Data and simulations about audiovisual asynchrony and predictability in speech perception

Jean-Luc Schwartz, Christophe Savariaux

Data and simulations about audiovisual asynchrony and predictability in speech perception

Schwartz Jean-Luc[1], *Savariaux Christophe*[1]

[1]GIPSA-Lab, Speech and Cognition Department, UMR 5216 CNRS Grenoble-Alps University, France

jean-luc.schwartz@gipsa-lab.grenoble-inp.fr, christophe.savariaux@gipsa-lab.grenoble-inp.fr

Abstract

Since a paper by Chandrasekaran et al. (2009), an increasing number of neuroscience papers capitalize on the assumption that visual speech would be typically 150 ms ahead of auditory speech. It happens that the estimation of audiovisual asynchrony by Chandrasekaran et al. is valid only in very specific cases, for isolated CV syllables or at the beginning of a speech utterance. We present simple audiovisual data on plosive-vowel syllables (pa, ta, ka, ba, da, ga, ma, na) showing that audiovisual synchrony is actually rather precise when syllables are chained in sequences, as they are typically in most parts of a natural speech utterance. Then we discuss on the way the natural coordination between sound and image (combining cases of lead and lag of the visual input) is reflected in the so-called temporal integration window for audiovisual speech perception (van Wassenhove et al., 2007). We conclude by a computational proposal about predictive coding in such sequences, showing that the visual input may actually provide and enhance predictions even if it is quite synchronous with the auditory input.

Index Terms: audiovisual asynchrony, temporal integration window, predictive coding, visual lead/lag, visual prediction

1. Introduction

Since a paper by Chandrasekaran et al. (2009), an increasing number of neuroscience papers capitalize on the assumption that visual speech would be typically 150 ms ahead of auditory speech. Let us mention a few quotations from these papers: “In most ecological settings, auditory input lags visual input, i.e., mouth movements and speech associated gestures, by ~150 ms” (Arnal et al., 2009; see also Arnal et al., 2011); “there is a typical visual to auditory lag of 150–200 ms in face-to-face communication (Musacchia & Schroeder, 2009); “articulatory facial movements are also correlated with the speech envelope and precede it by ~150 ms” (Zion-Golumbic et al., 2013).

The invoked natural audiovisual asynchrony is used in these papers in support to development on models and experiments assessing the so-called “predictive coding theory”. This theory posits that neural processing exploits a differential coding between predicted and incoming signals, with decreased activity when a signal is correctly predicted (Friston, 2005).

The assumption that image leads sound plays two different roles in the above mentioned neuroscience papers. It is sometimes used as a trick to demonstrate that the visual stimulus plays a role in modulating the neural auditory response, rightly capitalizing on a situation where a CV sequence (e.g. “pa” or “ta”) is produced after a pause. In this case, the preparatory movement of the mouth and lips is visible before any sound is produced, hence visual prediction can occur ahead of sound and actually modulates the auditory evoked potential measured by EEG or MEG (with a decrease in amplitude and delay of the first negative peak N1 around 100ms after the acoustic onset, Besle et al., 2004; van Wassenhove et al., 2005; Arnal et al., 2009).

The second role is more problematic. Considering that there would be a systematic and more or less stable advance of vision on audition around 150 ms, it is proposed that this situation would play a role in the ability to use the visual input to predict the auditory one all along the time. Audiovisual asynchrony is implicitly incorporated in a number of models and proposals.

However, the situation studied by Chandrasekaran et al. to propose that vision is ahead on audition is very specific, characteristic of a CV sequence produced in isolation or at the beginning of an utterance after a pause. In the remaining of this paper we present in Section 2 simple audiovisual data on plosive-vowel syllables (pa, ta, ka, ba, da, ga, ma, na) showing that audiovisual synchrony is actually almost perfect when syllables are chained in sequences, as they are typically in most parts of a natural speech utterance.

Then we discuss in Section 3 how natural coordination between sound and image (combining cases of lead and lag of the visual input) results in the so-called temporal integration window for audiovisual speech perception (van Wassenhove et al., 2007).

We present in Section 4 a computational proposal about predictive coding in such sequences, showing that the visual input may actually provide and enhance predictions even if it is quite synchronous with the auditory input. This is to show that the “visual lead” hypothesis, wrong in many cases, is actually not necessary to deal with audiovisual predictability. We propose various variants of such auditory or audiovisual prediction models. We discuss their properties in relation with experimental data showing that listeners do exploit audiovisual coherence properties in speech processing. This provides the basis for our final conclusion in Section 5.

2. Audiovisual synchrony vs. asynchrony in plosive-vowel sequences /Ca/

2.1. Distinguishing closing and opening events

In this section we focus on audiovisual temporal relationships in CV sequences where C is a voiced, unvoiced or nasal stop consonant that is, for English or French (the two languages considered in the paper by Chandrasekaran et al., 2009), one of the sounds /p t k b d g m n/, and V is the open vowel /a/. We shall consider more general phonetic material in the next section.

Consider what happens in an isolated /pa/ syllable. First you have to close your lips to prepare the “p”. This involves a visible

gesture described by Chandrasekaran and coll. by two temporal events, the initiation of the closing gesture, and the velocity peak of the lips during the closure phase. Then comes the release, which corresponds to a third visible event (not discussed by the authors) and to the first auditory event that is the acoustic burst for the plosive. Of course, the first visible event (closure gesture initiation) and the first auditory event (opening gesture initiation) are asynchronous, since closure must occur before opening! The temporal distance may reach 150 ms or even more: actually you can close any time before you open (imagine you want to stop your interlocutor by uttering "please", you prepare the "p" but don't succeed to interrupt him or her: you will stay with your lips closed for a while, and the temporal delay between visible lip closing and audible burst may reach very large values).

But in the largest part of the speech ecological material, syllables are chained and hence plosives are very often embedded between vowels. Consider the case of "apa". When you begin to close your lips, this is visible but it is also audible since it changes the formants and the intensity of the sound. At the end of the closing gesture the sound stops (or changes into intervocalic voicing in the case of "aba"). In such cases it is mistaken to characterize audiovisual coordination as the delay between closure gesture initiation for vision and opening gesture initiation for audition – though this is what Chandrasekaran et al. do in their Fig. 9 – because there is actually an audible and a visible event for both closure gesture and opening gesture initiation.

2.2. Audiovisual asynchronies in /aCa/ sequences

To show this more clearly, we have recorded a small database of 6 repetitions of the syllables /pa ta ka ba da ga ma na/ uttered by a French speaker either in isolation or in sequence, and we have labeled the corresponding auditory and visual events. The recording set up was based on the classical paradigm we use in Grenoble since years (Lallouache, 1990; Noiray et al., 2008) with blue make up applied in the lips, which enables automatic and precise detection of lip contours by applying a Chroma Key process extracting blue areas on the face, and hence allows precise positioning of visual events on the lip trajectories. The acoustic analysis was done on Praat (Boersma & Weenink, 2012). A typical display of the synchronized acoustic signal with its time-frequency analysis (including intensity and formants) and lip trajectory is presented on Fig. 1 for an isolated /pa/ vs. /pa/ embedded in a sequence (with a zoom on /apa/).

On such kinds of displays we have manually detected the corresponding events:

- on the acoustic signal and spectrogram, in the case of embedded sequences: the beginning of F1 decrease in the portion from the previous "a" to the next plosive (Closing onset for Audio Formant: CAF); the corresponding beginning of intensity decrease (Closing onset for Audio Intensity: CAI). And in all cases, for embedded as well as isolated sequences, the beginning of F1 increase in the portion from the plosive to the next "a" (Opening onset for Audio Formant: OAF) and the corresponding beginning of intensity increase, that is the burst onset (Opening onset for Audio Intensity: OAI).
- on the lip trajectory, in all cases: the beginning of lip area decrease in the portion from the previous "a" or from silence to the next plosive (Closing onset for Visible Lips: CVL) and the beginning of lip area increase at the plosive release towards the next vowel (Opening onset for Visible Lips: OVL).

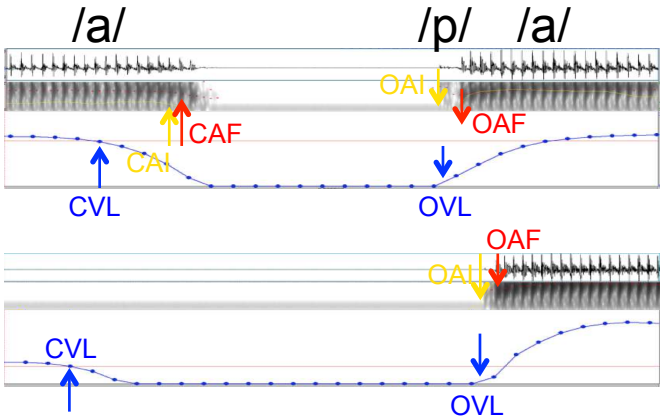

Figure 1: Acoustic signal, time-frequency analysis (intensity in yellow and formants in red) and lip trajectory in blue for /apa/ (top) and /pa/ (bottom). Blue arrows: lip events. Yellow arrows: intensity events. Red arrows: formant events. Up arrows: closing events. Down arrows: opening events.
CAF/OAF: Closing/Opening onset for Audio Formant
CAI/OAI: Closing/Opening onset for Audio Intensity
CVL/OVL: Closing/Opening onset for Visible Lips

We display in Fig. 2 the data about temporal coordination between audio and visual events for either closing (Fig. 2a) or opening (Fig. 2b) in the case of embedded sequences. The mean delay between visual and acoustic events at the closure (Fig. 2a) varies between -20ms and -40ms for intensity (CVL-CAI, in yellow) and reaches larger values from -40 to -80ms for formants (CVL-CAF, in red). This means that there is a small lead of the visual channel on the audio channel (where information is available on intensity before formants). But this lead is much smaller than the 150ms lead mentioned by Chandrasekaran et al. (2009), and there are actually cases where audio information arrives before video information, e.g. for /ad/ and /ag/ where the tongue gesture towards the voiced plosive decreases intensity while jaw may stay rather stable, and hence lip area does not decrease much – which prevents early video detection.

In the opening phase (Fig. 2b) the synchrony is even larger. Concentrating on the delay between labial and intensity events (OVL-OAI, in yellow) we actually observe an almost perfect synchrony for labials (/p b m/). This is trivial: as soon as the lips begin to open, the sound drastically changes, from silence (for /p/) or prevoicing (for /b/) or nasal murmur (for /m/) to the plosive burst. For velars /k g/ there is actually a clear lead of the audio channel, since the first tongue move producing the plosive release is done with no jaw movement at all and hence before any labial event is actually detectable: the audio lead may reach more than 20ms.

We display on Fig. 3 the data for isolated syllables. In this case, where there is no audible event for closure, we report the same measure as Chandrasekaran et al. (2009), that is the delay between the first visible event CVL and the first audible event, that is OAI or OAF. There is a very large anticipation, which actually reaches values much larger than 150 ms here (and which may reach 400ms in some cases).

In summary, while the "visual lead" proposal is true for isolated syllables with quantitative data similar to Chandrasekaran et al., for embedded sequences there is no big asynchrony, with

actually both cases where you can see what happens before you hear it (e.g. /aC/, whatever the consonant C) and reverse cases where you can hear before you see (e.g. /ka/ or /ga/).

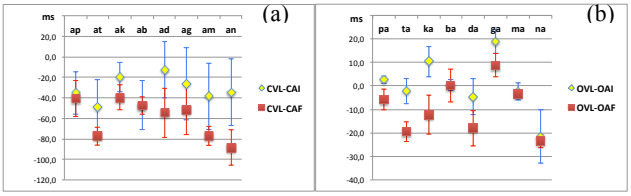

Figure 2: Delay between visual and auditory events: (a) in the closing phase (left), in /aC/ where C is a plosive in the set /p t k b d g m n/; (b) in the opening phase (right), in /Ca/ with the same plosives. In red: acoustic events for formants. In yellow: acoustic events for intensity. Signs point at mean values (over the 6 repetitions), and error bars correspond to the standard deviation.

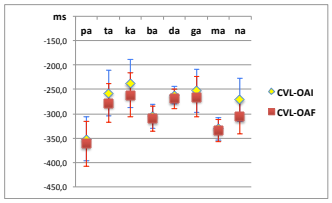

Figure 3: Delay between the first visual event (for the closing phase) and the first auditory event (for the opening phase) in isolated /Ca/. Same display as in Fig. 2.

3. Range of possible AV asynchronies and temporal integration window

Of course, speech utterances involve a range of phonetic configurations much larger than the /Ca/ sequences that were studied in Section 2. This variety of configurations leads to a variety of situations in terms of audiovisual asynchronies.

A first general property of speech concerns anticipatory coarticulation – much more relevant and general than preparatory movements discussed by Chandrasekaran et al. (2009). This relates to articulatory gestures towards a given phonetic target, which can begin within a previous target. Anticipatory coarticulation generally capitalizes on a property of the articulatory-to-acoustic transform, in which an articulatory gesture has sometimes no or weak effect on the sound and hence can be prepared in advance without audible consequences.

A typical example concerns the rounding gesture from /i/ to /y/ or /u/ in sequences such as /i$C_1C_2…C_n$y/ or /i$C_1C_2…C_n$u/ with a variable number of consonants $C_1…C_n$ not involving a specific labial control (e.g. /s t k r/) between the unrounded /i/ and the rounded /y/ or /u/. In this case the rounding gesture from /i/ towards /y/ or /u/ can begin within the sequence of consonants /$C_1C_2…C_n$/, and hence anticipate the vowel by 100 to 300 ms (Abry & Lallouache, 1996). Various sets of data and various theoretical models of this anticipatory coarticulation process have been proposed in the literature (e.g. Henke, 1966; Benguerel & Cowan, 1974, Bell-Berti & Harris, 1982; Perkell & Chiang, 1986; Perkell & Matthies, 1992; Abry & Lallouache, 1995, Abry et al., 1996, Roy et al., 2003). In such cases the rounding gesture can hence be visible well before it is audible.

So there are cases where vision leads audition (e.g. in /i$C_1…C_n$u/ sequences), others where vision and audition are quite synchronous (e.g. in /aCa/ sequences), and there are also cases where audition may actually lead vision. This was demonstrated by Troille et al. (2010) in sequences such as /izy/ in French, where the rounding gesture from /i/ to /y/ can be heard within the intervocalic fricative /z/, but is visually processed later because of the difficulty to disentangle what is due to the vowel gesture and what is due to the consonant. Troille et al. performed gating experiments on auditory, visual and audiovisual stimuli and displayed a lead of 40 ms of audition on vision.

In summary, there are actually a variety of situations from audio lead (estimated to 40 ms in Troille et al., 2010) to visual lead (which can reach more than 200 ms). In their study of mutual information between audio and video parameters on speech sequences, Feldhoffer et al. (2007) show that mutual information is maximal for some audio and video parameters when it incorporates a video lead up to 100ms. Czap (2011) obtains a smaller value (20 ms video lead) in audiovisual speech recognition experiments, recognition scores being higher with a small global video lead.

These global estimations are concordant with the classical view that "in average, the visual stream may lead the auditory stream", which is generally advocated by specialists of audiovisual speech perception (e.g. Bernstein et al., 2004; Kim & Davis, 2004). This has been conceptualized by the "audiovisual temporal integration window" (Munhall et al., 1996; Massaro et al., 1996; van Wassenhove et al., 2007), over which both simultaneity perceptual judgments and audiovisual fusion assessed by the McGurk effect seem to stay at their maximal value, and which happens to cover an asymmetric range between about 30 ms audio lead and 170 ms audio lag.

The temporal integration window is consistent with the view that there is actually a *range of possible asynchronies* (typically from 50 ms audio lead to 200 ms video lead) in natural speech, and that the perceptual system has internalized this range through a learning process. By contrast, the systematic "vision in advance to audition" stance is much too restricted and simplified to cover the ground truth of speech material. Brungart et al. (2008) actually showed that there does not seem to exist a clear relationship between the optimal delay for audiovisual fusion and the speech rate. This is more in line with a statistical relationship internalized in a perceptual process, than with a coordination driven by speech production with a visual lead that would quite likely change with speech rate.

4. AV predictability without AV asynchrony

4.1. Objective

Therefore there is NOT a video lead on the audio stream that would be stable around 150ms, and that would make video prediction on the auditory input easy and straightforward. There is rather a range of configurations with either audio lead, audiovisual synchrony or audio lag and this range results in the audiovisual temporal integration window. This rules out over-simplistic claims about audiovisual predictability. Does it raise a problem for predictability in general? The answer is clearly NO. The reason is that predictability does not require asynchrony. Actually, a pure auditory trajectory may provide predictions on its future stages, and the visual input may enhance these

predictions, since it is naturally in advance on *future* auditory events, though not systematically in advance on *present* ones.

In the remaining of this section, we shall propose a possible quantification of prediction abilities in the same kind of trajectories as the ones studied in Section 2. We shall show that there is a potential for pure auditory prediction in these trajectories, and that the visual input may significantly enhance these predictions in spite of the fact that the basic temporal events are close in time according to Section 2.

4.2. Methodology

The study is based on a corpus of 100 repetitions of sequences /aba/, /ada/ and /aga/ uttered at various rhythms in a spontaneous way within sequences such as "abadagabadagabadaga…", by a male French speaker. The recording and analysis setup is the same as the one described in Section 2. The material is processed in the following way:

- from the acoustic signal, spectrogram, formants and intensity are computed thanks to the tools available in Praat;
- a threshold at 50 dB is applied on the intensity signal to isolate /C_1aC_2/ sequences (the closure parts containing pure consonantal voicing being at an intensity below the threshold); /aC/ items are defined by taking the second half of these sequences in time;
- the values of formants F2 and F3 in the corresponding periods of time are extracted and temporally rescaled so that each /aC/ trajectory is described by a temporal sequence with 20 points, that is F2(1:20) and F3(1:20);
- the values of lip aperture L are automatically extracted thanks to the Chroma Key system; the same temporal extraction and rescaling process is applied, hence the lip trajectory for each /aC/ utterance is described by a temporal sequence with 20 points, that is L(1:20).

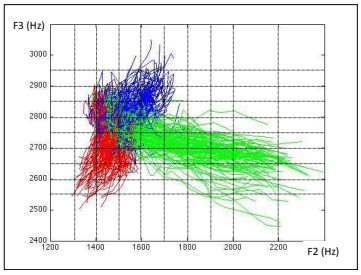

Figure 4: Trajectories of /ab/, /ad/, /ag/ in the F2-F3 plane

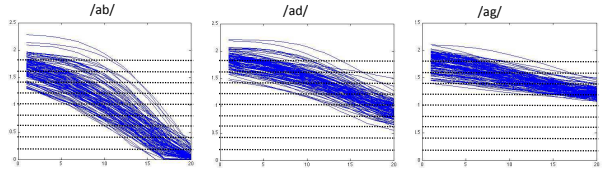

Figure 5: Variations of lip aperture for /ab/, /ad/, /ag/

The corresponding trajectories in the F2-F3 plane are depicted on Fig. 4, with classical shapes: joint F2 and F3 increase for /ad/, joint F2-F3 decrease for /ab/, F2 increase with F3 decrease producing an F2-F3 convergence for /ag/ (Schwartz et al., 2012). Variations of lip aperture L with time are displayed in Fig. 5, showing that the final L value is around 0 for /ab/ (though often not already at 0, since trajectories are stopped at the time when the intensity threshold is reached, which can happen before complete lip closure); the final value is higher for /ad/ (around 1 cm) and even higher for /ag/ (around 1.2 cm).

From these trajectories, a predictive coding model could attempt to provide guesses about the final point of the acoustic trajectory from a given point of the trajectory. We implemented such a model within a Bayesian probabilistic framework (Bessière et al., 2008). For this aim, we first discretized the (F2, F3) space into 100 values with 10 values for F2 and 10 for F3 sampling the acoustic space in Hz (see Fig. 4). We also discretized L with 10 values regularly sampling the lip opening space in cm (see Fig. 5). Then, from the 20 points of the 100 repetitions of the 3 types of stimuli (6000 points altogether) we learnt a joint audio probability function $p_A(C, C_{final})$. This is the probability to be in position C (from 1 to 100) at a given time and then in position C_{final} at the end of the trajectory. We also learnt an audiovisual probability function $p_{AV}(C, L, C_{final})$ that is the probability to be in position (C, L) at a given time and then in position C_{final} at the end of the trajectory.

The next step is the prediction model per se. We constructed two audio prediction models:

- Model A_{free}, for which the prediction about final position when the system is in C is provided by:

$C_{pred/Afree} = E(C_{final} / C) \quad \alpha \quad \Sigma_{Cfinal} C_{final} p_A(C, C_{final})$

(notice that in fact all the estimations are done independently on F2 and F3, that is mean positions for F2 and for F3 are first estimated, which enables to compute $C_{pred/Afree}$)

- Model $A_{entropy}$, for which the previous prediction is used under the constraint that entropy of the distribution of possible C_{final} positions stays at a low value (with a criterion using the sum of variances of possible F2 and F3 values of C_{final} positions). This means that if in a given position the possible C_{final} positions are too dispersed (as it is the case for example at the beginning of a trajectory around /a/) the $C_{pred/Afree}$ is not used and the true prediction stays at the position of the actual position C.

$C_{pred/Aentropy} = w\ C_{pred/Afree} + (1\text{-}w)\ C$ with w=f(entropy)

so that if entropy is low, $C_{pred/Aentropy}$ is close to $C_{pred/Afree}$, and if it is high, $C_{pred/Aentropy}$ is close to C.

Audiovisual prediction models are introduced in the same way:

- Model AV_{free}, for which the prediction about final position when the system is in C is provided by:

$C_{pred/AVfree} = E(C_{final} / C, L) \quad \alpha \quad \Sigma_{Cfinal} C_{final} p_{AV}(C, L, C_{final})$

Model $AV_{entropy}$, with:

$C_{pred/AVentropy} = w\ C_{pred/AVfree} + (1\text{-}w)\ C$ with w=f(entropy)

4.3. Evaluation of the simulation results

To evaluate these prediction models we used two criteria. The first one estimates the "size" of prediction. It is based on the Euclidian distance (in the discredited F2 and F3 values) between the actual position C and the predicted position C_{pred}, $d(C, C_{pred})$, for each of the four prediction models. For a given normalized time from 1 to 20, we estimate the mean of these distances for the 300 possible trajectories (100 for each consonant). This provides the $C_{size}(t)$ criterion for each prediction model: the larger $\mathcal{C}_{size}$ the "larger" the prediction at a given position of the audio or audiovisual trajectory.

The second criterion estimates the "efficiency" of the prediction. It is based on the difference between the distance between actual

position and true final position, and the distance between predicted position and true final position:
$d(C, C_{final}) - d(C_{pred}, C_{final})$

A positive difference expresses the fact that prediction is closer to final position than is actual position: the prediction is "efficient" in this case, and not if the difference is negative. $\mathscr{C}_{efficiency}(t)$ is computed for each normalized time between 1 and 20 as the mean of this difference for all trajectories.

The variations of $\mathscr{C}_{size}(t)$ are displayed for the four prediction models on Fig. 6. We observe that $\mathscr{C}_{size}$ is larger for "free" than for "entropy" models, which is expected since the entropy criterion precisely aims at decreasing prediction when it is unreliable. Actually, the major difference is for t values close to 0, at the beginning of the trajectory, when it is impossible to predict anything. $\mathscr{C}_{size}$ values for "entropy" models are actually small both at the beginning of the trajectory where predictions are unreliable and at the end where there is nothing more to predict. They reach a maximum in the second half of the trajectory. Audiovisual predictions produce larger $\mathscr{C}_{size}$ values than auditory prediction, particularly for "entropy" models: hence the visual component improves predictions.

On Figure 7 we display $\mathscr{C}_{efficiency}$ for the 4 prediction models. Here we draw both mean values (in solid lines) and minimum and maximum values (in dotted lines). Once again, "free" predictions produce larger mean values than "entropy" predictions but with a number of negative values, which show that predictions can be wrong, particularly at the beginning of the trajectory. On the contrary, "entropy" models almost never produce wrong predictions: most values are above zero. And for this criterion also, audiovisual predictions are much more efficient, particularly for "entropy" models.

4.4. Discussion

In summary, these figures and simulations show (1) how predictions can be made, (2) how their efficiency can be controlled thanks to an entropy-based criterion, (3) how they evolve in time in these configurations and (4) most importantly, that the visual input may strongly improve predictions, in spite of the close synchrony of basic temporal events in the auditory and visual streams, according to Section 2.

It is actually known since long that the auditory and video streams are related by a high level of cross-predictability as displayed by a number of studies about audio-visual correlations between various kinds of video (e.g. lip parameters, facial flesh points, DCT video features extracted from the face) and audio (acoustic envelope, band-pass filter outputs, LPC or LSP features) parameters: see e.g. Yehia et al., 1998; Barker & Berthommier, 1999; Grant & Seitz, 2000; Jiang et al., 2002; Berthommier, 2004; Chandrasekaran et al., 2009). Furthermore, a number of gating experiments (by e.g. Smeele, 1994; Munhall & Tohkura, 1998; de la Vaux & Massaro, 2004; Jesse & Massaro, 2010) suggest that the visual information about a given speech utterance, as a syllable or a by-syllabic word, is often present earlier than the acoustic information.

This confirms that there is actually a large amount of "visual predictability" present in the audiovisual input, even though the temporal relationship between audition and vision is less clear than proposed by Chandrasekaran et al. (2009). This predictability enables the perceptual system to improve speech detection in noise (Grant & Seitz, 2000; Kim & Davis, 2004) and to enhance speech intelligibility (Schwartz et al., 2004). It is the basis of "audiovisual binding mechanisms" that have been described recently in speech perception (Nahorna et al., 2012).

Of course, the prediction models presented in this paper are preliminary and over simplistic, including speech sequences very stereotyped, with only three variants for a single speaker. More powerful techniques will have to be explored to attempt to rescale this study towards more realistic corpora. The use of entropic predictions, as introduced in this paper, will be an important challenge for future studies in this framework.

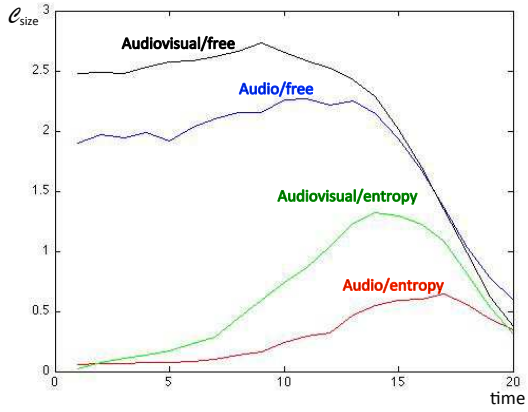

Figure 6: Variations of $\mathscr{C}_{size}$ for the 4 prediction models

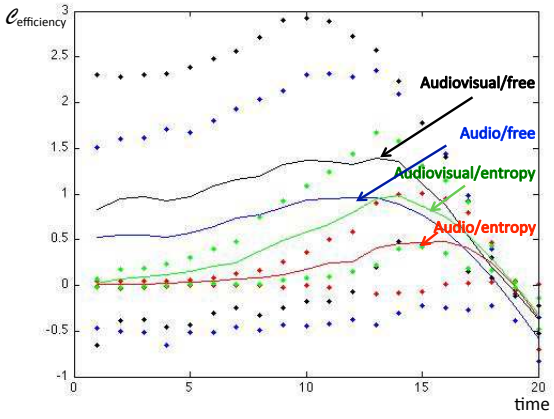

Figure 7: Variations of $\mathscr{C}_{efficiency}$ for the 4 prediction models Mean values in solid lines, maximum and minimum values in dotted lines, for each prediction model (see text).

5. Conclusion

This paper had two major objectives. Firstly, we made clear that the view that vision leads audition is oversimplified and often wrong. It should be replaced by the acknowledgement that the temporal relationship between auditory and visual cues is complex, including a range of configurations more or less reflected by the temporal integration window from 30 to 50 ms auditory lead to 170 to 200 ms visual lead.

Secondly, we showed that even if the visual input is not ahead of the auditory input, it may provide gains in predictability. For this

aim, we proposed prediction models in a Bayesian framework, including entropy constraints that appear to make predictions reliable and useful. Such quantitative proposals for predictive coding in speech perception do not exist yet, to our knowledge. We hope that the present work will provide an impulse towards realistic computational proposals for assessing auditory and audiovisual prediction coding models in speech perception.

6. References

Abry, C., & Lallouache, T. M. (1995). Modeling lip constriction anticipatory behaviour for rounding in French with the MEM. Proc. ICPhS'95, 152–155.

Abry C., & Lallouache T. (1996). Le MEM: Un modèle d'anticipation paramétrable par locuteur. Données sur l'arrondissement en français (MEM: A speaker parameterized anticipation model). Bul. de la Comm. Parlée 3, 85–99.

Abry C., Lallouache M. T., & Cathiard M. A. (1996). How can coarticulation models account for speech sensitivity to audio-visual desynchronization? In Stork D. and Hennecke M. (Eds.) Speechreading by Humans and Machines, NATO ASI Series F (vol. 150, pp. 247–255). Berlin: Springer-Verlag.

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, 13445-53.

Arnal, L.H., Wyart, V., & Giraud, A.L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature Neuroscience, doi:10.1038/nn.2810.

Barker, J. P., & Berthommier, F. (1999). Evidence of correlation between acoustic and visual features of speech. In Proceedings ICPhS '99 (pp. 199-202.) San Francisco: USA.

Bell-Berti, F., & Harris, K. S. (1982). Temporal patterns of coarticulation: Lip rounding. J. Acoust. Soc. Am. 71, 449–459.

Benguerel, A. P., & Cowan, H. A. (1974). Coarticulation of upper lip protrusion in French. Phonetica 30, 41–55.

Bernstein, L. E., Takayanagi, S., & Auer, E. T., Jr. (2004). Auditory speech detection in noise enhanced by lipreading. Speech Communication 44, 5-18.

Berthommier, F. (2004). A phonetically neutral model of the low-level audiovisual interaction. Speech Communication 44, 31-41.

Besle, J., Fort, A., Delpuech, C. & Giard, M-H. (2004). Bimodal Speech: Early Visual Effect in the Human Auditory Cortex. European Journal of Neuroscience, 20, 2225-2234.

Bessière, P., Laugier, C., & Siegwart, R. (Eds.) (2008). Probabilistic Reasoning and Decision Making in Sensory-Motor Systems. Springer Tracts in Advanced Robotics. Berlin: Springer-Verlag Series, vol. 46.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer (Version 5.3.04) [Computer program]. Retrieved May 2012, from http://www.praat.org.

Brungart, D.S., Iyer, N., Simpson, B.D., & van Wassenhove, V. (2008). The effects of temporal asynchrony on the intelligibility of accelerated speech. Proc. AVSP' 2008, 19-24.

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

Czap, László (2011). On the audiovisual asynchrony of speech, Proc AVSP'2011, 137-140.

Feldhoffer, G., Bárdi, T., Takács, Gy. & Tihanyi, A. (2007). Temporal asymmetry in relations of acoustic and visual features of speech. Proc. 15th European Signal Processing Conf., Poznan.

Friston, K. (2005). A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci, 360, 815-836.

Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. 108, 1197–1208.

Henke, W.L. (1966). Dynamic articulatory model of speech production using computer simulation. Dissertation Massachusetts Institute of Technology,

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. Attention, Perception & Psychophysics 72, 209-225.

Jiang, J., Alwan, A., Keating, P.A., Auer, E.T. Jr, & Bernstein, L.E. (2002). On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics. Eurasip Journal on Advances in Signal Processing 11, 1174-1188.

Kim, J., & Davis, C. (2004). Investigating the audio-visual speeh detection advantage. Speech Commun, 44, 19–30.

Lallouache, M.T. (1990). Un poste 'visage-parole'. Acquisition et traitement de contours labiaux. Proc. XVIII Journées d'Études sur la Parole, 282-286.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. J. Acoust. Soc. Am. 100, 1777–1786.

Munhall, K., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. Perception and Psychophysics 58, 351–362.

Munhall, K.G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. J. Acoust. Soc. Am. 104, 530–539.

Musacchia, C., & Schroeder, C.E. (2009). Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. Hearing Research 258, 72–79.

Nahorna, O., Berthommier, F., & Schwartz, J.L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. J. Acoust. Soc. Am. 132, 1061-1077.

Noiray, A., Cathiard, M.-A., Abry, C., Ménard, L. & Savariaux, C. (2008). Emergence of a vowel gesture control: Attunement of the anticipatory rounding temporal pattern in French children. In Kern S., Gayraud F. & Marsico E. (Eds.) Emergence of Language Abilities (pp. 100-117). Newcastle: Cambridge Scholars Pub.

Perkell, J. S., & Chiang, C. (1986). Preliminary support for a 'hybrid model' of anticipatory coarticulation. In Proc. 12th International Conference of Acoustics, A3–A6.

Perkell, J. S., & Matthies, L. M. (1992). Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability. J. Acoust. Soc. Am. 91, 2911–2925.

Roy, J.-R., Sock, R., Vaxelaire, B., & Hirsch, F. (2003). Auditory effects of anticipatory and carryover coarticulation. In Proc. 6th Int. Sem. Speech Production, Macquarie Series for Cognitive Sciences, 243-248.

Schwartz, J.L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. Cognition 93, B69–B78.

Schwartz J.-L., Boë L.-J., Badin P., & Sawallis R. T. (2012). Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial-coronal-velar stop series. Journal of Phonetics 40, 20-36.

Smeele, P. M. T. (1994). Perceiving speech: Integrating auditory and visual speech. Unpublished doctoral dissertation, Delft University of Technology.

Troille, E., Cathiard, M.A., & Abry, C. (2010). Speech face perception is locked to anticipation in speech production. Speech Communication 52, 513-524.

de la Vaux, S. K., & Massaro, D.W. (2004). Audiovisual Speech Gating: Examining Information and Information Processing. Cognitive Processing 5, 106-112.

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 Science, 102, 1181-1186.

van Wassenhove, V., Grant, K.W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. Neuropsychologia, 45, 598-607.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998) Quantitative association of vocal tract and facial behavior. Speech Communication 26, 23–43.

Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party. J Neurosci. 33, 1417-26.