affect the performance. They tried to solve the issue by applying average filtering to the output of their network, experiencing a rise of the PESQ score from 1.71 to 1.80 (not shown in Table D.3), comparable to [15], but still appreciably lower than the results we achieve. However, this filtering negatively affected the intelligibility of the produced speech signals, and was not used in the final system.

**Table D.1:** Architecture and training hyperparameters. Activation, batch normalisation and dropout omitted for brevity.

| **Input Size** | | | | | |
|---|---|---|---|---|---|
| $B\times S\times C\times F\times H\times W$ | | | | | |
| **Video Encoder** | | | | | |
| Layer | Input Channels | Output Channels | Kernel Size | Stride | Padding |
| Conv3D | 3 | 64 | (7,4,4) | (1,2,2) | (0,1,1) |
| Conv3D | 64 | 128 | (1,4,4) | (1,2,2) | (0,1,1) |
| Conv3D | 128 | 256 | (1,4,4) | (1,$d_1$,2) | (0,1,1) |
| Conv3D | 256 | 512 | (1,4,4) | (1,2,2) | (0,1,1) |
| Conv3D | 512 | 128 | (1,$d_2$,6) | (1,1,1) | (0,0,0) |
| **Recursive Module** | | | | | |
| Layer | Input Size | Hidden Size | | | |
| GRU | 128 | 128 | | | |
| **Spectral Envelope (SP) Decoder** | | | | | |
| Layer | Input Channels | Output Channels | Kernel Size | Stride | Padding |
| ConvT2D | 128 | 256 | (1,6) | (1,1) | (0,0) |
| ConvT2D | 256 | 128 | (2,4) | (1,2) | (0,0) |
| ConvT2D | 128 | 64 | (4,4) | (1,2) | (0,0) |
| ConvT2D | 64 | 1 | (4,2) | (1,2) | (0,0) |
| **Aperiodic Parameter (AP) Decoder** | | | | | |
| Layer | Input Channels | Output Channels | Kernel Size | Stride | Padding |
| ConvT2D | 128 | 128 | (4,1) | (1,1) | (0,0) |
| ConvT2D | 128 | 64 | (3,3) | (1,1) | (0,0) |
| ConvT2D | 64 | 1 | (3,3) | (1,1) | (0,0) |
| **Voiced-Unvoiced (VUV) Decoder** | | | | | |
| Layer | Input Size | Output Size | | | |
| Linear | 128 | 8[a] | | | |
| **Fundamental Frequency (F0) Decoder** | | | | | |
| Layer | Input Size | Output Size | | | |
| Linear | 128 | 8[a] | | | |
| **Visual Speech Recognition (VSR) Decoder** | | | | | |
| Layer | Input Size | Output Size | | | |
| Linear | 128 | 28[b] | | | |
| **Extra Information** | | | | | |
| The system is implemented in Pytorch [33] and trained for $N$ iterations using the Adam optimizer [34] with a learning rate of 0.0001, $\beta_1$=0.5 and $\beta_2$=0.9. The model that performs the best in terms of PESQ on the validation set is used for testing.<br>$S$=75 (sequence length). $C$=3 (image channels).<br>$F$=7 (consecutive video frames). $W$=96 (video frame width).<br>If the full face is used as input:<br>$B$=16 (batch size). $H$=128 (video frame height). $d_1$=3. $d_2$=5.<br>If only the mouth is used as input:<br>$B$=24 (batch size). $H$=64 (video frame height). $d_1$=2. $d_2$=4.<br>In the speaker dependent case, the dropout probability of each dropout layer is $p_d$=0.2. $N$=300000.<br>In the speaker independent case, $p_d$=0.5 for the video encoder and the GRU, and $p_d$=0.2 for the rest. $N$=185000. | | | | | |

[a] Eight is the number of the output audio frames corresponding to the video frame used as input (together with its context).

[b] The 28 CTC characters consist of the 26 letters of the English alphabet, one space character and one blank token.

**Table D.2:** Systems used in this study.

| | Input | |
|---|---|---|
| | Mouth | Face |
| w/o VSR Decoder | vid2voc-M | vid2voc-F |
| w/ VSR Decoder | vid2voc-M-VSR | vid2voc-F-VSR |

**Table D.3:** Results for the speaker dependent and the speaker independent cases. Best performance (except WORLD) in bold.

| | Speaker Dependent | | | Speaker Independent | | |
|---|---|---|---|---|---|---|
| Mean Scores | PESQ ↑ | ESTOI ↑ | WER ↓ | PESQ ↑ | ESTOI ↑ | WER ↓ |
| Approach in [15][a] | 1.82 | - | - | - | - | - |
| Approach in [17] | 1.71 | 0.329 | - | 1.24 | 0.198 | - |
| vid2voc-M | 1.89 | 0.448 | - | 1.20 | 0.214 | - |
| vid2voc-M-VSR | **1.90** | **0.455** | 15.1% | 1.23 | **0.227** | **51.6%** |
| vid2voc-F | 1.85 | 0.439 | - | 1.19 | 0.202 | - |
| vid2voc-F-VSR | 1.88 | 0.447 | **14.4%** | **1.25** | 0.210 | 69.3% |
| WORLD[b] | 3.06 | 0.759 | - | 3.03 | 0.759 | - |

[a] Value taken from the experiments in [17].

[b] WORLD indicates the reconstruction retrieved from the vocoder features of the clean speech signals and it is a performance upper bound of our systems.

Among the systems we developed (cf. Table D.2), we observe that including the VSR decoder in the pipeline is beneficial for the speech reconstruction task (see Table D.3). Moreover, the use of the mouth as input not only is sufficient to synthesise speech, but it also allows to achieve higher estimated speech quality and intelligibility if compared to the models that use the whole face of the speaker as input. This might be explained by the fact that handling an input with a larger dimensionality is harder if we want to keep roughly the same deep architecture with a similar number of parameters. However, when the whole face is used as input, the WER is slightly lower, indicating that there might be a performance trade-off between VSR and speech reconstruction that should be further investigated in future work in relation with other multi-task learning techniques.

## 3.2 Speaker Independent Case

Regarding the speaker independent scenario (cf. right part of Table D.3), we observe that the performance gap between the approach in [17] and our systems is not as large as for the speaker dependent case. Although our models appear to perform slightly better than [17] in terms of ESTOI, the PESQ scores are similar. This can be explained by the fact that some speech characteristics, e.g. F0, cannot be easily estimated for unseen speakers. Since it is reasonable to think that people having similar facial characteristics (e.g. due to gender,

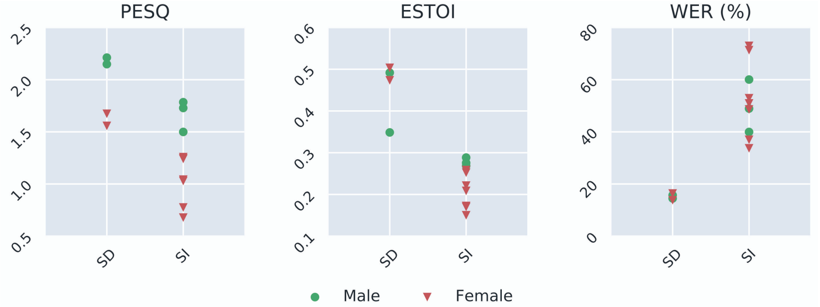


**Fig. D.2:** Results of the vid2voc-M-VSR models for the speaker dependent (SD) and the speaker independent (SI) cases. Each marker indicates the mean score of a speaker.

age etc.) have similar speech characteristics (cf. [37], where the face of a person was predicted from a speech signal), we expect that training a network with a dataset that includes more speakers might be beneficial: such a network can produce an average voice of speakers from the training set that share similar facial traits with an unseen talking face.

Among the systems we developed, the presence of the VSR decoder still gives an advantage for speech reconstruction. Unlike the speaker dependent case, the WER for the model that uses the whole face as input is higher than the system using only the mouth. This is due to the early stopping technique that we adopt, which tends to favour speech reconstruction over VSR, indicating again the trade-off between these two tasks.

Finally, Figure D.2 shows the results for the vid2voc-M-VSR models by speaker. We can see that the spread of the scores is much higher for the speaker independent case in particular for WER. This is in line with the observations reported in [17], suggesting the different performance between the estimated speech of subjects whose facial traits substantially differ from the speakers in the training set and the others.

# 4 Conclusion

In this study, we reconstructed speech from silent videos using a deep model that estimates WORLD vocoder features. We tested our approach in both speaker dependent and speaker independent scenarios. In both cases, we were able to obtain speech signals with estimated speech quality and intelligibility generally higher if compared to a recently proposed GAN-based approach. In addition, we designed our system to simultaneously perform visual speech recognition by using a decoder that estimates CTC characters from a given video sequence.

Future work includes: (a) the adoption of self-paced multi-task learning techniques; (b) the improvement of the visual speech recognition performance, e.g. with a beam search decoding scheme; (c) the design of a system that can generalise well to unseen speakers in noncontrolled environments.

## 5 Acknowledgment

The authors would like to thank Konstantinos Vougioukas, Stavros Petridis, Pritish Chandna and Merlijn Blaauw.

This research is partially funded by: the William Demant Foundation; the TROMPA H2020 project (770376); the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Program (MDM-2015-0502) and the Social European Funds; the MICINN/FEDER UE project (PGC2018-098625-B-I00); the H2020-MSCA-RISE-2017 project (777826 NoMADS).

## References

[1] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 79–, 2014.

[2] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. of CVPR*, 2016.

[3] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in *Proc. of CVPR*, 2018.

[4] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," in *Proc. of GTC*, 2017.

[5] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Proc. of Interspeech*, 2017.

[6] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. of CVPR*, 2017.

[7] J. Sotelo *et al.*, "Char2wav: End-to-end speech synthesis," in *Proc. of ICLR Workshop*, 2017.

[8] W. Ping *et al.*, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. of ICLR*, 2018.

[9] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. of ICASSP*, 2018.

[10] T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. of Interspeech*, 2015.

[11] ——, "Generating intelligible audio speech from visual speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, 2017.

[12] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[13] A. Ephrat and S. Peleg, “Vid2speech: Speech reconstruction from silent video,” in *Proc. of ICASSP*, 2017.

[14] A. Ephrat, T. Halperin, and S. Peleg, “Improved speech reconstruction from silent video,” in *ICCV Workshop on Computer Vision for Audio-Visual Media*, 2017.

[15] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, “Lip2audspec: Speech reconstruction from silent lip movements video,” in *Proc. of ICASSP*, 2018.

[16] T. Chi, P. Ru, and S. A. Shamma, “Multiresolution spectrotemporal analysis of complex sounds,” *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[17] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, “Video-driven speech reconstruction using generative adversarial networks,” in *Proc. of Interspeech*, 2019.

[18] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[19] Y. Kumar, M. Aggarwal, P. Nawal, S. Satoh, R. R. Shah, and R. Zimmermann, “Harnessing AI for speech reconstruction using multi-view silent video feed,” in *Proc. of ACM-MM*, 2018.

[20] Y. Kumar, R. Jain, M. Salik, R. R. Shah, R. Zimmermann, and Y. Yin, “Mylipper: A personalized system for speech reconstruction using multi-view visual feeds,” in *Proc. of ISM*, 2018.

[21] Y. Kumar, R. Jain, K. M. Salik, R. R. Shah, Y. Yin, and R. Zimmermann, “Lipper: Synthesizing thy speech using multi-view lipreading,” in *Proc. of AAAI*, 2019.

[22] S. Uttam *et al.*, “Hush-hush speak: Speech reconstruction using silent videos,” in *Proc. of Interspeech*, 2019.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. of ICML*. ACM, 2006.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[25] A. Camacho, “SWIPE: A sawtooth waveform inspired pitch estimator for speech and music,” Ph.D. dissertation, University of Florida Gainesville, 2007.

[26] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, 2016.

[27] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” *Applied Sciences*, vol. 7, no. 12, p. 1313, 2017.

[28] P. Chandna, M. Blaauw, J. Bonada, and E. Gomez, “A vocoder based method for singing voice extraction,” in *Proc. of ICASSP*, 2019.

[29] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks),” in *Proc. of ICCV*, 2017.

[30] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. of ICML*, 2015, pp. 448–456.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proc. of EMNLP*, 2014.

[33] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in NeurIPS*, 2019.

[34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. of ICLR*, 2015.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs,” in *Proc. of ICASSP*, 2001.

[36] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[37] T.-H. Oh *et al.*, “Speech2face: Learning the face behind a voice,” in *Proc. of CVPR*, 2019.

# Paper E

## An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang,
Yong Xu, Meng Yu, Dong Yu, Jesper Jensen

The paper has been submitted to
*IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

This page intentionally left blank.

This paper is currently under review.
A pre-print is available at the following link:
`https://arxiv.org/abs/2008.09586`

This page intentionally left blank.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-698-4

AALBORG UNIVERSITY PRESS