Speech dominoes and phonetic convergence

Gérard Bailly and Amélie Lelong

Département Parole & Cognition, GIPSA-lab, UMR 5216 CNRS/Université de Grenoble
{gerard.bailly,amelie.lelong}@gipsa-lab.grenoble-inp.fr

Abstract

Interlocutors are known to mutually adapt during conversation. Recent studies have questioned the adaptation of phonological representations and kinematics of phonetic variables such as loudness, speech rate or fundamental frequency. Results are often contradictory and the effectiveness of phonetic convergence during conversation is still an open issue. This paper describes an original experimental paradigm – a game played in primary schools known as verbal dominoes - that enables us to collect several hundreds of syllables uttered by both speakers in different conditions: alone, in ambient speech or in full interaction. Speech recognition techniques are then applied to globally characterize phonetic convergence if any. We hypothesize here that convergence of phonetic representations such as vocalic dispersions is not immediate especially when considering common words of the target language.

Index Terms: phonetic convergence, spectral distances, speaker adaptation, and interaction.

1 Introduction

The Communication Accommodation Theory (CAT) introduced by Giles et al. [9] postulates that individuals accommodate their communication behavior to each other either by becoming more similar (convergence) or by exaggerating their differences (divergence). Conversational partners notably adapt to each other's choice of words, particularly referring expressions [4] and converge on certain syntactic choices [15]. Zoltan-Ford [21] has also shown that users of dialog systems converge lexically and syntactically to the spoken responses of the system. The aim of this alignment [19] may have several benefits such as ease the task of exchanging meaning with highly context-dependent messages [14], disclose ability and willingness to perceive, understand or accept new information [1] as well maintain social glue or resonance [12].

Researchers have questioned adaptation of phonetic dimensions such as f0 [11], speech rate [7, 13], loudness [13] or dispersions of vocalic targets [5]. The results of these studies evidence rather small or sometimes no adaptation at all. The perceptual study conducted by Pardo [18] evidences disparities between talkers that he attributes to various dimensions such as social settings, communication goals and varying roles in the conversation.

This emerging field of research is central to the comprehension of adaptive behavior during free conversation and to versatile speech technologies that aim at substitute one partner with an artificial conversational agent. The survey of the literature shows that two main challenges are quite open: (a) the need of original experiments that allow us to collect sufficient phonetic material to study/isolate the impact of the numerous factors influencing adaptation; (b) the search of automatic techniques for characterizing the degree of that convergence if any.

2 State of the art

Studies by Pardo [18] and Delvaux and Soquet [5] are cited in most recent studies because their states of the art, methologies, results and comments summarize almost all we know so far on phonetic adaptation.

2.1 Description of key studies

Pardo [18] examined whether pairs of talkers converged in phonetic repertoire over the course of a single conversational interaction. Six same-sex pairs of talkers interacted to solve a series of 5 map tasks where their role – instruction giver or receiver – are exchanged. The interest of the map task landmark is to collect landmark label phrases that are uttered several times by each interlocutor for the receiver to replicate the itinerary known by the giver. One and two weeks before, talkers read aloud the set of map task landmark labels in order to get reference pronunciations. Following interactions, this reading was also performed to test persistence of convergence, i.e. distinguishing stimulus-dependent mimicry from mimesis that supposes a deeper change of phonetic representations [6]. 30 listeners judged perceptual similarity between pronunciations of pre-, map- and post-task landmark labels in a AXB test, X being a map-task utterance and (A,B) pre-, map- or post-task of the same utterance pronounced by the corresponding partner. Result of this forced choice evidences a significant effect of expose and persistence but with important effects of role and sex: givers converge notably more to receivers than the opposite, female givers even more than male givers.

Delvaux and Soquet [5] studied influence of ambient speech to the pronunciation of keywords. In this non interactive experiments subjects described a simple scene "*It s in X that there are N Y*", where X are locations, N numbers and Y objects. This description is either done by the speaker or recorded speakers of same or different Belgian dialects. Pre- and post-tasks are also performed. The phonetic analysis focussed on the realization of two sounds that are used in the two possible labels X. The allophonic variations of these sounds are typical of the two dialects and the authors seek for unintentional imitation. A tentative characterization of the amplitude of that change is performed by comparing durations and computing spectral distance. In most cases, small but significant displacements towards the prototypes of the other ambient dialect are observed for both sounds. Similar unconscious imitation of characteristics of ambient speech has also been observed by Gentilucci et al [8] for audiovisual stimulations.

2.2 Comments

These studies show that pronunciations of keywords belonging to the speakers' common vocabulary significantly narrow. The collection of a significant number of occurrences of these few words is however lengthy. Aubanel and Nguyen [3] have proposed an original paradigm to collect dense

interactive corpora with uncommon proper nouns but again these words are under strong focus and enter rapidly in the mutual lexicon. Moreover, Delvaux and Soquet as well as Aubanel and Nguyen study the mutual influence between regiolects, northern/southern Belgian or French accents, that could be part of the subjects' experience. In any case, we study here phonological/pronunciation convergence - e.g. choice of open/closed mid-vowels for open syllables such as /le/ in northern French vs. /lɛ/ in southern French for the work lait (milk) - rather than more subtle phonetic convergence within the same idiolect.

Table 1. Number of phones collected for each speaker during the dominoes' game. 350 CV or CVC syllables are pronounced in total.

phones	a	ɛ	e	i	y	u	o	ɔ	others
#items	47	48	45	43	44	40	43	31	9

3 Material and method

In our experiments, speakers are instructed to choose and pronounce words displayed on a computer screen.

3.1 Dominoes' game

During the interaction task, they have to select the word in the word list which begins with the same syllable as the word previously pronounced by the interlocutor. Such rhyme games – called speech dominoes – are part of the children's folklore and widely used in primary schools for language learning [2]. We choose here to chain simple dissyllabic words such as in:

bateau [bato], taudis [todi], diffus [dify], furie [fyri], etc.

The words are chosen so that to uniformly collect allophonic variations of certain sounds of the target language (here French). We choose here to study the eight peripheral oral vowels: [a], [ɛ], [e], [i], [y], [u], [o], [ɔ]

The word list offers alternatives built so as the speaker can not guess the next domino he has to utter on his sole information. He has effectively to pay attention to the word uttered by his interlocutor to decide what "domino" to utter next. As an example, if starting with [bato], he will be presented with at least two alternatives [dify] and [rozo], [todi] and [toro] being two valid and equiprobable common words. Special attention should be paid in general to lexical frequency in order to ease pronunciation and not bias attention.

In order to limit the speakers' cognitive load and ease the successive sessions, the number of alternatives was limited to two. A series of 350 chained dominoes was thus established that collects almost 40 exemplars of each peripheral oral vowel (see Table 1).

spk 1	spk 2	spk 1	spk 2	spk 1	spk 2	spk 1	
rotɔr	tordy	ʃimi	ʃema	ʒilɛ	leto	geri	…
	bɛrly	dyre	repi	pile	kepi	todi	

Figure 1. First speech dominos used in the interactive scenario. Interlocutors have to choose and utter alternatively the rhyming words. Correct rhymes in each pair are enlightened with a dark background. Note that alternatives may be triggered by valid words - i.e. [torʃi], [dyʃe], [reʒi], [pike], [lege] - if they would have been part of the previous alternatives of the interlocutor.

3.2 Conditions

Each speaker's phonetic space is characterized by recording dominoes before interaction: all words that will be pronounced by either interlocutor are uttered during a pre-test session that takes place before any dialog with the interlocutor. The pre-test reference condition will be used to measure the amplitude of adaption if any.

Different sessions and recording conditions can be added such as ambient speech where speakers play with pre-recorded dominoes of one unique or various interlocutors. A post-test can also be added to address the question of the depth of the cognitive adaptation: imitation or mimesis.

Figure 2: Face-to-face interaction. The scene is captured by a unique camera thanks to a mirror positioned at the left hand side of one interlocutor. Head movements were monitored during this experiment.

3.3 Experiments

We will here only contrast pre-test and interactive conditions in three series of experiments.

1. Experiment I "unknowns" is performed by pairs of subjects that have been never talked to each other.
2. Experiment II "friends" is performed in contrast by pairs who are good friends, knowing and working together for years.
3. Experiment III "face-to-face" friends sat across a table where two screens are placed back-to-back for displaying alternatives. Clicks on one unique moose used alternatively by each subject forward turns.

For experiments I and II, interlocutors sat in different rooms and communicated by sets of microphones and earphones. They are instructed to avoid speech overlaps and repairs so as to ease automatic segmentation and alignment. Signals are digitized at 16 kHz by a high-quality stereo sound card.

3.4 Characterization

Delvaux and Soquet [5] noticed that global automatic analysis of spectral distributions by MFCC lead to quasi-identical but more robust characterization of convergence than a more detailed semi-automatic phonetic analysis such as formant tracking. Aubanel and Nguyen [3] similarly use automatic recognition techniques to recognize idiolects.

We trained phone-sized context-independent HMMs using pre-test data (using HTK [20]) and performed various forced alignments to compare the distribution of normalized self vs. other's recognition scores (see Figure 4). Paired T-tests are performed to compare changes of distributions of scores of vowels produced in the same words (here 175 words for each speaker).

Note that semi-automatic segmentation of each word performed by aligning self HMMs is checked by hand.

Devoiced or creaky vowels - often high vowels [i], [y] in unvoiced context - are discarded. Idiolectal variations are also considered: allophones are labeled with open or closed mid-vowels according to the majority perceptual identification of the isolated syllable by three independent labelers.

4 Results

We compare the distributions of normalized recognition scores of the pre-test and interactive utterances. These utterances are recognized both by the speaker's own HMMs and the HMMS of his interlocutor. For pre-test data, we expect by construction high scores for HMMs tested on own training data and lower scores for HMMs of the interlocutor. The difference between the scores reflects somehow the inter-speaker distance. Convergence of interactive speech if any is cued by a joint decrease of scores by self HMMs and increase of scores by other's HMMs.

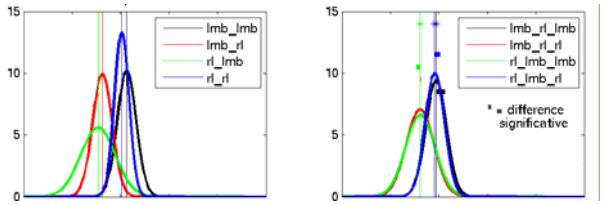

Figure 3. Distribution of recognition scores for the vowels of disyllabic words produced by two unknowns. The recognition is performed by their own HMM models and by the HMM models of their interlocutor. Left: scores for word lists read aloud in isolation; this speech data is used to train the speaker-specific HMM models. Right: same words pronounced in a verbal domino game. Only small adjustments are observable.

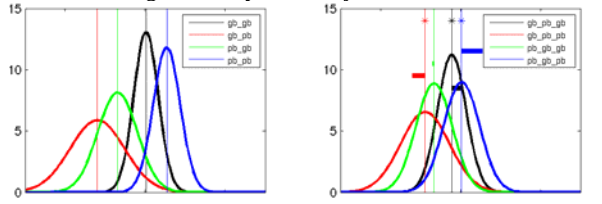

Figure 4. Same of Figure 3 but for disyllabic words produced by two old friends. Larger convergence is observed.

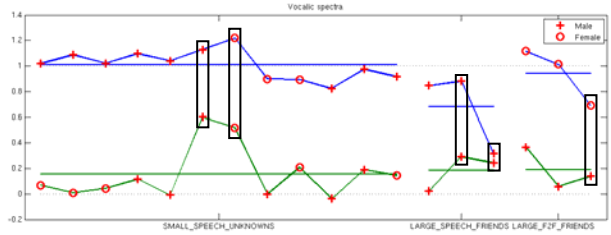

Figure 5: Average convergence rate of vocalic targets of interlocutors for all conditions. Top rates are for game initiators. Largest convergences (see frames) are observed between pairs of same sex (circles for female vs. crosses for males).

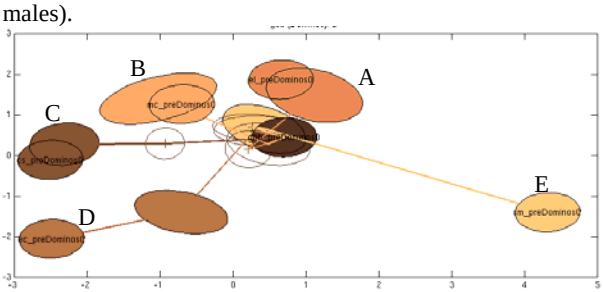

Figure 6: Projection - on the first discriminant space - of MFCC targets for [a] produced by speaker gbb (dark dispersion ellipsis for the pre-test at the center) interacting successively with interlocutors A, B, C, D and E (pre-test ellipsis at the periphery). Realizations for interactions are displayed with unfilled ellipsis for gbb and filled ellipsis with same color as pre-test for interlocutors. While A, B and C are very conservative, D converges partially and E totally to gbb. Conversely gbb adapts partially only to C (unfilled ellipses).

4.1 Distributions of recognition scores

Figure 3 and Figure 4 displays the distribution of recognition scores for the vowels of disyllabic words produced by two dyads. While Figure 3 small but significant convergence of one speaker to the other, Figure 4 characterizes the occurrence of a large and significant convergence of both interlocutors towards one with the other.

Figure 5 shows the average convergence rate for all dyads recorded in the three conditions computed as the relative MFCC distance between pretest vs. interaction vocalic targets (central state of the HMM alignment). Convergence is not systematic. The largest convergence is observed in our date between pairs of same sex except one pair (a young couple living for 3 years together).

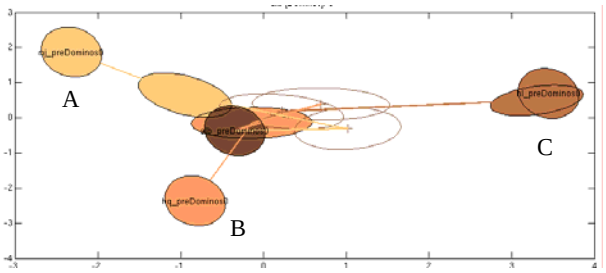

Figure 7: Same as Figure 6 but for the vowel [ɔ] the female speaker alb interacting with 3 different interlocutors A, B & C. While A and B converge to alb, alb and C do not adapt.

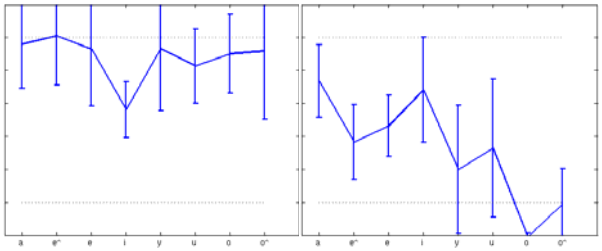

Figure 8: Vowel-dependent convergence patterns. Left : only the distance between spectra of the spread vowel [i] are significantly minimized. Right: for that dyad, rounded vowels do converge more than others.

4.2 Convergence of vocalic targets

Game initiators typically interacted with 2 to 5 different interlocutors in order to question the interlocutor-specific adaptation strategies. Figure 6 and Figure 7 give examples of target- and interlocutor-specific behaviors. When considering each vowel separately (40 occurrences in average, see Table 1), we do observe cases of full convergence (see speaker E in Figure 6 and speakers A and B in Figure 7).

Target-specific strategies seem to be structured as amplified by Figure 8. It would be worth testing on a larger database if these strategies are governed by the adaptation within compact acoustic regions or consistent articulatory maneuvers that could be enhanced by face-to-face interaction [8].

Please note however that our analysis is based on relative distances and should take into account the whole structure of

the vocalic space of the interlocutors. Speakers notably fill differently their available acoustic space, especially between mid-vowels [16, 17].

4.3 Prosody

Figure 9 shows that prosody is relatively unaffected by the interaction. This is probably due to the task that focuses on phonemic adaptation. Delvaux and Soquet [5] recommend to discard final syllables for studying phonetic convergence. In our case, we did not find any difference between full and partial statistics except the stronger convergence for the durations of vocalic nuclei of initial syllables.

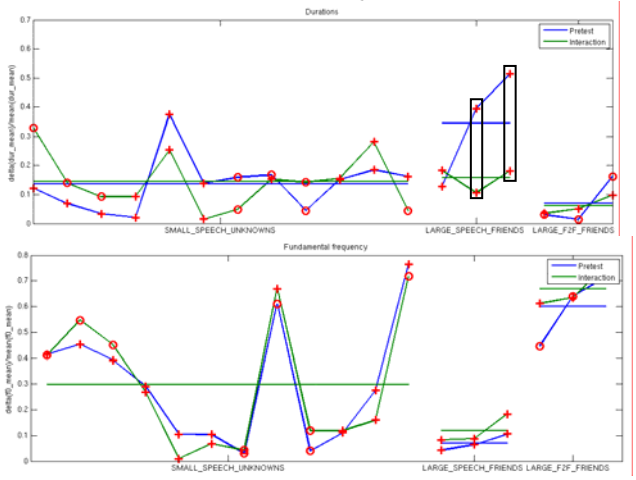

Figure 9: Mean changes of relative difference between durations (top) and F0 (bottom). The only significant changes - occurring on durations - are framed.

5 Conclusions and perspectives

We propose here an original speech game that collects rapidly many instances of target sounds with a mutual influence that force interlocutors to engage into active action-perception loops. Distribution of target sounds can be explicitly controlled to observe convergence in action if any.

We indeed find occurrences of strong convergence with few instances of modest divergence. This convergence depends strongly on the dyads – with strongest convergence observed for pairs of the same sex - and seems to be phoneme-dependent.

We will now use this gaming paradigm to select subjects and dyads who exhibit the strongest adaptation abilities and study more complex conversational situations. This data will be used to train speech synthesis engines that will implement these adaptation strategies. Such interlocutor-aware components are certainly crucial for creating social rapport between humans and virtual conversational agents [10].

6 Acknowledgments

This work has been financed by ANR Amorces and by the Cluster RA ISLE. We thank Frederic Elisei and Loïc Martin for their help.

7 References

[1] Allwood, J., *Bodily communication - dimensions of expression and content*, in *Multimodality in Language and Speech Systems*, B. Granström, D. House, and I. Karlsson, Editors. 2002, Kluwer Academic Publishers: Dordrecht. p. 7-26.

[2] Arléo, A., *Un jeu de dominos verbal: Trois p'tits chats, chapeau d'paille*, in *Chants enfantins d'Europe*, A. Arléo, et al., Editors. 1997, L'Harmattan: Paris. p. 33-68.

[3] Aubanel, V. and N. Nguyen, *Automatic recognition of regional phonological variation in conversational interaction.* Speech Communication, 2010: p. to appear.

[4] Brennan, S.E. and H.H. Clark, *Lexical choice and conceptual pacts in conversation.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 1996. **22**: p. 1482-1493.

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

[6] Donald, M., *Origins of the Modern Mind: three stages in the evolution of culture and cognition.* 1991, Cambridge, MA: Harvard University Press.

[7] Edlund, J., M. Heldner, and J. Hirschberg. *Pause and gap length in face-to-face interaction*. in *Interspeech*. 2009. Brighton. p. 2779–2782.

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

[9] Giles, H. and R. Clair, *Language and Social Psychology*. 1979, Oxford: Blackwell.

[10] Gratch, J., N. Wang, J. Gerten, E. Fast, and R. Duffy. *Creating rapport with virtual agents*. in *Intelligent Virtual Agents (IVA)*. 2007. Paris, France. p. 125-138.

[11] Gregory, S.W., *Social psychological implications of voice frequency correlations: analyzing conversation partner adaptation by computer*. Social psychology quarterly, 1986. **49**(3): p. 237-246.

[12] Kopp, S., *Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors.* Speech Communication, in press.

[13] Kousidis, S., D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, C. McDonnell, and E. Coyle. *Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues.* in *Interspeech*. 2008. Brisbane. p. 1692-1695.

[14] Lakin, J., V. Jefferis, C. Cheng, and T. Chartrand, *The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry.* Nonverbal Behavior, 2003. **27**(3): p. 145–162.

[15] Lockridge, C.B. and S.E. Brennan, *Addressees needs influence speakers early syntactic choices.* Psychonomic Bulletin and Review, 2002. **9**: p. 550-557.

[16] Ménard, L., J.-L. Schwartz, and J. Aubin, *Invariance and variability in the production of the height feature in French vowels.* Speech Communication, 2008. **50**(1): p. 14-28.

[17] Neagu, A., *Analyse articulatoire du signal de parole: caractérisation des syllabes occlusive-voyelle en Français*. 1998, Institut National Polytechnique: Grenoble - France.

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

[19] Pickering, M., H. Branigan, A. Cleland, and A. Stewart, *Activation of syntactic priming during language production.* Journal of Psycholinguistic Research, 2000. **29**(2): p. 205–216.

[20] Young, S., D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. 1999, Cambridge, United Kingdom: Entropic Ltd.

[21] 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.