Characterizing phonetic convergence with speaker recognition techniques

Amélie Lelong, Gérard Bailly

CHARACTERIZING PHONETIC CONVERGENCE WITH SPEAKER RECOGNITION TECHNIQUES

Amélie Lelong and Gérard Bailly

GIPSA-lab, UMR 5216 CNRS/INPG/UJF/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 adaptation 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 ie. 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.

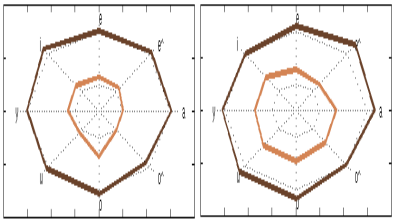

Figure 1. *Convergence rates C_{LDA} for the 8 vowels exchanged by two dyads. Circles with dotted lines (radius 1 and 4) feature default vocalic representations of the two speakers. Left: no convergence was found; Right: convergence of one speaker (C_{LDA}=0.38) towards the other, the orange line is getting closer to the outer circle representing the default vocalic representation of the partner.*

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:

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

where $p(y|H)$ is the probability density function for the hypothesis H evaluated for the speech segment y and θ 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 $p(y|H_S)$ and $p(y|H\neg_S)$ with the following law:

$$p(y \mid H) = \sum_{i=1..M} w_i N(y \mid \mu_i, \Sigma_i) \qquad (2)$$

where wi, μi and Σi 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:

$$LLR_{l1l2}(Y) = \sum_{y \in Y} \log\left(\frac{p(y \mid H_{l1})}{p(y \mid H_{l2})}\right) \qquad (3)$$

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

These GMMs are trained in order to maximize $LLR_{s1s2}(P_{s1})+LLR_{s2s1}(P_{s2})$[(1)] 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}).

$$C_{LLR}(s1, s2) = \frac{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(I_{s1s2})}{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(P_{s2})} \qquad (4)$$

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 Dominoes" [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*

(1) Since $LLR_{s1s2}(P_{s1}) = - LLR_{s2s1}(P_{s1})$, $LLR_{s1s2}(P_{s1}) + LLR_{s2s1}(P_{s2}) = LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(P_{s2})$

words, i.e. about 175 words. Figure 2 represents the first speech dominoes used in the interactive scenario. Interlocutors have to choose and utter alternatively the rhyming words. At each turn, speaker has to wait for his interlocutor to utter the correct word in order to choose what to pronounce next since the alternatives given to him are equally probable (e.g. both words [tɔrdy] and [tɔrʃi] exist in French and have roughly the same lexical frequency). Our reference subject first pronounces [rotɔr] to begin the game, then our tested subject will have to choose between [tɔrdy] and [bɛrly] the one that begins with [tɔr]: he will thus utter [tɔrdy] 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 bi-syllabic words chosen to collect equal numbers of allophonic variations (about 20 tokens per speaker) of the eight peripheral oral French vowels: [a], [ɛ], [e], [i], [y], [u], [o], [ɔ].

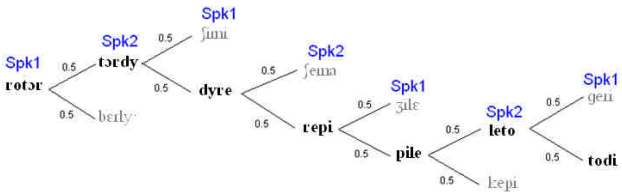

Figure 2. *First speech dominoes used in the interactive scenario. Correct rhymes in each pair are enlightened in bold.*

Table 1. *Convergence rates computed for each member of our 27 pairs by LDA (first column and second column of interact and pretest corresponds respectively to the mean convergence rate and its standard deviation) and LLR. Significant data (p<0.1 for LDA distributions for Interact vs. Pretest) are in bold.*

Game	Dyad	Initiator							Respondent						
		Sex	LDA				p	LLR	Sex	LDA				p	LLR
			Interact		Pretest					Interact		Pretest			
Session 1: 186 dominos - strangers	1	M	.03	.10	.03	.01	.89	.11	F	.03	.09	.04	.02	.79	.18
	2	M	.01	.05	.03	.02	.20	.14	F	.04	.10	.04	.02	.90	.01
	3	M	-.01	.09	.04	.02	.19	.20	F	.11	.14	.04	.02	.16	.09
	4	M	.04	.13	.09	.06	.32	.05	**M**	**.13**	**.08**	**.07**	**.03**	**.07**	**.08**
	5	M	.07	.14	.08	.05	.89	.25	**M**	**.28**	**.20**	**.07**	**.06**	**.01**	**-.05**
	6	M	.06	.19	.06	.03	.95	.16	**M**	**.31**	**.17**	**.05**	**.03**	**.00**	**.15**
	7	F	.01	.13	.09	.05	.14	.07	F	.15	.15	.08	.05	.21	.08
	8	F	.10	.19	.09	.06	.84	.05	M	.08	.11	.07	.06	.87	.07
	9	**F**	**.41**	**.38**	**.11**	**.06**	**.04**	**.30**	F	.18	.29	.08	.04	.33	.15
	10	M	.17	.24	.09	.08	.37	.19	M	.15	.11	.09	.09	.23	.06
	11	M	.08	.19	.07	.01	.78	.21	M	.04	.14	.07	.03	.46	.10
	12	M	.08	.09	.04	.02	.23	.09	**F**	**-.04**	**.07**	**.03**	**.02**	**.02**	**.07**
Session 2: 350 dominos - friends	13	**M**	**.41**	**.31**	**.07**	**.03**	**.01**	**.21**	M	.14	.12	.07	.02	.16	.18
	14	M	.15	.19	.05	.03	.16	.25	M	.13	.16	.04	.03	.16	.07
	15	**M**	**.40**	**.29**	**.07**	**.06**	**.01**	**.44**	M	.03	.22	.07	.05	.60	.16
	16	F	.00	.11	.03	.02	.49	.13	M	-.03	.12	.02	.01	.22	.11
	17	F	.00	.09	.03	.02	.50	.12	M	.04	.07	.03	.02	.46	.07
	18	F	.06	.12	.02	.01	.32	.13	**M**	**.10**	**.10**	**.02**	**.01**	**.04**	**.16**
	19	F	.14	.24	.06	.04	.38	.38	F	.13	.31	.08	.06	.66	.11
	20	**F**	**.39**	**.24**	**.06**	**.03**	**.00**	**.46**	F	.00	.14	.07	.02	.16	.08
	21	F	.16	.31	.04	.02	.30	.23	**F**	**.22**	**.25**	**.05**	**.03**	**.07**	**.13**
	22	F	.07	.33	.10	.07	.80	-.10	**F**	**.28**	**.20**	**.09**	**.06**	**.02**	**.16**
	23	**F**	**.15**	**.15**	**.06**	**.02**	**.09**	**.20**	**F**	**.18**	**.11**	**.06**	**.02**	**.01**	**.18**
	24	F	.22	.43	.08	.04	.36	.23	F	.34	.54	.09	.05	.21	.51
	25	F	.12	.16	.05	.03	.26	.28	**F**	**.15**	**.13**	**.06**	**.04**	**.08**	**.29**
	26	**F**	**.12**	**.14**	**.01**	**.01**	**.05**	**.39**	M	-.03	.11	.01	.01	.32	-.03
	27	**F**	**.34**	**.24**	**.06**	**.02**	**.01**	**.48**	F	.22	.35	.06	.02	.23	.27

5.2. Speakers' models, reference and test data

Only half of the pre-test data are used to train the speakers' models. The other half is used as reference data. We used a simple cross validation procedure: the convergence rates are the mean values of relative distances between reference and test data over 10 random partitions between training and reference data.

In a first series of experiments with 186 dominos [15], we noted that phonetic convergence was higher for dyads who already knew each other and particularly for women [10], as shown in the first 12 pairs in Table 1 and Figure 3. During this condition, speakers were in two different rooms and communicated through microphones and headphones. This setup was easy to realize thanks to the MICAL platform of our laboratory (two rooms separated with a tinted mirror). In a second series involving good friends exchanging a larger number of dominoes (350), 3 male dyads, 4 mixed dyads and 8 female dyads have been tested. 5 men from 24 to 54 years old and 11 women from 18 to 26 years old participated. In this case, people were engaged in a real face-to-face interaction.

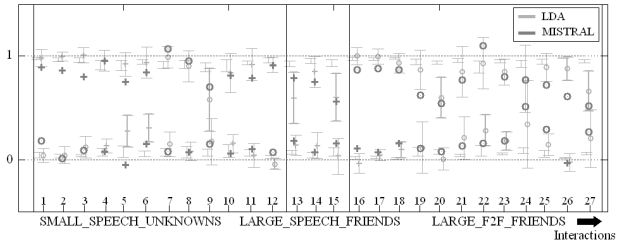

Figure 3. *Comparison of convergence rates for the two methods (1 is for the initiator and 0 for the respondent).The dark gray color represents the results obtained with MISTRAL. The results obtained with the different methods are quite similar. During interactions 1 to 15, people were communicated thanks to microphones and headphones while they were in a face-to-face condition for interactions 16 to 27.*

6. Results

Table 1 displays convergence rates C_{LDA} and C_{LLR} computed for all dyads. An ANOVA test was performed to test significant deviations between reference and test C_{LDA}. Distributions with significant convergence rates ($p<0.1$) are noted in bold.

Convergence is not systematic. Moreover, we can see that the phenomenon is amplified for same sex pairs (see pairs 4-6, 9, 13-15, 20-23 and 25-27) and particularly for women. This observation led us to select mostly women for the final interactions, and the results largely confirm this tendency.

We found a significant correlation of .64 ($p<0.05$) between these two coefficients for initiators and of .73 ($p<0.005$) for the respondents in the case of the large corpus (last 15 interactions in Table 1 and Figure 3). The correlations calculated on the first 12 interactions are lower. We do think that this is the consequence of insufficient training data provided by the 186 dominos.

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].

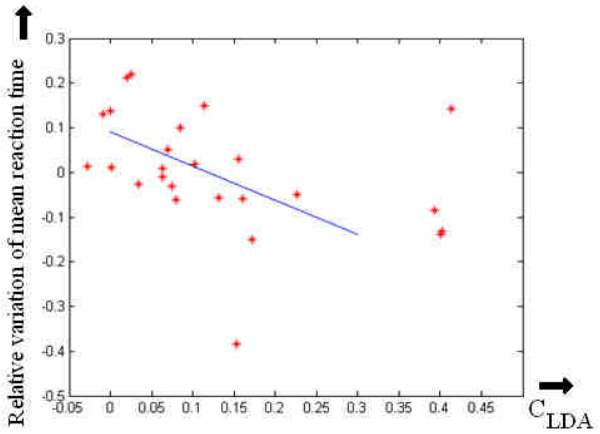

Figure 4. *Relative variation of mean turn-taking time of the respondent as a function of C_{LDA} of the initiator. Test subjects increase the rhythm of the interaction (decrease turn-taking time) in response of the phonetic convergence of their interlocutor.*

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

[1] Giles, H., J. Coupland, and N. Coupland, *Contexts of Accommodation: Developments*. Applied Sociolinguistics. 1991, Cambridge: Cambridge University Press.

[2] Zoltan-Ford, E., *How to get people to say and type what computers can understand.* International Journal of Man-Machine Studies, 1991. **34**: p. 527-547.

[3] Ward, A. and D. Litman. *Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora*. in *SLaTE Workshop on Speech and Language Technology in Education*. 2007. Farmington, PA.

[4] Jones, J.A. and K.G. Munhall, *Perceptual calibration of F0 production: Evidence from feedback perturbation.* Journal of the Acoustical Society of America, 2000. **108**: p. 1246-1251.

[5] Fowler, C.A., et al., *Cross language phonetic influences on the speech of French-English bilinguals.* Journal of Phonetics, 2008. **36**(4): p. 649-663.

[6] Gregory, S.W., S. Webster, and G. Huang, *Voice pitch and amplitude convergence as a metric of quality in dyadic interviews.* Language and Communication, 1993. **13**: p. 195-217.

[7] Gentilucci, M. and P. Bernardis, *Imitation during phoneme production.* Neuropsychologia, 2007. **45**(3): p. 608-615.

[8] Sato, M., et al. *Converging to a common speech code: automatic imitative and perceptuo-motor recalibration processes in speech communication.* in *Second Neurobiology of Language Conference*. 2010. San Diego, USA.

[9] Delvaux, V. and A. Soquet, *The influence of ambient speech on adult speech productions through unintentional imitation.* Phonetica, 2007. **64**: p. 145-173.

[10] Pardo, J.S., *On phonetic convergence during conversational interaction.* Journal of the Acoustical Association of America, 2006. **119**(4): p. 2382–2393.

[11] Aubanel, V. and N. Nguyen, *Automatic recognition of regional phonological variation in conversational interaction.* Speech Communication, 2010. **52**: p. 577-586.

[12] Charton, E., et al. *Mistral: an open source biometric platform* in *25th Symposium on Applied Computing (SAC)*. 2010. Switzerland.

[13] Reynolds, D., *Speaker identification and verification using Gaussian mixture speaker models.* Speech Communication - special issue on Face-to-Face Communication, 1995. **17**(1): p. 91-108.

[14] Bailly, G. and A. Lelong. *Speech dominoes and phonetic convergence*. in *Interspeech*. 2010. Tokyo. p. 1153-1156.

[15] Lelong, A. and G. Bailly, *Study of the phenomenon of phonetic convergence thanks to speech dominoes* in *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue*, A. Esposito, et al., Editors. 2011, Springer Verlag: Berlin. p. 280-293.

[16] Pardo, J.S., I. Cajori Jay, and R.M. Krauss, *Conversational role influences speech imitation.* Attention, Perception, & Psychophysics, 2010. **72**: p. 2254-2264.

[17] Goldinger, S.D., *Echoes of echoes? An episodic theory of lexical access.* Psychological Review, 1998. **105**: p. 251-279

[18] Gratch, J., et al. *Creating rapport with virtual agents.* in *Intelligent Virtual Agents (IVA)*. 2007. Paris, France. p. 125-138.