



HAL
open science

Original objective and subjective characterization of phonetic convergence

Amélie Lelong, Gérard Bailly

► **To cite this version:**

Amélie Lelong, Gérard Bailly. Original objective and subjective characterization of phonetic convergence. ISICS 2012: International Symposium on Imitation and Convergence in Speech, Sep 2012, Aix-en-Provence, France. pp.O1:2. hal-00741686

HAL Id: hal-00741686

<https://hal.science/hal-00741686>

Submitted on 15 Oct 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Original objective and subjective characterization of phonetic convergence

Amélie Lelong and Gérard Bailly

GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal

{amelie.lelong,gerard.bailly}@gipsa-lab.grenoble-inp.fr

Introduction

Individuals accommodate their communication behavior either by increasing similarity with their interlocutors (i.e. convergence) or on the contrary by increasing their differences (i.e. divergence). Speech accommodation has been observed at both linguistic and non linguistic levels. Several studies have been conducted on phonetic dimensions such as pitch, speech rate, loudness or dispersions of vocalic targets with various experimental paradigms ranging from close-shadows of prerecorded stimuli to more ecological face-to-face conversations.

Multiple objective and subjective characterizations of phonetic convergence have been proposed. This paper discusses limitations of current proposals, notably in terms of top-down strategies that may be used by labelers and listeners when characterizing/perceiving the stimuli. We put forward and evaluate here two novel techniques: objective characterization by speaker recognition techniques and subjective characterization by a novel paradigm named “speaker switching”.

We will illustrate these techniques with stimuli collected during an original experimental paradigm called verbal dominoes (Lelong and Bailly, 2011), a speech game that can be played by several interlocutors and consisting in chaining rhyming words.

Objective characterization

Objective characterizations of convergence between two audio stimuli often involve the calculation of distances or correlations between time-aligned patterns. These characterizations are thus bounded to an a priori segmentation and labeling of relevant segments of interest, ranging from specific phonetic events (Fowler *et al.* 2008) to whole words (Delvaux and Soquet, 2007; Kim *et al.*, 2011). Distribution of various phonetic cues – VOT, formant frequencies, spectral tilts, durations of segments, etc. – are then collected in these segments and compared.

The identification, segmentation and labeling of segments of interest may provide interesting insights in phonological (i.e. cross-categorical) vs. phonetic (i.e. intra-categorical) accommodation issues. However this distinction is often neglected and difficult to disentangle – particularly in studies involving dialectal variations (see for example Aubanel and Nguyen, 2010) – by manual as well as automatic procedures.

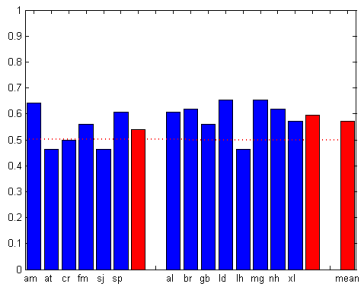
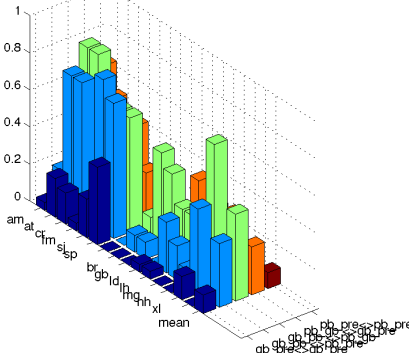
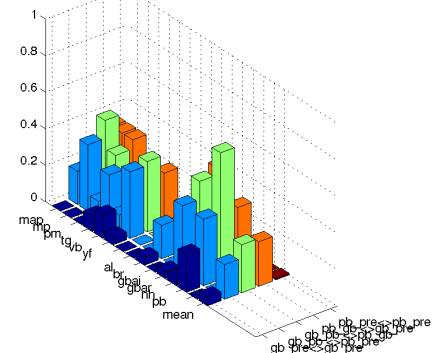
On the contrary, speaker recognition techniques often consider a global characterization of the phonetic space of each speaker without any a priori knowledge on the phonological variants used by speakers. We have demonstrated (Lelong and Bailly, 2012) that GMM-based speaker recognition scores correlates significantly with a more detailed analysis of the distributions of speaker-specific vocalic spaces. The correlation increases with the corpus size: we have found a significant correlation of .66 ($p < 0.01$) for the two objective measures of convergence in the case of large chains of 350 dominoes.

Subjective characterization

The AXB test introduced by Goldinger (1998) is the most widely used test for subjective characterization of phonetic convergence: listeners hear three versions of the same lexical items and judge which item produced by one talker, A or B, “sounds like a better imitation of” or “is more similar to” (Namy *et al.*, 2002) the X item produced by another talker. A reduced perceptual distance between X and one of the A or B items is then interpreted as a convergence of A/B towards X.

These results are often significant nevertheless, the size of the effects reported in the literature are often small with a preference for items produced in interaction with X rather than the ones produced with no interaction around 60% (e.g. reading) . We tested real and synthetic convergence (**created thanks to adaptive synthesis with the harmonic plus noise model interpolating parameters at 0% and 20% between both speakers**). Our conclusions about AXB tests are very disappointing. Subjects had trouble to remember A when hearing B even in the easiest case (e.g. when contrasting items with objective convergence rates of 0%

versus 20%). This led them to develop strategies unrelated to the task – such as focusing on prosodic variations or background noises – to ease decision. The final results mirror this difficulty (see Figure 1). We have recently tested a novel perceptual test that we named *speaker switching*¹. This test consists of generating a continuous signal where we randomly switch between items uttered by two speakers in different conditions, e.g. in isolation, imitating or interacting with one another. The listeners' task is simply to press a key each time they perceive/suspect a speaker switch. We considered that a switch was detected when the key hit occurred between the onset of the current item and the onset of the next. We experimentally set the ISI at 1000 ms. This is a rather rapid but comfortable presentation rate that favors immediate on-line processing and provides much more information and control data than the AXB test. We report preliminary results of two *speaker switching* experiments (see Figures 2 & 3) where we switched between 4 conditions: items read in isolation by two speakers and items uttered by the same speakers during a domino game. The stimuli are the same as for the AXB test. All subjects reported that this task was much easier than the AXB decision task.

Synthetic data: simulation of a symmetrical 20% convergence between speakers	Real data with a convergence between speakers close to 20%	
 <p data-bbox="151 1025 513 1361">Figure 1. Results of an AXB test with 12 listeners. X are dominoes pronounced by speaker 1 and synthesized at 20% to speaker 2. A and B are the same dominoes pronounced by speaker 2 either synthesized at 0% or 20% to speaker 1. Only the last 8 listeners knew the speakers. Despite a large objective phonetic convergence, listeners have difficulties in rating closeness (ratings close to 50%)</p>	 <p data-bbox="542 1093 952 1361">Figure 2. Speaker switching. Percentage of false detections for various transitions. Data from six listeners knowing the two speakers are displayed front. Data from other five listeners not knowing the speakers displayed at the back exhibit more confusion but display similar behavior: interaction increases misdetections.</p>	 <p data-bbox="981 1093 1423 1361">Figure 3. Same as Figure 2 but for real data. From left to right between: read items from speaker 1, interactive speech from speaker 1 and read items of speaker 2, interactive speech of both speakers, interactive speech from speaker 2 and read items of speaker 1 and read items from speaker 2. Data from six listeners knowing the two speakers are displayed front. Data from other six listeners not knowing the speakers displayed at the back.</p>

References

- Aubanel, V. and N. Nguyen (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Communication*, 52, 577-586.
- Delvaux, V. and A. Soquet (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, 64, 145-173.
- Fowler, C. A., V. Sramko, D. J. Ostry, S. A. Rowland and P. Halle (2008). Cross language phonetic influences on the speech of French-English bilinguals. *Journal of Phonetics*, 36(4), 649-663.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251-279.
- Kim, M., W. S. Horton and A. R. Bradlow (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, 2, 125-156.
- Lelong, A. and G. Bailly (2011). Study of the phenomenon of phonetic convergence thanks to speech dominoes, *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue*. A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud and A. Nijholt. Berlin, Springer Verlag, pp 280-293.

¹ We acknowledge Jason A. Shaw from UWS for his fruitful suggestions concerning this idea.

- Lelong, A. and G. Bailly (2012). Characterizing phonetic convergence with speaker recognition techniques. *Listening Talker Workshop*. Edinburgh.
- Namy, L. L., L. C. Nygaard and D. Sauerteig (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, 21, 422–432.