Evaluating the Potential Gain of Auditory and Audiovisual Speech-Predictive Coding Using Deep Learning

Thomas Hueber, Eric Tatulli, Laurent Girin, Jean-Luc Schwartz

LETTER **Communicated by Karl Friston**

Evaluating the Potential Gain of Auditory and Audiovisual Speech-Predictive Coding Using Deep Learning

Thomas Hueber
thomas.hueber@gipsa-lab.grenoble-inp.fr
Eric Tatulli
eric.tatulli@gmail.com
Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Laurent Girin
laurent.girin@gipsa-lab.grenoble-inp.fr
Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France, and Inria Grenoble-Rhône-Alpes, 38330 Montbonnot-Saint Martin, France

Jean-Luc Schwartz
jean-luc.schwartz@gipsa-lab.grenoble-inp.fr
Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Sensory processing is increasingly conceived in a predictive framework in which neurons would constantly process the error signal resulting from the comparison of expected and observed stimuli. Surprisingly, few data exist on the accuracy of predictions that can be computed in real sensory scenes. Here, we focus on the sensory processing of auditory and audiovisual speech. We propose a set of computational models based on artificial neural networks (mixing deep feedforward and convolutional networks), which are trained to predict future audio observations from present and past audio or audiovisual observations (i.e., including lip movements). Those predictions exploit purely local phonetic regularities with no explicit call to higher linguistic levels. Experiments are conducted on the multispeaker LibriSpeech audio speech database (around 100 hours) and on the NTCD-TIMIT audiovisual speech database (around 7 hours). They appear to be efficient in a short temporal range (25–50 ms), predicting 50% to 75% of the variance of the incoming stimulus, which could result in potentially saving up to three-quarters of the processing power. Then they quickly decrease and almost vanish after 250 ms. Adding information on the lips slightly improves predictions, with a 5% to 10% increase in explained variance. Interestingly the visual gain vanishes more slowly, and the gain is maximum for a delay of 75 ms between image and predicted sound.

Neural Computation **32, 596–625** (2020)
https://doi.org/10.1162/neco_a_01264

1 Introduction

1.1 The Predictive Brain. The concept of "predictive brain" progressively emerged in neurosciences in the 1950s (Attneave, 1954; Barlow, 1961). It assumes that the brain is constantly exploiting the redundancy and regularities of the perceived information, hence reducing the amount of processing by focusing on what is new and eliminating what is already known. After half a century of experimental developments, the predictive brain has been mathematically encapsulated by Friston and colleagues into a powerful framework based on Bayesian modeling (Friston, 2005), assimilating such concepts as perceptual inference (Friston, 2003), reinforcement learning (Friston, Daunizeau, & Kiebel, 2009), and optimal control (Friston, 2011). In this framework, it has been proposed that the minimization of free energy, a concept from thermodynamics, could provide a general principle associating perception and action in interaction with the environment in a coherent, predictive process (Friston, Kilner, & Harrison, 2006; Friston, 2010). A number of recent neurophysiological studies confirm the accuracy of the predictive coding paradigm for analyzing sensory processing in the human brain (e.g., Keller & Mrsic-Flogel, 2018).

Actually, predictive coding is a general methodological paradigm in information processing that consists of analyzing the local regularities in an input data stream in order to extract the predictable part of these input data. After optimization of the coding process, the difference signal (i.e., prediction error) provides an efficient summary of the original signal. Technically, this can be cast in terms of minimum description lengths (MDL; Grünwald, Myung, & Pitt, 2005) and accompanying (variational) free energy minimization. The predictive brain is conceived as an inference engine with a fixed structure whose parameters are tuned to provide optimal predictions, that is, predictions of incoming signals from past ones, optimizing mutual information between both sets. This principle was introduced by Barlow (1961) in terms of redundancy reduction.

The information processing system can then focus on the difference between input data and their prediction. In a very general manner, whatever the processing system is, there are two main advantages to processing the difference signal over directly processing the input signal. First, if the prediction is efficient, the difference signal is generally of (much) lower energy than the original signal, which leads to energy consumption saving in subsequent processes and resource saving for representing the signal with a given accuracy (e.g., bit rate saving in an audio or a video coder). In short, this reduces the "cost" of information processing. Second, there is a concentration of novelty or unpredictable information in the difference signal, which is exploitable for, for example, the detection of new events. Because of these advantages, predictive coding has been largely exploited in technological applications, in particular in signal processing for telecommunications (Gersho & Gray, 1992; Jayant & Noll, 1984).

1.2 Predictions in Speech. Speech involves different linguistic levels from the acoustic-phonetic level up to the lexical/syntactic/semantic and pragmatic levels. Each level of language processing is likely to provide predictions (Manning & Schütze, 1999), and globally, automatic speech recognition (ASR) systems are based on statistical predictive models of the structure of speech units in the acoustic input (Jelinek, 1976; Rabiner, 1989; Deng & Li, 2013). Generative grammar traditionally conceives linguistic rules as providing the basis of language complexity at all levels (Jackendoff, 2002), and the existence of correlations between linguistic units at various ranges has been the focus of a large amount of scientific research—for example, by Kaplan and Kay (1994) and Berent (2013) for phonology, Heinz and Idsardi (2011) for lexicon or syntax, and Oberlander and Brew (2000) and Altmann, Cristadoro, and Degli Esposti (2012) for semantics in textual chains. The search for dependencies between linguistic units at various scales may also be related to the more general framework describing long-term dependencies in symbolic sequences (Li, 1990; Li & Kaneko, 1992; Ebeling & Neiman, 1995; Montemurro & Pury, 2002; Lin & Tegmark, 2017).

These different levels of prediction in speech correspond to different temporal scales, ranging from a few tens or hundreds of milliseconds for the lowest linguistic units (i.e., phoneme/syllable) up to a few hundred milliseconds or seconds for the highest ones (i.e., word/phrase/utterance). In the human brain, exploiting these different levels is done by hierarchically organized computational processes that correspond to a large network of cortical areas, as described in a number of recent review papers (e.g., Friederici & Singer, 2015). In this network, fast auditory and phonetic processing is supposed to occur locally in the auditory cortex (superior temporal sulcus/gyrus, STS/STG), while slower lexical access and syntactic processing involve information propagation within a larger network associating the temporal, parietal, and frontal cortices (Hickok & Poeppel, 2007; Giraud & Poeppel, 2012; Friederici & Singer, 2015).

Importantly, Arnal and Giraud (2012) have identified rapid cortical circuits that seem to provide predictions at low temporal ranges in the human brain. They propose that the predictions in time ("when" something important would happen) are based on a coupling between low-frequency oscillations driven by the syllabic rhythm in the delta-theta channel of neural firing (around 2–8 Hz) and midfrequency regulation in the beta channel of neural firing (12–30 Hz). The "what" information would combine top-down predictions conveyed by the beta channel with analysis of the sensory input providing prediction errors to be conveyed to higher centers in a bottom-up process through the gamma channel (30–100 Hz). Such local gamma-theta-beta auditory structures typically operate at relatively short temporal scales (up to a few hundreds of milliseconds) characteristic of phonetic processes and likely to operate at the level of auditory cortical areas in the STS/STG region (Gagnepain, Henson, & Davis, 2012; Mesgarani, Cheung, Johnson, & Chang, 2014). These local circuits mostly operate without

lexical and postlexical processes that would require both larger temporal scales and longer cortical loops.

To our knowledge, such predictions occurring at the acoustic-phonetic level have never been quantified. Still, it is of real importance to evaluate what is the nature and amount of phonetic predictions that can be made locally in the speech input. This letter is focused on the quantitative analysis of temporal predictions in speech signals at a phonetic, sublexical level, using state-of-the-art machine learning models.

1.3 Predictions in Speech Coding Systems. In essence, the largest and historically prominent family of computational models for speech signal predictions is found in speech coding techniques for telecommunications. The vast majority of standardized predictive speech codecs apply prediction of a speech signal waveform sample from a linear combination of the preceding samples in the range of about 1 ms. This is the basis of the famous linear predictive coding (LPC) technique and LPC family of speech coders (Markel & Gray, 1976). The predictor coefficients are calculated over successive so-called short time frames of signal of a few tens of milliseconds (typically 20–30 ms), every 10–20 ms. Globally, LPC techniques may be related to the general principle of MDL minimization mentioned in section 1.1, where the MDL model is the set of prediction coefficients (Kleijn & Ozerov, 2007).

The prediction power of the LPC technique within a single short-term frame has been largely quantified in the speech coding literature. However, this literature has quite poorly considered the prediction of speech at the level of one to several short time frames ahead (i.e., a few tens to a few hundred milliseconds—in other words, an intermediary timescale in between the speech sample level and the lexical level). This is mostly due to constraints on latency in telecommunications. For example, only a very few studies have applied some form of predictive coding on vectors of parameters encoding a short-term speech frame. This has been done using differential coding (Yong, Davidson, & Gersho, 1988), recursive coding (Samuelsson & Hedelin, 2001; Subramaniam, Gardner, & Rao, 2006), or Kalman filtering (Subasingha, Murthi, & Andersen, 2009). Yet these approaches are limited to one-step frame prediction. A few other "unconventional" studies (Atal, 1983; Farvardin & Laroia, 1989; Mudugamuwa & Bradley, 1998; Dusan, Flanagan, Karve, & Balaraman, 2007; Girin, Firouzmand, & Marchand, 2007; Girin, 2010; Ben Ali, Djaziri-Larbi, & Girin, 2016) have proposed "long-term" speech coders, which aim at exploiting speech signal redundancy and predictability over larger time spans, typically in the range of a few hundreds of milliseconds.[1] However, these

[1] Such coders are limited to speech storage since interactive communication is not feasible with the resulting high latency.

methods actually implement a joint coding of several short-term frames (basically, by using trajectory models or projections) but do not apply any explicit prediction of a frame given past frames. In short, to the best of our knowledge, no study has yet attempted to systematically quantify the predictability of the acoustic speech signal at the phonetic to syllable timescale (i.e., one to several short-term frames). A first objective of this letter is to address this question thanks to a (deep) machine learning approach that we present.

1.4 Visual Potential Contribution to Phonetic Predictions in Speech. Importantly, the visual input can also convey relevant information for acoustic-phonetic predictions. As a matter of fact, pioneer studies such as Besle, Fort, Delpuech, and Giard (2004) and Van Wassenhove, Grant, and Poeppel (2005) showed that the visual component of an audiovisual speech input (e.g., "ba") could result in decreasing the first negative peak N1 in the auditory event-related potential pattern in electroencephalographic (EEG) data. Peak decrease has been related to the ability of the visual input to provide predictive cues likely to suppress the auditory response displayed in N1. The potential predictive role of vision is supported by behavioral data showing that vision of the speaker's face may indeed provide cues for auditory prediction (e.g., Sánchez-García, Alsius, Enns, & Soto-Faraco, 2011; Venezia, Thurman, Matchin, George, & Hickok, 2016).

It has been claimed that the predictive aspect of visual speech information might be enhanced by the fact that there is often an advance of image on sound in natural speech (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Actually, this remains a matter of controversy (Schwartz & Savariaux, 2014). Still, studies on audiovisual speech coders capable of exploiting correlation between audio and visual speech are extremely sparse (though see the pioneering studies in Rao & Chen, 1996, and Girin, 2004). Hence, here also, no systematic quantification of the potential role of the visual input in the predictive coding of speech stimuli has been realized yet. Providing such a quantification, using a machine learning approach, is the second objective of this study.

1.5 Our Contribution: Modeling and Assessing Mid-Term Predictability in Acoustic and AudioVisual Speech. The goal of this study is to quantify what is really predictable online from the speech acoustic signal and the visual speech information (mostly lip movements). To this aim, we propose to use a series of computational models based on artificial (deep) neural networks trained to predict future acoustic features from past information. We focus on midterm prediction, that is, a prediction at the level of sequences of multiple consecutive short-term frames (from 25 ms to 450 ms in our experiments) and with no explicit access to lexical or postlexical information. Depending on the representation (audio or audiovisual), different network architectures such as feedforward neural

networks and convolutional neural networks are used to learn sequences of acoustic and visual patterns. In order to generalize the network speech prediction capabilities across many speakers, these networks are trained on large multispeaker audio and audiovisual speech databases. More specifically, we use the LibriSpeech corpus (Panayotov, Chen, Povey, & Khudanpur, 2015), one of the largest publicly available acoustic speech databases, and the NTCD-TIMIT corpus (Abdelaziz, 2017), one of the largest publicly available audiovisual speech databases.

The choice of a statistical framework based on deep learning was motivated by its ability to build successive levels of increasingly meaningful abstractions in order to learn and perform complex (e.g., nonlinear) mapping functions. By combining different types of generic layers (e.g., fully connected, convolutional, recurrent) and training their parameters jointly from raw data, deep neural networks provide a generic methodology for feature extraction, classification, and regression. Deep learning–based models have led to significant performance improvement in many speech processing problems, for example, acoustic automatic speech recognition (ASR; Abdel-Hamid et al., 2014), speech enhancement (Wang & Chen, 2018), audiovisual and visual ASR (Mroueh, Marcheret, & Goel, 2015; Wand, Koutník, & Schmidhuber, 2016; Tatulli & Hueber, 2017), articulatory-to-acoustic mapping (Bocquelet, Hueber, Girin, Savariaux, & Yvert, 2016), and more generally for tasks involving speech-related biosignals (Schultz et al., 2017). Thus, deep-learning models are here considered as providing an accurate evaluation of the amount of information and regularities present in the auditory and visual inputs and likely to intervene in speech-predictive coding in the human brain. Though substantially different from biological neural networks, artificial deep neural networks provide a computational solution to cognitive questions and may thus provide some insight into the nature of biological processes (Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018).

The proposed computational models of predictive speech coding enabled us to address the following questions:

- How much of the future speech sounds can be predicted from the present and the past ones? What is the temporal range at which acoustic-phonetic predictions may operate, and how much of past information do they capitalize on for predicting future events?
- If the visual input (i.e., information on the speaker's lip movements) is added to the acoustic input (i.e., the speech sound), how much gain can occur in prediction, and what temporal window of visual information is typically useful for augmenting auditory predictions? Crucially, can audiovisual predictions confirm the assumption that visual information would be available prior to auditory information in the predictive coding of speech, and by what temporal amount?

2 Materials and Methods

2.1 Database. Two publicly available data sets were used in this study. The first is the LibriSpeech corpus, which is derived from read audiobooks from the LibriVox project (Panayotov et al., 2015). In this study, we used the "train-clean-100" subset of LibriSpeech, which contains 100.6 hours of read English speech, uttered by 251 speakers (125 female speakers and 126 male speakers). The second data set is the NTCD-TIMIT data set (Abdelaziz, 2017), which contains audio and video recordings of 59 English speakers, each uttering the same 98 sentences extracted from the TIMIT corpus (Garofolo et al., 1993)—5782 sentences in total, representing around 7 hours of speech). NTCD-TIMIT contains both clean and noisy versions of the audio material. In our study, we used only the clean audio signals. As for the video material, NTCD-TIMIT provides a postprocessed version of raw video sequences of the speaker's face focusing on the region of interest (ROI) around the mouth. This includes cropping, rotation, and scaling of the extracted ROI so that the mouths of all speakers approximately lie on the same horizontal line and have the same width. Each ROI image is finally resized as a 67×67 pixels 8-bit gray-scale image (Abdelaziz, 2017).

Librispeech is our favored data set here for quantifying the auditory speech prediction from audio-only input. Experiments conducted on the NTCD-TIMIT corpus aim more specifically at quantifying the potential benefit of combining audio and visual inputs (over audio-only input) for such prediction. In spite of its reduced size compared to Librispeech (around 7 hours and 59 speakers versus 100 hours and 251 speakers), it remains one of the largest publicly available audiovisual data sets of continuous speech.

2.2 Data Preprocessing. For the LibriSpeech corpus, no specific preprocessing of the audio signal was done. For the NTCD-TIMIT corpus, each audiovisual recording was first cropped in order to reduce the amount of silence before and after each uttered sentence. Temporal boundaries of silence portions were extracted from the phonetic alignment file provided with the data set. In order to take into account anticipatory lip gestures, a safe margin of 150 ms of silence was kept intact before and after each recorded sentence.

A sliding window was used to segment each waveform into short-term acoustic frames. A classical frame length of 25 ms was used in our study (400 samples at 16 kHz). Importantly, a frame shift of 25 ms was chosen in order to avoid any overlap between consecutive frames (i.e., the frame shift was set equal to the frame length). This aimed at preventing the introduction of artificial correlation due to shared samples, which could introduce some bias in the midterm prediction (i.e., the prediction of a speech frame given the preceding ones).

The discrete Fourier transform (DFT) was applied on each frame to represent its spectral content. The overall process is referred to as the

short-term Fourier transform (STFT) analysis, and the resulting signal representation is the STFT (complex-valued) spectrogram. In our study, a 512-point fast Fourier transform (FFT) was used to calculate each DFT (each 400-sample short-term frame was zero-padded with 112 zeros and was then applied a Hanning analysis window). Only the 257 first coefficients in the frequency dimension, corresponding to positive frequencies, are retained. Then we computed the log magnitude of the STFT spectrogram (on a dB scale) and rescaled the resulting values to the range $[0, -80]$ dB for each sentence of the data set (the maximum value over each sentence was set to 0 dB and all values below -80 dB were set to -80 dB). Finally, the short-term speech spectrum was converted into a set of so-called Mel-frequency cepstral coefficients (MFCC). Such coefficients were obtained by integrating subbands of the log power spectrum using a set of 40 triangular filters equally spaced on a nonlinear Mel frequency scale and converting the resulting 40-dimensional Mel-frequency log spectrum into a 13-dimensional vector using the discrete cosine transform (DCT). The resulting representation for a complete utterance (a sequence of frames) is referred to as the MFCC spectrogram.

MFCC coefficients are widely used in many fields such as automatic speech recognition (ASR; Rabiner, 1989) and music information retrieval (e.g. classification of musical sound; Kim et al., 2010). MFCC analysis can be seen as a high-level biologically inspired process related to psychoacoustics (i.e., simulating the cochlear filtering). Moreover, MFCC analysis leads to a compact representation of the short-term speech spectrum, which may be of significant interest in the context of statistical learning since it may limit the number of free model parameters to estimate. All the above audio analysis procedures were performed using the Librosa Python open-source library, release 0.6.0 (McFee et al., 2018).

As concerns the video sequences (for the NTCD-TIMIT corpus), a linear interpolation across successive images in the pixel domain was performed in order to adjust the video frame rate (originally 30 fps) to the analysis rate of the audio recordings (40 Hz). Each frame of 67×67 pixels was then resized to 32×32 pixels using linear interpolation. Eight-bit integer pixel intensity values were divided by 255 in order to work with normalized values in the $[0, 1]$ range. Video analysis was performed using the *openCV2* Python open-source library (Bradski, 2000; release 3.4.0.12).

2.3 Computational Models of Speech Prediction from Audio-Only Data.

2.3.1 Task. We denote $\mathbf{x}_t$ a D_x-dimensional (column) vector of spectral audio features, which in this study is a 13-dimensional vector of MFCC coefficients. The frame index t is an integer and corresponds to time tH where $H = 25$ ms is the frame spacing (see the previous section). The predictive coding problem using past audio information can be formulated as

computing

$$\widehat{\mathbf{x}}_{t+\tau_f} = f(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-\tau_p}), \tag{2.1}$$

where $\widehat{\mathbf{x}}$ denotes an estimated (predicted) value of $\mathbf{x}$ and τ_f and τ_p denote a time lag in the future and in the past, respectively (in number of frames). In summary, τ_f labels the depth into the future of the prediction, based on a sequence of past observations from the current time step t to $t - \tau_p$. These time indices correspond to 25 ms of clock time. We propose to model the nonlinear predictive function f by an artificial (deep) neural network, described below.

2.3.2 Architectures. To process the MFCC spectrograms, we use a standard feedforward deep neural network (FF-DNN). An FF-DNN is composed of a cascade of fully connected layers. Each neuron of a given fully connected layer performs a nonlinear transformation of a weighted sum of its inputs, which are the outputs of the previous layer. In the case of a regression task (as we consider here with prediction and as opposed to classification), the last (output) layer is directly the weighted sum of its inputs; it has a linear activation function. To process the portion of MFCC spectrogram composed of frames $t - \tau_p$ to t, the latter has to be vectorized, that is, concatenated into a single larger vector, and equation 2.1 is here reformulated as

$$\widehat{\mathbf{x}}_{t+\tau_f} = f_{\text{FF-DNN}}\left(\left[\mathbf{x}_t^T \; \mathbf{x}_{t-1}^T \; \ldots \; \mathbf{x}_{t-\tau_p}^T\right]^T\right), \tag{2.2}$$

where ^{T} denotes vector transpose.

2.3.3 General Methodology for Model Training. All parameters (i.e. weights) of FF-DNNs are learned from data, usually by stochastic gradient descent and backpropagation. Briefly, this consists of iterating the following process: (1) evaluating a loss function, which measures the average discrepancy between the prediction of the network and the ground-truth value for a subset of the training data (called a minibatch), and (2) calculating the gradient of this loss function with respect to all the network weights, starting from the output layer and backpropagating it through all the hidden layers, then (3) updating all weights using the gradient in order to decrease the loss function. This process is applied over all minibatches of the training data and repeated a certain number of times, called epochs, until the loss function no longer significantly evolves.

In addition to this general process, three strategies are often used to prevent model overfitting and accelerate training convergence: (1) early stopping, which consists of monitoring the loss function on a validation data set and stopping the training as soon as its value stops decreasing after a given number of epochs; (2) batch normalization, which consists of

applying a transformation so that the inputs to each layer have zero mean and unit variance (Ioffe & Szegedy, 2015); and (3) dropout, which consists of not updating a random fraction of neurons in a given layer during training. In our study, we combined these three processes.

2.3.4 Model Selection and Training. As in many modeling studies based on deep learning, complex architectures require setting a large number of hyperparameters, mostly related to the sizing of the network, a process known as model selection. It also requires setting several training settings. An extensive search for the optimal combination for these hyperparameters and settings is out of range. Therefore, we optimized only some of them on a subset of each database. We tested combinations of 1, 2, 3, and 4 layers with either 128, 256, or 512 neurons each. This converged to the same architecture for the two data sets, with three groups of 256-neuron fully connected layers. This model is represented in Figure 1a. All models were trained using the Adam optimizer, a popular variant of the stochastic gradient descent (Kingma & Ba, 2014), on minibatches of 256 observations. The Leaky ReLU was used as an activation function (for the neurons of the hidden layers). It is defined as $f(x) = \alpha x$ for $x < 0$ and $f(x) = x$ for $x \geq 0$ (with $\alpha = 0.03$ in our experiments). The mean squared error (MSE) was used as the loss function. In each experiment, 66% of the data (randomly partitioned) were used for training, and the remaining 33% were used for testing. Twenty percent of the training data were used for validation (early stopping). The number of epochs in early stopping was set to 10.

After model selection, the optimal set of hyperparameters and the same training settings were then used to train and evaluate the final computational models of speech prediction from audio-only data. Two separate series of experiments were conducted. These models were trained and evaluated using the entire train-clean-100 subset of the LibriSpeech corpus (around 100 hours, 251 speakers). They were also trained and evaluated on the audio data of the entire NTCD-TIMIT audiovisual corpus (around 7 hours, 59 speakers). The latter series of experiments were mostly done for comparison with their audiovisual counterpart.

Technical implementation of all models was performed using the Keras open-source library (Chollet et al., 2015, release 2.1.3). All models were trained using GPU-based acceleration.

2.4 Computational Models of Speech Prediction from Both Audio and Visual Data.

2.4.1 Task. We denote $\mathbf{I}_t$ the lip image at frame t (in our case a gray-scale image of 32×32 pixels). The predictive coding problem using both audio and visual past information can be formulated as

$$\widehat{\mathbf{x}}_{t+\tau_f} = f(\mathbf{x}_t, \mathbf{I}_t, \mathbf{x}_{t-1}, \mathbf{I}_{t-1}, \ldots, \mathbf{x}_{t-\tau_p}, \mathbf{I}_{t-\tau_p}). \tag{2.3}$$

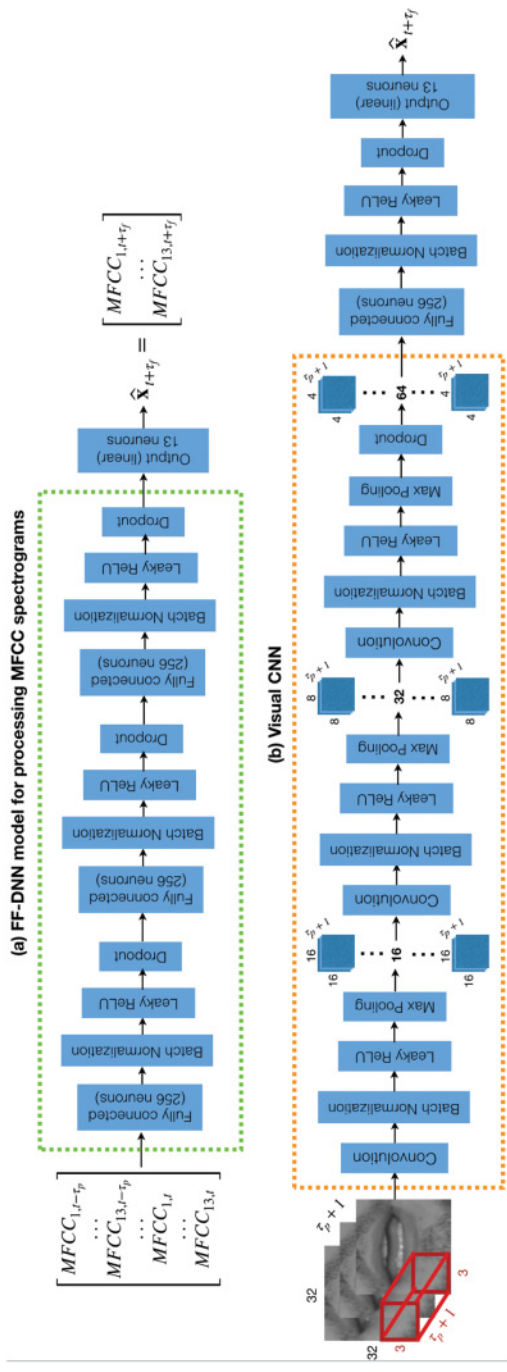

Figure 1: Selected architectures for the audio model and the visual model. (a) FF-DNN model for processing MFCC spectrograms. (b) CNN model for predicting an MFCC vector from a sequence of lip images. In panel b, the red cube represents the 3D filters of the convolutive layers. The dotted rectangles indicate the subnetworks to be used in the audiovisual model (see Figure 2).

2.4.2 Architectures. Integration of audio and visual speech information has been largely considered for automatic audiovisual speech recognition (Potamianos, Neti, Gravier, Garg, & Senior, 2003; Mroueh et al., 2015) and also (though much less extensively) for other applications, such as speech enhancement (Girin, Schwartz, & Feng, 2001) and speech source separation (Rivet, Girin, & Jutten, 2007). Basically, the general principle is that integration can be processed at the input signal level (concatenation of the input data from each modality, that is, *early integration*), at the output level (combination of the outputs obtained separately from each modality, that is, *late integration*), or somewhere in between those extremes (after some separate processing of the inputs and before final calculation of the output, that is, *midlevel integration*) (Schwartz, Robert-Ribes, & Escudier, 1998). Artificial neural networks provide an excellent framework for such multimodal integration, since it can be easily implemented with a fusion layer receiving the inputs from different streams and generating a corresponding output. Moreover, the fusion layer can be placed arbitrarily close to the input or the output.

In this study, we propose a computational model of auditory speech from both audio and visual inputs based on artificial neural networks. We adopt the midlevel fusion strategy, which enables to benefiting from the design and training of the audio (FF-DNN) network used for predictive coding based on audio-only input presented in the previous section, and the design and training of a visual model dedicated to process the lip images.

A convolutional neural network (CNN; LeCun, Bengio, & Hinton, 2015) was used as the core of this visual model. A CNN is a powerful network architecture well adapted to process 2D data for classification and regression. It can extract a set of increasingly meaningful representations along its successive layers. It is thus widely used in image and video processing—for example, object detection (Szegedy et al., 2015), gesture recognition (Baccouche, Mamalet, Wolf, Garcia, & Baskurt, 2012; Ji, Xu, Yang, & Yu, 2013; Karpathy et al., 2014; Simonyan & Zisserman, 2014), or visual speech recognition (Noda, Yamaguchi, Nakadai, Okuno, & Ogata, 2014; Tatulli & Hueber, 2017).

Technically, a CNN is a deep (multilayer) neural network classically composed of one or several convolutional layers, pooling layers, fully connected layers, and one output layer. In a nutshell, a convolutional layer convolves an input 2D image with a set of so-called local filters and then applies a nonlinear transformation to the convolved image. The output is a set of so-called feature maps. Each feature map can be seen as the (nonlinear) response of the input image to the corresponding local filter. One important concept in the convolutional layer is weight sharing, which states that the parameters of each filter remain the same whatever the position of the filter in the image. This allows the CNN to exploit spatial data correlation and build translation-invariant features. A pooling layer then downsamples each feature map in order to build a scale-invariant representation. For

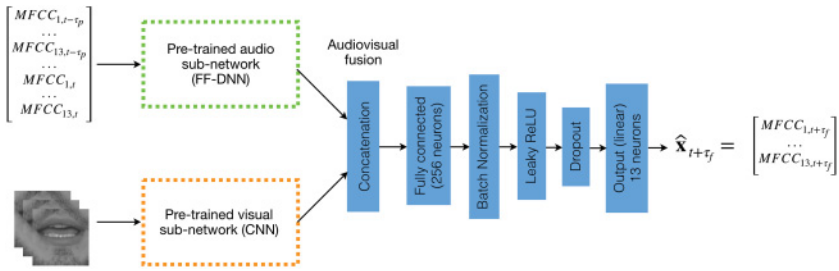

Figure 2: Selected architecture for the audiovisual model. Audio and visual pre-trained subnetworks from Figure 1 are merged using a 256-neuron fully connected fusion layer.

example a so-called max-pooling layer outputs the max value observed on subpatches of a feature map. The convolutional + pooling process can be cascaded several times. A CNN generally ends up with a series of fully connected layers that have the same function as in a standard feedforward deep neural network, as described in section 2.3. Note that in a CNN, the first fully connected layer usually operates over a vectorized form of the downsampled feature maps provided by the last pooling layer.

We thus first designed and trained such a visual CNN for efficient visual speech feature extraction from speakers' lip images. Its architecture is represented in Figure 1b. This visual CNN maps a sequence of a speaker's lip images into the corresponding future MFCC vector. Because we process a sequence of $(\tau_p + 1)$ images, the 2D convolution is extended to a 3D convolution, including the temporal dimension, as illustrated by the red cube in Figure 1b. Then, the convolutional + pooling layers of the visual CNN and the fully connected layers of the MFCC FF-DNN were selected. These subnetworks were merged using a fully connected fusion layer, which is followed by other usual layers. The resulting network regressing audio and visual data into audio data is represented in Figure 2. Each portion of the present and past MFCC spectrogram and associated sequence of lip images is mapped into an output predicted (future) MFCC vector. This audiovisual model was trained anew with the audiovisual training data.

2.4.3 Model Selection and Training. As concerns the CNN model processing the speakers' lip images (used to initialize the audiovisual model), we tested one, two, or three groups of convolutional or pooling layers. As often done in computer vision tasks involving CNNs (e.g., Noda et al., 2014), the number of filters was incremented in each layer: 16 for the first layer, 32 for the second, 64 for the third. Filters of size 3×3, 5×5, and 10×10 were tested. The pooling factor was fixed to 2×2.

For the audiovisual model (the one jointly processing audio and visual data to predict audio), we used the subnetworks of the selected audio and visual networks, and we varied only the number of fully connected layers $N_{FC} \in \{1, 2, 3\}$, with either 256 or 512 neurons each.

The selected visual CNN has three groups of convolution + pooling layers, with 16, 32, and 64 filters of size $3 \times 3 \times (\tau_p + 1)$, and a single 256-neuron fully connected layer, as represented in Figure 1b. Finally, the audiovisual model merges the subnetworks from the selected audio and visual models using a 256-neuron, fully connected fusion layer, as represented in Figure 2.

Finally, after model selection, both the visual-only model of Figure 1b and the audiovisual model of Figure 2 were (separately) trained on the entire NTCD-TIMIT data set. In each experiment, the settings of the training were very similar to the ones used for training the MFCC spectrogram FF-DNNs (e.g., use of the Adam optimizer, use of 66% of the data set for training, test on the remaining 33%, validation with early stopping on 20% of the training data).

2.5 Metrics. Two metrics were used to assess the prediction performance of the different models: (1) the mean squared error (MSE) between the predicted audio vector and the corresponding ground-truth audio vector (this MSE was also used as a loss function to train the different models) and (2) the weighted explained variance (EV) regression score, evaluating the proportion to which the predicted coefficients account for the variation of the actual ones.

For each pair (τ_p, τ_f) of past context lag and prediction lag, the MSE is first defined per MFCC coefficient (indexed by d) and per test sentence (indexed by k) as

$$MSE_{\tau_p,\tau_f,k,d} = \frac{1}{T_k - \tau_f - \tau_p} \sum_{t=\tau_p+1}^{T_k-\tau_f} (\widehat{\mathbf{x}}_{k,d,t+\tau_f} - \mathbf{x}_{k,d,t+\tau_f})^2, \tag{2.4}$$

where T_k is the number of audio vectors (i.e., the number of acoustic short-term frames) in sentence k, and $\widehat{\mathbf{x}}_{k,d,t}$ and $\mathbf{x}_{k,d,t}$ are the dth entry of, respectively, the predicted and ground-truth vectors at frame t for the kth sentence. Then it is averaged across MFCC coefficients and across sentences:

$$MSE_{\tau_p,\tau_f,k} = \frac{1}{D} \sum_{d=1}^{D} MSE_{\tau_p,\tau_f,k,d}, \tag{2.5}$$

$$MSE_{\tau_p,\tau_f} = \frac{1}{N_{test}} \sum_{k=1}^{N_{test}} MSE_{\tau_p,\tau_f,k}, \tag{2.6}$$

where $D = 13$ and N_{test} is the number of test sentences. Assuming a gaussian distribution of the errors, a 95% confidence interval of MSE_{τ_p,τ_f} is defined as

$$CI^{MSE}_{\tau_p,\tau_f} = MSE_{\tau_p,\tau_f} \pm \frac{1.96}{\sqrt{N_{test}-1}}\sigma(MSE_{\tau_p,\tau_f,k}), \tag{2.7}$$

where $\sigma(.)$ denotes the empirical standard deviation evaluated over the set of test sentences.

The weighted explained variance (EV) regression score is defined as

$$EV_{\tau_p,\tau_f} = \frac{1}{N_{test}D}\sum_{k=1}^{N_{test}}\sum_{d=1}^{D} w_{\tau_p,\tau_f,k,d} EV_{\tau_p,\tau_f,k,d}, \tag{2.8}$$

with

$$EV_{\tau_p,\tau_f,k,d} = 1 - \frac{\mathrm{Var}\left\{\widehat{\mathbf{x}}_{k,d,t+\tau_f} - \mathbf{x}_{k,d,t+\tau_f}\right\}}{\mathrm{Var}\left\{\mathbf{x}_{k,d,t+\tau_f}\right\}} \tag{2.9}$$

and

$$w_{\tau_p,\tau_f,k,d} = \frac{\mathrm{Var}\left\{\mathbf{x}_{k,d,t+\tau_f}\right\}}{\sum_{d'=1}^{D}\mathrm{Var}\left\{\mathbf{x}_{k,d',t+\tau_f}\right\}}, \tag{2.10}$$

where Var denotes the empirical variance evaluated along the time dimension (the different frames of a sentence). For each sentence, the contribution of the dth coefficient to the explained variance is weighted by the normalized variance of each individual coefficient (see equation 2.10). This avoids a coefficient with very low variance to yield a very large (negative) EV value for that coefficient, which would pollute the average EV value.

The weighted EV is within the interval $]-\infty, 1]$. A value close to 1 indicates that the error between a predicted and ground-truth data is small compared to the ground-truth data themselves—hence, a strong correlation between them. This corresponds to a large prediction gain (larger than 1). An EV value close to 0 generally indicates a poor correlation and a very weak prediction gain (close to 1). Negative EV values indicate that the error is larger than the ground-truth data, hence very inefficient predictions.

Note that in contrast to the EV, the MSE is not weighted and not normalized in any way. Therefore, it is expected to be more sensitive than the EV to potential differences in data sets (e.g., recording material, waveform scaling to avoid clipping). This means that EV values can be more easily compared across our two data sets than MSE values.

Both MSE and EV are strongly related to a key metric of the predictive coding theory, which is the prediction gain (Gersho & Gray, 1992; Markel & Gray, 1976). Indeed, the latter is defined as the ratio of the ground-truth

signal power and the prediction error power (i.e, the MSE):

$$G_{\tau_p,\tau_f,k,d} = \frac{\sum_{t=\tau_p+1}^{T_k-\tau_f}(\mathbf{x}_{k,d,t+\tau_f})^2}{\sum_{t=\tau_p+1}^{T_k-\tau_f}(\widehat{\mathbf{x}}_{k,d,t+\tau_f} - \mathbf{x}_{k,d,t+\tau_f})^2} = \frac{\sum_{t=\tau_p+1}^{T_k-\tau_f}(\mathbf{x}_{k,d,t+\tau_f})^2}{(T_k - \tau_f - \tau_p)MSE_{\tau_p,\tau_f,k,d}}. \tag{2.11}$$

The lower the MSE, the higher the prediction gain. Moreover, in the case where both ground-truth signal and predicted signal are zero mean, we have

$$EV_{\tau_p,\tau_f,k,d} = 1 - \frac{1}{G_{\tau_p,\tau_f,k,d}}. \tag{2.12}$$

Therefore, the higher the prediction gain, the closer to 1 the expected variance is. Those relations are given here for each MFCC coefficient and each sentence. Depending on how averaging across coefficients and sentences is performed, they can become more intricate after averaging. In our study, we define an average prediction gain such as

$$G_{\tau_p,\tau_f} = \frac{1}{1 - EV_{\tau_p,\tau_f}}. \tag{2.13}$$

3 Results and Discussion

The prediction performances of the audio models trained and evaluated on LibriSpeech are presented in Figure 3. The prediction performances of both audio and audiovisual models trained and evaluated on NTCD-TIMIT are presented in Figure 4. Note that in this section, we express the time lags τ_p and τ_f in milliseconds for convenience of discussion. For example, $EV_{75,50}$ denotes the weighted explained variance obtained when predicting a 25 ms audio frame 50 ms in the future, looking 75 ms in the past—that is, using the current frame and the three previous past frames.

3.1 Speech Prediction from Audio Data.

3.1.1 General Trends. The results show that it is indeed possible to predict, to a certain extent, the spectral information in the acoustic speech signal in a temporal range of 200 ms following the current frame. As expected, the accuracy of such predictions decreases rapidly when the temporal horizon τ_f increases. This evolution more or less follows a logarithmic shape for MSE_{τ_p,τ_f} and an exponential decay toward 0 for EV_{τ_p,τ_f}. These general trends are observed on both LibriSpeech and NTCD-TIMIT. The audio-only predictive models trained on the large-scale LibriSpeech corpus are globally slightly more accurate than the ones trained on the smaller data set,

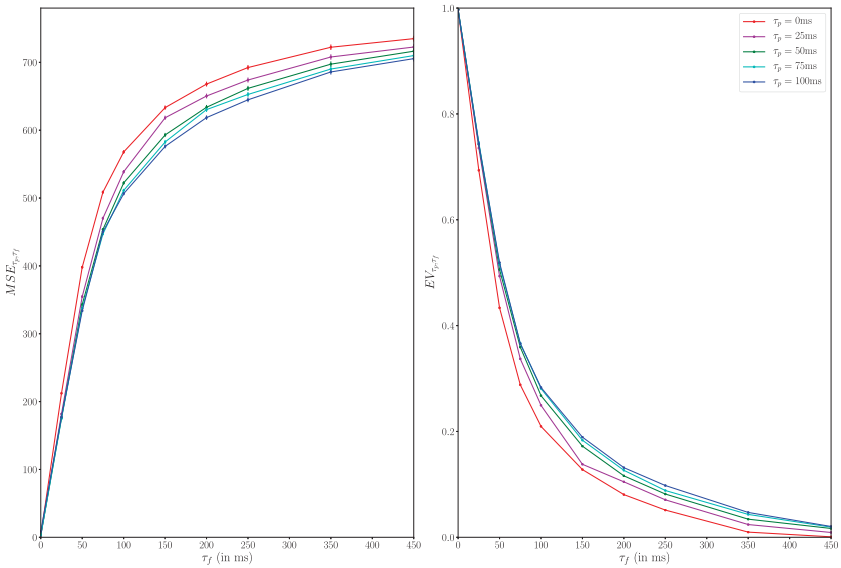

Figure 3: Prediction performance of the audio models using LibriSpeech. Mean square error with 95% confidence interval, represented by the error bars (left), and explained variance regression score (right).

NTCD-TIMIT. This is likely due to the better generalization capacity of the networks when the data set is larger (in terms of number of speakers and speech material per speaker).

At $\tau_f = 25$ ms, the weighted explained variance is about 0.75 for the audio-only model trained on LibriSpeech and about 0.65 for the one trained on NTCD-TIMIT (e.g., for $\tau_p = 75$ ms, $EV_{75,25} = 0.67$ on NTCD-TIMIT and $EV_{75,25} = 0.76$ on LibriSpeech; compare the cyan solid lines in Figures 3 and 4). This corresponds to an average predictive coding gain G_{τ_p,τ_f} around 2.8 and 4, respectively. This provides a rough estimate of the factor by which the power of the error signal (input minus predicted) to transmit by neural processes is reduced compared to the original input. It thus provides some quantification of the amount of biological energy that the system might gain in exploiting a short-range (25 ms) predictive process.

At $\tau_f = 50$ ms, the weighted explained variance is about 0.5 on LibriSpeech (e.g., $EV_{50,50} = 0.51$) and about 0.4 on NTCD-TIMIT (e.g., $EV_{50,50} = 0.41$), which corresponds to an average prediction gain between 1.7 and 2. For the audio-only models trained on LibriSpeech, prediction becomes poor above $\tau_f = 250$ ms: the explained variance goes below 0.1 and keeps on decreasing toward 0. For models trained on NTCD-TIMIT, the performance degradation occurs a bit sooner: the explained variance goes below 0.1 for

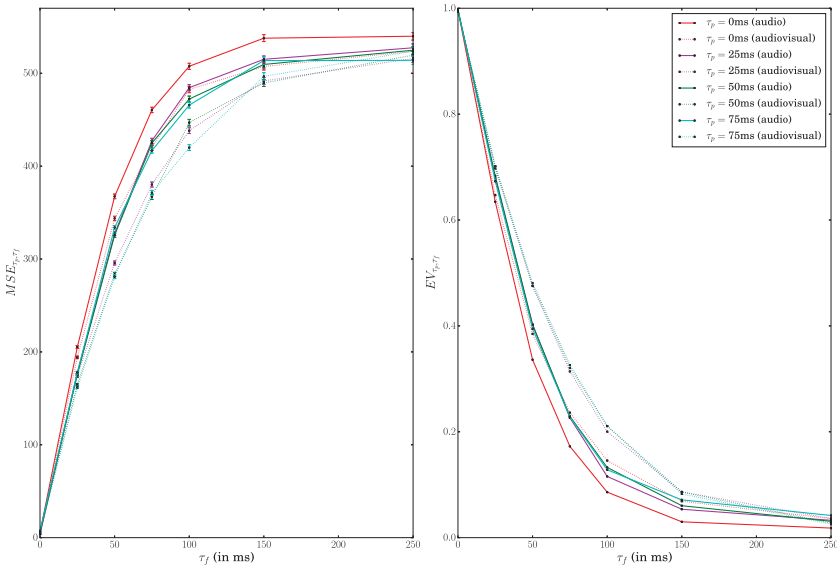

Figure 4: Prediction performance of the audio and audiovisual models using NTCD-TIMIT. Mean square error with 95% confidence interval (represented by the error bars) (left) and explained variance regression score (right).

τ_f between 100 ms and 150 ms, and prediction keeps on decreasing toward 0. Again, the difference between the results obtained with the two data sets is likely due to their difference in size and thus to the resulting difference in generalization properties of the corresponding models. Nevertheless, these results provide a rather coherent estimation of the temporal window in which acoustical predictions are available, typically around the duration of a syllable.

3.1.2 Impact of Past Information. Another aspect of acoustic prediction concerns the role of the temporal context. Unsurprisingly, adding one context frame to the current one provides significant improvement in the prediction of the next frame for both data sets (e.g., $EV(0, 75) = 0.28$ and $EV(25, 75) = 0.33$ on LibriSpeech, $EV(0, 75) = 0.17$ and $EV(25, 75) = 0.22$ on NTCD-TIMIT). Adding a second context frame is also beneficial for the large LibriSpeech corpus (e.g., $EV(50, 75) = 0.36$) though more marginally for the smaller NTCD-TIMIT corpus (e.g., $EV(50, 75) = 0.23$). Such past information may enable the model to evaluate speech trajectories and extract relevant information on the current dynamics, related, for example, to formant transitions, known to be crucial in speech perception. Adding a third frame of past context ($\tau_p = 75$ ms) marginally improves prediction, but only

for τ_f larger than about 75 ms and only for LibriSpeech data. Adding a fourth past frame ($\tau_p = 100$ ms) provides no further gain. While being related to a different task, such results may be compared to classical ones in automatic speech recognition, where adding first and second derivatives of the spectral parameters is classically considered as the optimal choice for reaching the best performance.

3.1.3 Prediction Accuracy per Class of Speech Sound. A fine-grained analysis of the prediction accuracy for five major classes of speech sounds is presented in Figure 5 (this analysis was conducted on the NTCD-TIMIT data set for which phonetic alignment is available).

Interestingly, the prediction accuracy at $\tau_f = 25$ ms is in a comparable range for vowels, fricatives, nasals, and semivowels but significantly lower for plosive sounds (i.e., plosives exhibit a significantly larger MSE; see Figure 5). This may be explained by the difficulty of predicting the precise timing of the occlusion release within the plosive closure and the shape of the corresponding short-term spectrum from the prerelease signal. This pattern is also visible but decreased for predictions at $\tau_f = 50$ ms and $\tau_f = 75$ ms, probably due to a ceiling effect of the prediction power at these temporal horizons.

3.2 Speech Prediction from Audio and Visual Data. The performance of visual-only models (i.e., predictive models of acoustic speech that rely only on lip images, trained and tested on the NTCD-TIMIT corpus) is presented in Figure 6 (left).

As expected, the information provided by the visual modality is real though limited. For example, the best performance obtained at $\tau_f = 0$ ms is $EV_{75,0} = 0.37$ only, which corresponds to a prediction gain of 1.59. This result can be put in perspective with respect to the literature on automatic lip-reading (also known as visual speech recognition), where a typical performance of a visuo-phonetic decoder that does not exploit any high-level linguistic knowledge (via statistical language models) is between 30% and 40% (i.e., 60% to 70% phone error rate). Similarly to the audio-only models, adding past context frames to the current one provides significant improvement in the prediction accuracy. Most of this improvement is observed when considering past information at $\tau_p = 25$ ms—one additional lips image. Adding another past context frame (i.e., $\tau_p = 50$ ms) only marginally improves prediction, and going to three past frames does not provide further improvement.

Interestingly, the performance of visual-only models decreases relatively slowly in comparison with the rapid decrease in prediction accuracy for the audio models (for the NTCD-TIMIT data set) in the same range of time lags. For example, for $\tau_f = 50$ ms, we have $EV_{50,50} = 0.32$, and for $\tau_f = 75$ ms, we have $EV_{50,75} = 0.25$. However, on average, visual modality does not seem to convey useful information above $\tau_f = 100$ ms (e.g., at

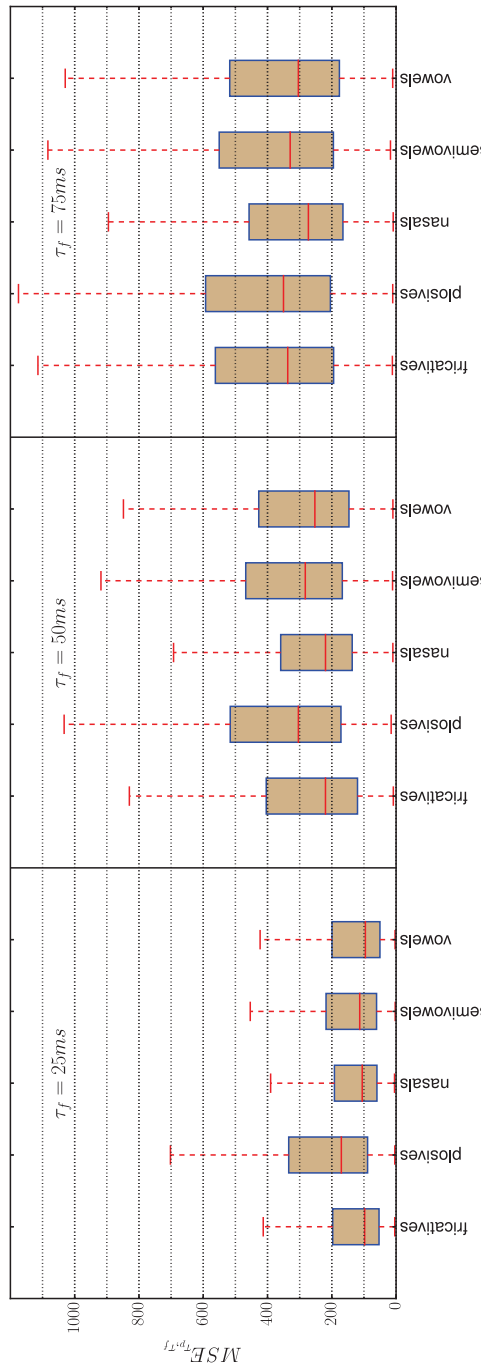

Figure 5: Prediction accuracy per class of speech sound. MSE for $\tau_f \in [25, 50, 75]$ ms and $\tau_p = 75$ ms; audio-only predictive model trained on the NTCD-TIMIT data set.

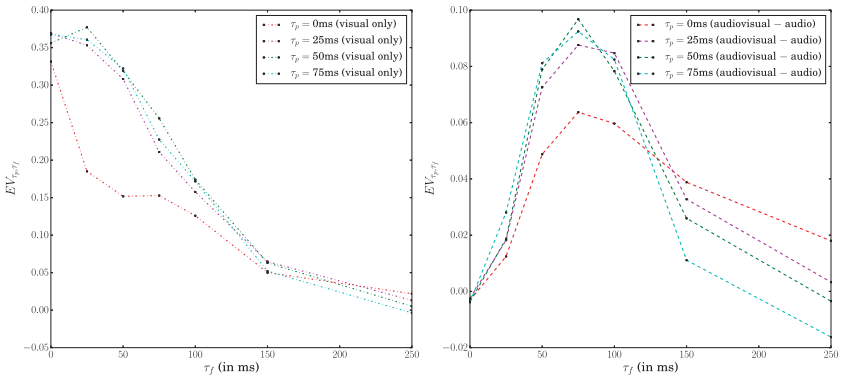

Figure 6: Performance of the visual-only and audiovisual models using NTCD TIMIT. Explained variance obtained when considering only the visual modality (left) and difference of explained variance between audio and audiovisual models (right).

$\tau_f = 150$ ms, $EV_{50,150} = 0.06$). These results may contribute to the debate in the neuroscience literature on the fact that the lip movements could be in advance on the sound because of anticipatory processes in speech production (see Chandrasekaran et al., 2009; Golumbic, Cogan, Schroeder, & Poeppel, 2013). The prediction of the spectral parameters from the lip information is maximal for the frame synchronous with the current input lip image (i.e., $\tau_f = 0$ ms)—or for the next frame ($\tau_f = 25$ ms) only for $\tau_p = 50$ ms—and then decreases smoothly with time. This is not in agreement with a stable advance of lips on sound.

As illustrated in Figure 4, combining audio with visual information improves the prediction over using audio only (compare solid lines with dashed lines). The gain is small but real, increasing the weighted explained variance by up to 0.1 depending on τ_f and τ_p. In order to better illustrate the dynamics of the gain brought by the visual input, we displayed in Figure 6 (right) the difference of weighted explained variance between audiovisual models and audio-only models. Importantly, results show a peak in the gain provided by the visual input for $\tau_f = 75$ ms. Therefore, even if there is no systematic lead of lips on sounds, there is a temporal window between 50 ms and 100 ms where the use of visual information is most helpful.

3.2.1 Qualitative evaluation. All of these quantitative results were averaged over many test sentences and speakers. Here, we finally discuss from a qualitative point of view the accuracy of the predicted spectral content at the utterance level. An example of a prediction error at $\tau_f = 100$ ms using either an audio-only or audiovisual predictive model is shown in Figure 7.

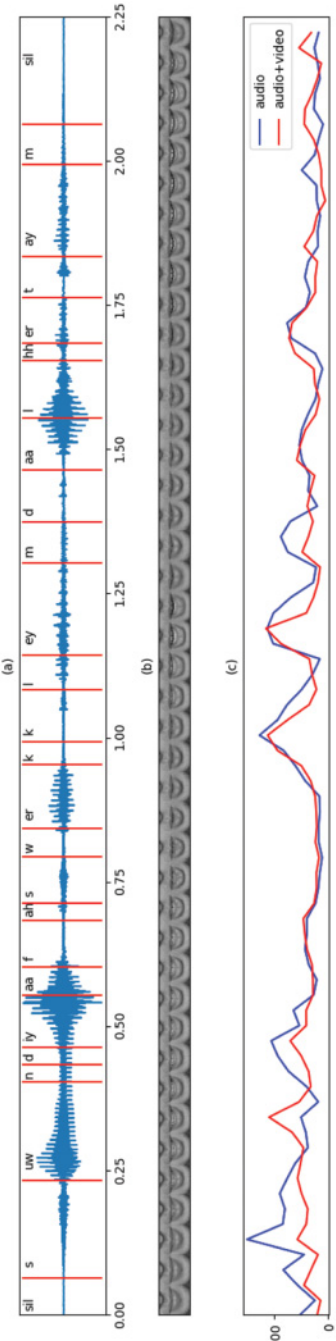

Figure 7: Example of prediction errors for $\tau_p = 75$ ms and $\tau_f = 100$ ms using an audio-only versus an audiovisual model (a) original audio waveform with phonetic segmentation (with TIMIT phonetic labels) for sentence *sil100* recorded by speaker *57M* of NTCD-TIMIT corpus, (b) corresponding lip image stream (downsampled to 15 fps to improve figure readability), (c) evolution of MSE_{τ_p,τ_f} over the utterance for both audio and audiovisual predictive models.

For the audio-only model (see the blue in plot c), peaks in prediction errors are mainly observed at either the vowel onset of consonant-vowel sequences (e.g., [d-iy], [l-(hh)-er]) or at the onset of the consonant of vowel-consonant sequence (e.g., [er-m], [er-t]) for which the precise initiation of the trajectory after a period of relative stability is hardly predictable. As concerns the audiovisual model, the average gain is accompanied by a large range of variations, leading to fluctuations between large gains and large losses provided by lip movements. A substantial gain from the visual input may occur when the speaker produces preparatory lip gestures before beginning to speak as in the [s] onset after the silence at the beginning of the utterance. Visible though poorly audible gestures as the closure for [m] in [l-ey-m] also lead to a visual gain in prediction. Another source of gain could be related to a coarticulation effect, when lips anticipate the upcoming vowel as during the first [d] in the utterance where the stretching gesture starts before the onset of the [iy]. Conversely, cases of error increase due to the visual input concern occurrences of nonvisible tongue gestures, such as intensity decrease in the vowel [uw] due to displacement of the tongue apex in the dental region in the following [nd] cluster around 0.4 s that is detected in the auditory stream but not in the visual stream.

3.3 A Database of Prediction Errors for Future Neurocognitive Experiments. A number of recent neurophysiological experiments have tested the existence and characteristics of predictive patterns in the audio and audiovisual responses to speech in the human brain (Van Wassenhove et al., 2005; Arnal, Morillon, Kell, & Giraud, 2009; Tavano & Scharinger, 2015; Ding, Melloni, Zhang, Tian, & Poeppel, 2016) and a recent theoretical review of predictive processes (Keller & Mrsic-Flogel, 2018). Importantly, these experiments lack a ground-truth basis on the natural predictive structure of audio and audiovisual speech, which can lead to misinterpretations or overgeneralizations of observed patterns (Schwartz & Savariaux, 2014). Our study could provide an interesting basis for future studies, providing a quantitative knowledge on the amount of "predictability" available in the physical signals considered in the simulations. The source code used to format the data and train and evaluate both audio and audiovisual predictive models on LibriSpeech and NTCD-TIMIT data sets, as well as all simulation results, has been made publicly available on https://github.com/thueber/DeepPredSpeech (for source code) and https://zenodo.org/record/3528068 (for data, simulation results on NTCD-TIMIT and pretrained models, doi:10.5281/zenodo.1487974). We believe that such results could be of interest in future neurophysiological experiments aiming at testing neural predictions in speech processing in the human brain.

4 Conclusion

In the general framework of predictive coding in the human brain, this study aimed at quantifying what is predictable online from the speech acoustic signal and the visual speech information (mostly lip movements). We proposed a set of computational models based on artificial (deep) neural networks that were trained to predict future audio observations from past audio or audiovisual observations. Model training and evaluation were performed on two large and complementary multispeaker data sets, respectively for audio and audiovisual signals, both publicly available. The key results of our study are these:

- It is possible to predict the spectral information in the acoustic speech signal in a temporal range of about 250 ms. At 25 ms, prediction enables reducing the power of the signal to transmit by neural processes (i.e., the error signal instead of the input signal) by a factor up to four. But the accuracy of the prediction decreases rapidly with future time lag (e.g., with average prediction gain obtained on the larger of our two tested data sets around 2 (EV ≈ 0.5) at 50 ms (with $\tau_p = 75$ ms), 1.6 at 75 ms (EV $= 0.37$) and almost 1 (EV $= 0.05$), i.e., no gain, at around 350 ms).
- The information provided by the visual modality is real but limited. The prediction accuracy of the predictive model based on visual-only information does not evidence a stable advance of lips on sound (as sometimes stated in the literature). The maximum average gain provided by the visual input in addition to the audio input is about $+0.1$ of explained variance and is obtained for a prediction at 75 ms.
- Best prediction accuracy is obtained when considering 50 to 75 ms of past context, for both audio and audiovisual models.
- Plosives are more difficult to predict than other types of speech sound.

This study hence provides a set of quantitative evaluations of auditory and audiovisual predictions at the phonetic level, likely to be exploited in predictive coding models of speech processing in the human auditory system. These evaluations are based on a specific class of statistical models based on deep learning techniques. Of course, it can be envisioned that the number of regularities in the speech signal might be actually larger than what has been captured by deep learning techniques in this study. Still, the large amount of data exploited here and the acknowledged efficacy of deep learning techniques make us confident that the estimations provided in this work constitute a reasonable estimation of the order of magnitude of possible regularities captured by statistical models.

As we stated in section 1, predictive coding should operate at a number of higher stages in speech neurocognitive processing, related to lexical,

syntactic, and semantic/pragmatic levels exploiting wider temporal scales. Our study should hence be considered as just a first stage in the analysis of predictive coding in speech processing. It provides a baseline along which further studies on higher-level predictive stages can be evaluated quantitatively, comparing the number of additional predictions that can occur from linguistic models to this audiovisual phonetic reference. Future work will focus on the integration of this linguistic level in a more complete neurocognitive architecture for speech predictive coding.

Acknowledgments

This work has been supported by the European Research Council under the European Community Seventh Framework Programme (FP7/2007-2013 grant agreement 339152, Speech Unit(e)s).

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545.

Abdelaziz, A. H. (2017). NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition. In *Proc. Interspeech* (pp. 3752–3756). International Speech Communication Association.

Altmann, E. G., Cristadoro, G., & Degli Esposti, M. (2012). On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, *109*(29), 11582–11587.

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, *29*(43), 13445–13453.

Atal, B. (1983). Efficient coding of LPC parameters by temporal decomposition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (vol. 8, pp. 81–84). Piscataway, NJ: IEEE.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2012). Spatiotemporal convolutional sparse auto-encoder for sequence classification. In *Proceedings of the British Machine Vision Conference*. Durham, UK: British Machine Vision Association.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Ben Ali, F., Djaziri-Larbi, S., & Girin, L. (2016). Low bit-rate speech codec based on a long-term harmonic plus noise model. *Journal of the Audio Engineering Society*, *64*(11), 844–857.

Berent, I. (2013). The phonological mind. *Trends in Cognitive Sciences*, *17*(7), 319–327.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234.

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., & Yvert, B. (2016). Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLoS Computational Biology*, *12*(11), e1005119.

Bradski, G. (2000). The OpenCV Library: Dr. Dobb's journal of software tools. https://opencv.org

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436.

Chollet, F. et al. (2015). *Keras*. https://github.com/fchollet/keras

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060–1089.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158.

Dusan, S., Flanagan, J. L., Karve, A., & Balaraman, M. (2007). Speech compression by polynomial approximation. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(2), 387–395.

Ebeling, W., & Neiman, A. (1995). Long-range correlations between letters and sentences in texts. *Physica A: Statistical Mechanics and Its Applications*, *215*(3), 233–241.

Farvardin, N., & Laroia, R. (1989). Efficient encoding of speech LSP parameters using the discrete cosine transformation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (vol. 1, pp. 168–171). Piscataway, NJ: IEEE.

Friederici, A. D., & Singer, W. (2015). Grounding language processing on basic neurophysiological principles. *Trends in Cognitive Sciences*, *19*(6), 329–338.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127.

Friston, K. (2011). What is optimal about motor control? *Neuron*, *72*(3), 488–498.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS One*, *4*(7), e6421.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology–Paris*, *100*(1–3), 70–87.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615–621.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., & Zue, V. (1993). *TIMIT acoustic phonetic continuous speech corpus ldc93s1*. Web Download. Philadelphia: Linguistic Data Consortium. https://catalog.ldc.vpenn.edu/LDC93s1w

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Amsterdam: Kluwer.

Giraud, A.-L., & Poeppel, D. (2012). Speech perception from a neurophysiological perspective. In D. Poeppel, T. Overath, A. Popper, & R. Fay (Eds.), *The human auditory cortex* (pp. 225–260). New York: Springer.

Girin, L. (2004). Joint matrix quantization of face parameters and LPC coefficients for low bit rate audiovisual speech coding. *IEEE Transactions on Speech and Audio Processing*, *12*(3), 265–276.

Girin, L. (2010). Adaptive long-term coding of LSF parameters trajectories for large-delay/very-to ultra-low bit-rate speech coding. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*(1), 597039.

Girin, L., Firouzmand, M., & Marchand, S. (2007). Perceptual long-term variable-rate sinusoidal modeling of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(3), 851–861.

Girin, L., Schwartz, J.-L., & Feng, G. (2001). Audio-visual enhancement of speech in noise. *Journal of the Acoustical Society of America*, *109*(6), 3007–3020.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *Journal of Neuroscience*, *33*(4), 1417–1426.

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Heinz, J., & Idsardi, W. (2011). Sentence and word complexity. *Science*, *333*(6040), 295–297.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393.

Ioffe, S., & Szegedy, C. (2015). *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. arXiv:1502.03167v3.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. New York: Oxford University Press.

Jayant, N. S., & Noll, P. (1984). *Digital coding of waveforms: Principles and applications to speech and video*. Englewood Cliffs, NJ: Prentice Hall.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 221–231.

Kaplan, R. M., & Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics*, *20*(3), 331–378.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L., (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1725–1732). Piscataway, NJ: IEEE.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., . . . Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proceedings of the International Symposium on Music Information Retrieval* (pp. 255–266). International Society for Music Information Retrieval.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv:1412.6980v9.

Kleijn, W. B., & Ozerov, A. (2007). Rate distribution between model and signal. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 243–246). Piscataway, NJ: IEEE.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Li, W. (1990). Mutual information functions versus correlation functions. *Journal of Statistical Physics*, *60*(5–6), 823–837.

Li, W., & Kaneko, K. (1992). Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. *Europhysics Letters*, *17*(7), 655.

Lin, H., & Tegmark, M. (2017). Critical behavior in physics and probabilistic formal languages. *Entropy*, *19*(7), 299.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Markel, J. D., & Gray, A. J. (1976). *Linear prediction of speech*. New York: Springer.

McFee, B., McVicar, M., Balke, S., Thomé, C., Lostanlen, V., Raffel, C., . . . Holovaty, A. (2018). *Librosa toolkit*. 10.5281/zenodo.1342708.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.

Montemurro, M. A., & Pury, P. A. (2002). Long-range fractal correlations in literary corpora. *Fractals*, *10*(4), 451–461.

Mroueh, Y., Marcheret, E., & Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing* (pp. 2130–2134). Piscataway, NJ: IEEE.

Mudugamuwa, D. J., & Bradley, A. B. (1998). Optimal transform for segmented parametric speech coding. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing* (vol. 1, pp. 53–56). Piscataway, NJ: IEEE.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2014). Lipreading using convolutional neural network. In *Proc. Interspeech* (pp. 1149–1153). International Speech Communication Association.

Oberlander, J., & Brew, C. (2000). Stochastic text generation. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, *358*(1769), 1373–1387.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing* (pp. 5206–5210). Piscataway, NJ: IEEE.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, *91*(9), 1306–1326.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rao, R. R., & Chen, T. (1996). Cross-modal predictive coding for talking head sequences. In *Proc. of the International Conference on Acoustics, Speech, and Signal processing* (pp. 2058–2061). Piscataway, NJ: IEEE.

Rivet, B., Girin, L., & Jutten, C. (2007). Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(1), 96–108.

Samuelsson, J., & Hedelin, P. (2001). Recursive coding of spectrum parameters. *IEEE Transactions on Speech and Audio Processing*, *9*(5), 492–503.

Sánchez-García, C., Alsius, A., Enns, J. T., & Soto-Faraco, S. (2011). Cross-modal prediction in speech perception. *PLOS One*, *6*(10), e25198.

Schultz, T., Wand, M., Hueber, T., Krusienski, D. J., Herff, C., & Brumberg, J. S. (2017). Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2257–2271.

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 85–108). Hove, UK: Psychology Press.

Schwartz, J.-L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLOS Computational Biology*, *10*(7), e1003743.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27* (pp. 568–576). Red Hook, NY: Curran.

Subasingha, S., Murthi, M. N., & Andersen, S. V. (2009). Gaussian mixture Kalman predictive coding of line spectral frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(2), 379–391.

Subramaniam, A. D., Gardner, W. R., & Rao, B. D. (2006). Low-complexity source coding using gaussian mixture models, lattice vector quantization, and recursive coding with application to speech spectrum quantization. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 524–532.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. In *Proc. of the Conference on Computer Vision and Pattern Recognition* (pp. 1–9). Piscataway, NJ: IEEE.

Tatulli, E., & Hueber, T. (2017). Feature extraction using multimodal convolutional neural networks for visual speech recognition. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing* (pp. 2971–2975). Piscataway, NJ: IEEE.

Tavano, A., & Scharinger, M. (2015). Prediction in speech and language processing. *Cortex*, *68*, 1–7.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186.

Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., & Hickok, G. (2016). Timing in audiovisual speech perception: A mini review and new psychophysical data. *Attention, Perception, and Psychophysics*, *78*(2), 583–601.

Wand, M., Koutník, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing* (pp. 6115–6119). Piscataway, NJ: IEEE.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726.

Yong, M., Davidson, G., & Gersho, A. (1988). Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction (speech coding). In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing* (pp. 402–405). Piscataway, NJ: IEEE.

Received June 6, 2019; accepted November 7, 2019.