Characterizing phonetic convergence with speaker recognition techniques

Amélie Lelong, Gérard Bailly

To cite this version:
Amélie Lelong, Gérard Bailly. Characterizing phonetic convergence with speaker recognition techniques. LISTA 2012 - The Listening Talker Workshop (LISTA 2012), May 2012, Édimbourg, United Kingdom. pp.28-31. hal-00695558

HAL Id: hal-00695558
https://hal.science/hal-00695558
Submitted on 9 May 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.
CHARACTERIZING PHONETIC CONVERGENCE WITH SPEAKER RECOGNITION TECHNIQUES

Amélie Lelong and Gérard Bailly

GIPSA-lab, UMR 5216 CNRS/INPG/UFJ/U. Stendhal
{amelie.lelong,gerard.bailly}@gipsa-lab.grenoble-inp.fr

Abstract
Speakers are known to accommodate to each other’s behavior when interacting. Information circulates via multiple sensory-motor loops operating at various levels of the interaction and this closed-loop process induces modifications in all levels of representation from social and psychological evaluation to low-level gestural behaviors such as gaze, respiratory patterns, or speech. Various authors have proposed that these representations tend to converge or diverge according to cognitive demand. While quite plausible, claimed observations of such behavior in speech are extremely controversial. The effects are rather small, and are difficult to capture and characterize objectively. This paper focuses on the study of the convergence between phonetic representations – spectral realizations of speech sounds – using automatic classification techniques developed for speech and speaker recognition. Using data collected during a novel language game we term ‘verbal dominoes’, we show that scores are comparable between global techniques and a more fine-grained analysis focused on vocalic segments.

Index Terms: speaker recognition, phonetic convergence, speech adaptation

1. Introduction

Individuals accommodate their communication behavior [1] either by becoming 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 several levels. Researchers have in fact conducted studies on convergence of phonetic dimensions such as pitch, speech rate, loudness or dispersions of vocalic targets. The supposed goals and benefits of this accommodation include: easing comprehension, facilitating the exchange of highly context-dependent messages, disclosing ability and willingness to perceive, understanding or accepting new information, and maintaining social glue. Zoltan-Ford [2] has also shown that users of dialog systems tend to converge lexically and syntactically to the spoken responses of the system. Moreover, Ward et al[3] demonstrated that adaptive systems mimicking this behavior facilitate learning. But the phenomenon depends on several factors and most objective studies show only limited convergence, if any.

This emerging field of research is nonetheless central for two projects: the study of adaptive behavior during unconstrained conversation, and the substitution of an artificial conversational agent for a live partner.

This paper addresses two main topics: (a) we document a new method of collecting phonetic material to study and isolate the impact of the various factors influencing adaptation; and (b) we evaluate automatic techniques for quantifying any extant degree of convergence.

2. Observing and characterizing phonetic convergence

2.1. Scenarios

Researchers have used a variety of paradigms to characterize adaptation at different levels.

Perturbation of auditory feedback: Evidence that speakers tend to compensate for perturbation of their auditory feedback (see [4] for f0) lead some researchers to infer an internal sensory-motor speech representation towards which speakers tend to return in response to external excitations (or in their absence).

Imitation: Repetition and shadowing paradigms demonstrate convergence effects on Voice Onset Time (VOT) [5], F0 distribution [6], and articulation [7]. Sato et al. [8] showed that unintentional and voluntary imitation during the production of vowels used almost the same cognitive resources and resulted in similar behavior.

Ambient production: Delvaux and Soquet [9] tested the influence of ambient speech on pronunciation of certain keywords during a description task. They show small but significant effects on the spectra of target sounds when uttered in alternation with recordings of same vs. different Belgian dialects of French.

Interaction: Finally, researchers have studied phonetic convergence during unconstrained interaction. Pardo [10] examined the pronunciation of target words exchanged during a map task between pairs of same-sex talkers. Her perceptual experiments show that interactive speech decreases inter-subject distances. Aubanel and Nguyen [11] tested a new method of collecting dense interactive corpora with uncommon proper nouns, and they found a number of significant signatures of dialectal and phonetic convergence.

2.2. Computing degree of convergence

Quantification of convergence requires a baseline for comparison, so the default phonetic characteristics of speech segments (words, syllables, allophones) that will be analyzed during interaction are thus often collected through reading [10-11] or playing games alone [9] in a so-called pre-test session. Phonetic characteristics of the two speaker’s productions before interaction are then compared to those of speech segments uttered during the interaction or after (post-test).

To characterize phonetic convergence, most authors use spectral cues (formants, Mel Frequency Cepstral Coefficients i.e. MFCC) in the central part of particular segments of target words (mostly vowels & fricatives). The calculation of RMS distances between speaker-specific allophones are sometimes preceded by linear discriminant analysis [11].
To our knowledge, no results have been published addressing more global acoustic characterizations.

3. Speech recognition

The method for calculating convergence rates used by Delvaux & Soquet [9] and numerous researchers consists in computing an average distance between vocalic spaces using linear discriminant analysis (LDA). MFCC are first extracted for each vocalic target. Discriminant spaces are then computed for the central frames of each target sound for each pair of speakers. These frames are finally projected on the first discriminant axis separating speaker-specific spaces for the pronounced vowel and convergence rates for each target sound are calculated by normalizing the distance between speakers during interaction by the distance between vowels uttered during the pre-test. The convergence rate - noted as \( C_{LDA} \) - is then taken as the mean of these sound-specific rates (see Figure 1).

This method requires prior segmentation, labeling and clustering of specific target sounds (here vowels), the pronunciation of which speakers are supposed to mutually accommodate.

4. Speaker recognition

This paper compares the previous approach with a speaker recognition technique that compares the more global shape of the acoustic spaces. The experiments were performed using the MISTRAL platform [12]. We choose to perform speaker recognition by the Gaussian mixtures models (GMM), one of the most popular techniques for text-independent speaker recognition [13]. The speaker decision task mainly consists in a basic statistical test between two hypotheses:

- \( H_S: \) the speech characteristics \( y \) has been produced by the hypothesized speaker \( S \)
- \( H_{\neg S}; y \) is not from the hypothesized speaker \( S \) (often called the model of the “world”)

The decision uses a likelihood ratio (\( LR_S \)) test given by:

\[
LR_S(Y) = \prod_{y \in Y} \frac{p(y \mid H_S)}{p(y \mid H_{\neg S})} < \theta \tag{1}
\]

where \( p(y \mid H) \) is the probability density function for the hypothesis \( H \) evaluated for the speech segment \( y \) and \( \theta \) is the decision threshold for accepting or rejecting \( H_S \).

With MISTRAL, the log likelihood ratio (\( LLR \)) is computed over a test set of frames \( Y \). Two GMM respectively describe \( p(y \mid H_S) \) and \( p(y \mid H_{\neg S}) \) with the following law:

\[
p(y \mid H) = \sum_{i=1}^{M} w_i N(y \mid \mu_i, \Sigma_i) \tag{2}
\]

where \( w_i, \mu_i \) and \( \Sigma_i \) are the weights, mean vectors and covariance matrix of the \( M \) components of the mixture.

In our case, \( H_S \) and \( H_{\neg S} \) are the models of the two speakers of the dyad: the “world” \( \neg S \) corresponds to the interlocutor’s model. We then note:

\[
LLR_{\neg S}(Y) = \sum_{y \in Y} \log \left( \frac{p(y \mid H_S)}{p(y \mid H_{\neg S})} \right) \tag{3}
\]

GMMs here have \( M=64 \) components and the \( y \) components are MFCC coefficients estimated every 10ms.

These GMMs are trained in order to maximize \( LLR_{\neg S}(P_{s1})+LLR_{\neg S}(P_{s2}) \) over the set of frames \( P_{s1} \) and \( P_{s2} \) uttered respectively by speakers \( s1 \) and \( s2 \) during the pretest. This sum corresponds to the global distance between acoustic spaces of the two speakers.

Initialized using vector quantization, GMM parameters are refined by the iterative Expectation-Maximization (EM) algorithm in order to increase the likelihood of the estimated model for the observed feature vectors. Five to ten iterations are sufficient to get a correct estimation of each speaker’s model.

The convergence rate of \( s1 \) “towards” \( s2 \), noted \( C_{LLR}(s1 \rightarrow s2) \) is then taken as the relative quotient between the difference of a speaker’s \( LLR \) (here \( s1 \)) calculated with his own model on frames \( P_{s1} \) and during interaction \( I_{s1s2} \) and the difference of \( LLR \) calculated with the two interlocutor’s model on the pre-test \( P_{s1} \).

\[
C_{LLR}(s1, s2) = \frac{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(I_{s1s2})}{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(P_{s2})} \tag{4}
\]

where \( I_{s1s2} \) is the set of frames uttered by speaker \( s1 \) when interacting with speaker \( s2 \). So, if we don’t have any convergence, \( I_{s1s2}=P_{s1} \) and \( C_{LLR}(s1 \rightarrow s2) = 0 \).

5. Data

5.1. Speech dominos

A novel technique called “Speech Dominos” [14] was used to collect rich phonetic data on interactive speech. The rule of the game is quite simple: speakers had to choose between several alternatives the word that begins with the same syllable as the final one of the word previously uttered by the interlocutor (see Figure 2). The experiment was divided into two phases. Intrinsic references were gathered for each speaker during a pre-test session, where the speaker reads aloud a list of 350 words before any interaction with others. The pre-test words are those used during the dominoes’ game. During the game, each interlocutor pronounces respectively half of the pre-test
oral French vowels: \[a\], \[e\], \[i\], \[y\], \[u\], \[o\], \[r\].

Variations (about 20 tokens per speaker) of the eight peripheral syllabic words chosen to collect equal numbers of allophonic As our first analyses are focused on vowels [15], we select bi-the cognitive load and ease the running of successive sessions.

We chain here simple dissyllabic words in order to limit the length of the test and make pronunciation easier. The alternatives given to him are equally probable (e.g. both correct word in order to choose what to pronounce next since the respondent has to choose and utter alternatively the rhyming words. At the begin of the game, our tested subject will have to choose between [tordy] and [berly] the one that begins with [t]: he will thus utter [tordy] and so on.

We chain here simple dissyllabic words in order to limit the cognitive load and ease the running of successive sessions. As our first analyses are focused on vowels [15], we select bisyllabic words chosen to collect equal numbers of allophonic variations (about 20 tokens per speaker) of the eight peripheral oral French vowels: \([\text{a}], [\text{e}], [\text{i}], [\text{y}], [\text{u}], [\text{o}], [\text{r}]\).
6.1. Convergence and performance

We define turn-taking time (TTT) as the time delay between the onset of the last vowel of the domino pronounced by one speaker and the onset of the first vowel of the next domino uttered by his partner. Figure 4 shows the main impact of convergence on the evolution of TTT during the interaction: for moderate convergence rates, the degree of convergence of the initiator towards his partner correlates with increased TTT speed for the latter ($r = -.77$). We do not find this effect for the initiator’s turns. This tends to confirm that the role and background of each participant has a strong impact on behavior and performance [16].

![Relative variation of mean turn-taking time](image)

Figure 4. Relative variation of mean turn-taking time of the respondent as a function of $C_{lda}$ of the initiator. Test subjects increase the rhythm of the interaction (decrease turn-taking time) in response of the phonetic convergence of their interlocutor.

7. Conclusions

We have shown here that speaker recognition techniques provide a reliable estimate of the global degree of phonetic convergence without the need of phonetic segmentation or a procedure for part of speech pairing. For almost all pairs analyzed so far, few cases of divergence have been observed. On the contrary, large convergence rates have been found. Such impoverished phonetic contrasts between interacting speakers should be considered in automatic speaker tracking.

Our interaction paradigm offers other potential applications as well, regarding for instance the impact of word frequencies on the convergence [17] or rhythmical coupling across interlocutors. A perceptual validation of the large convergence effects found here is also called for.

This method will be used to characterize adaptation in less controlled conditions, to investigate the impact of conditions and linguistic content and study the dynamics of phonetic convergence. We plan to train statistical speech synthesis engines to implement the dynamics of the observed adaptation strategies. Such interlocutor-aware components are certainly crucial for creating social rapport between humans and virtual conversational agents [18].

8. Acknowledgements

This work was supported by ANR Amorces and by the Cluster RA ISLE. We thank Frederic Elisei, Sascha Fagel and Loïc Martin for their help.

9. References