Measuring dynamics of mimicry by means of prosodic cues in conversational speech

Céline de Looze, Catharine Oertel, Stéphane Rauzy, Nick Campbell

MEASURING DYNAMICS OF MIMICRY BY MEANS OF PROSODIC CUES IN CONVERSATIONAL SPEECH

Céline De Looze[a], *Catharine Oertel*[a], *Stéphane Rauzy*[a] & *Nick Campbell*[b]

[a]Speech Communication Laboratory, Trinity College Dublin, Ireland;
[b]Laboratoire Parole et Langage, CNRS UMR 6057, Université de Provence, France
deloozec@tcd.ie; oertelgc@ tcd.ie;
nick@tcd.ie; stephane.rauzy@lpl-aix.fr

ABSTRACT

This study presents a method for measuring the dynamics of mimicry in conversational speech by means of prosodic cues. It shows that the more speakers are involved in a conversation, the more they intend to mimic each other's speech prosody. It supports that mimicry in speech is part of social interaction and that it may be implemented into spoken dialogue systems in order to improve their efficiency.

Keywords: mimicry, prosody, modeling, social interaction

1. INTRODUCTION

What makes a conversation, an interactive dialogue, are the dynamic changes involved in spoken interaction. Interlocutors do not remain involved to the same degree over the whole course of a conversation; they may change from being inactive to talking, going through phases such as listening, thinking, arguing a point or giving feedback. It can be assumed that the phenomenon of mimicry - also found under the terms of "convergence", "accommodation", "alignment", "imitation", "entrainment" or "synchrony" - undergoes dynamic changes in spoken interaction. Burgoon, et al [2] define mimicry as "*The situation where the observed behaviours of two inter-actants although dissimilar at the start of the interaction are moving towards behavioral matching*". This implies that speakers tend to imitate over the course of the interaction. However, it can be assumed, according to the functions it may convey, that the phenomenon of mimicry tends to vary throughout the conversation, resulting in phases of mimicry and phases of non-mimicry. Kousidis, et al [7] for instance found that speakers imitate each other's prosodic features over time; this can be attributed to the fact that participants are involved in a cooperation task.

A growing number of studies have investigated mimicry in speech in terms of speech sounds, lexicon, syntax as well as prosody (for a review see [9]). Mimicry strength was either measured for one speaker in different dyads or for two speakers in one dyad. In most of these studies however, with the exception of [3, 6], the methodologies or metrics developed failed to capture the temporal dynamics of mimicry. In this paper we measure mimicry in speech by means of prosodic cues; we distinguish between measurements of the whole interaction and at various points of the conversation.

We also investigate the extent to which phases of mimicry are correlated with speaker's level of agreement or degree of involvement since mimicry has been reported to be linked to speakers' attitudes and topic discussed [5].

We hypothesize that (1) automatic extraction of prosodic cues (pitch level and span, speech rate and voice intensity) can be used to detect phases of mimicry; (2) the higher the level of agreement and (3) the higher the degree of involvement, the higher the strength of mimicry.

We argue that taking into account the temporal dynamics of mimicry improves the modeling of social interaction and hence the efficiency of spoken dialogue systems.

2. EXPERIMENT

2.1. Data

Our study was based on the D64-corpus [8]. The D64 corpus consists of the conversational speech of 5 speakers and was collected in a domestic apartment; this corresponds to a total of 8 hours of recordings. For our analyses, sections where only two participants took part in the conversation were selected. The first section consists of the conversation of speaker 1 (S1; male) and speaker 2 (S2; female). The topic under discussion was S2's master thesis, S1 supervising S2's work. In the

second section, S1 again participates in the conversation but this time with a male colleague (speaker 3; S3). They exchanged personal experiences and opinions (e.g. politics, travel, etc.). Each interaction lasts about half an hour.

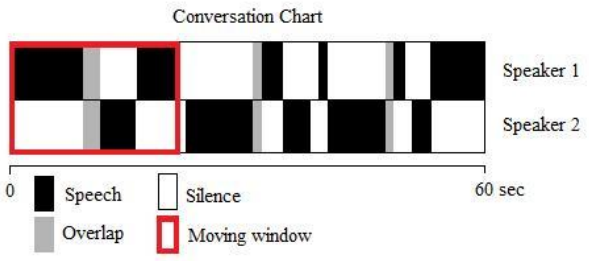
2.2. Segmentation and measurements

The prosodic parameters under investigation are pitch level and span, number and mean duration of pauses and voice intensity. Acoustic measurements were obtained using the phonetic software Praat [1]. Pitch level and span were measured by calculating the F0-median and the log2(F0max − F0min) respectively. The F0-median is given on a linear scale (i.e. Hertz) while F0-max/min is given on a logarithmic scale (i.e. octave). Silent pauses were detected automatically and corrected manually. Filled pauses, laughters and overlaps were excluded from the analyses. The intensity of the voice was expressed as the root mean square (RMS) amplitude (rms-Int) and standard deviation Intensity (sd-Int).

2.3. Prosodic cues extraction

The difficulty encountered when measuring speech mimicry is that it is not time-aligned. To resolve this, the TAMA method, as did Kousidis, et al [7] or Edlund, et al [3], was used. Average values of prosodic cues were automatically extracted from a series of overlapping windows of fixed length (20 seconds) using a time step of 10 seconds. This means that prosodic cues were extracted for each speaker every 10 seconds.

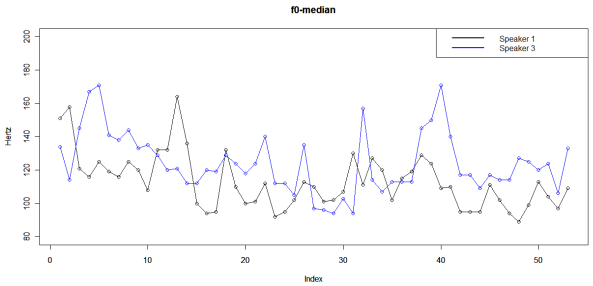
Figure 1: Parts of the conversation chart for speakers 1 and 2 interaction.



Average values were calculated proportionally to the utterances' length within the window giving a bigger weight to the prosodic cue whose utterance's length is the longest. **Figure 1** shows the moving window along speakers' interaction (represented by a conversation chart).

Average values were plotted to visually investigate whether speakers tend to mimic each other's speech. **Figure 2** represents the two times series of f0-median average values obtained for S1 and S3.

Figure 2: Time series of f0-median average values (represented here in Hertz) obtained for S1 and S3 interaction (17 minutes).



In order to account for speaker differences in pitch level and span, number and mean duration of pauses as well as voice intensity, data was normalised by a log- (except for f0-span) and an additional z-score transformation.

2.4. Mimicry strength: measurement and significance

The mimicry strength is herein measured by computing the Pearson's correlation coefficient (I) of the two time series representing the variation of a given prosodic parameter for the two speakers in the interaction. It is expected that mimicry strength is equal to zero during non-mimicry phases and becomes positive when the two speakers imitate each other. To decide whether mimicry strength is significant or rather due to sampling fluctuations, the Fisher's transformation is applied to I (F(I)).

I and F(I) were first calculated for the whole interaction, then for individual sections in order to measure temporal variations of mimicry. Calculating the coupling of the two time series over time allows for detecting phases of mimicry and non-mimicry. To calculate temporal variations of mimicry, a series of overlapping windows of a fixed length (20 points) and a time step of 5% of the window's length were used. It is however expected that statistical significance may not be reached since for a given prosodic parameter, the number of independent measurements falls to 10 with our choice of window size.

2.5. Annotation of agreement and involvement

The data was annotated for agreement, disagreement and neutral speech (i.e. monologues). Values of -2 were attributed to sections of disagreement, 0 to neutral speech and +2 to

sections of agreement. The data was also annotated for involvement. By *involvement* we refer to the general involvement of a group of speakers rather than the involvement of an individual in a conversation [X]. Involvement was annotated on a scale from 0 to 10 for 5 second intervals; 0 being the smallest degree of involvement and 10 the highest. Only involvement values between 4 and 9 were chosen for the here selected parts of the corpus. In order to validate the annotation schema a perception test was conducted in which 20 participants took part. Inter-annotator agreement was found to have a kappa value of 0.56.

3. RESULTS

3.1. Mimicry and prosodic cues

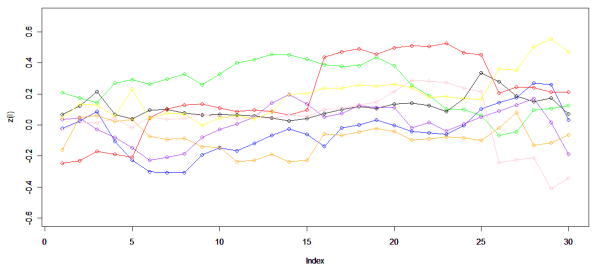
3.1.1. Measuring imitation for the whole interaction

For the interaction S1/S2, mimicry is not detected on the whole interaction. However, for interaction S1/S3 results show that S1 and S3 modulate their voice intensity level and variation (rms-Intensity, p=0.01539 & sd-Intensity, p=0.00212), mean pause duration (dpauses, p=0.00256) as well as the ceiling of their pitch range (f0-max, p=0.01044) to imitate each other. To investigate whether non-mimicry between S1 & S2 and mimicry between S1 & S3 is true over time, temporal variations of mimicry were then measured.

3.1.2. Measuring temporal variations of mimicry

Figure 3 gives an example of I-series obtained for each parameter for the interaction S1/S2.

Figure 3: Representation of (I) for f0-min (blue), f0-max (green), f0-median (black), f0-max/min (purple), rms-Int (red), sd-Int (orange), npauses (pink) and dpauses (yellow) obtained for each moving window (Interaction S1/S2).

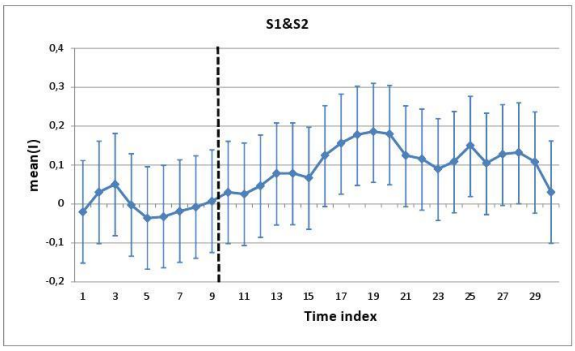


For interaction S1/S2, it is shown that S1 & S2 tend to imitate each other's voice intensity (rms-Int), which enables to define one phase of mimicry in this interaction. For the interaction S1/S3, speakers tend to mimic their speech in terms of f0-max and dpauses, which enables to define two phases of mimicry ($p<0.05$).

In order to reduce error bars amplitude, mean(I), i.e. (I) obtained from the set of the 8 prosodic parameters, was also calculated.

Temporal variations of mean(I) for the interaction S1/S2 show a trend towards mimicry from point 9 towards the end (**Figure 4**), the mimicry strength being significant from point 17 to 20. Temporal variations of mean(I) for interaction S1/S3 enables to detect one phase of mimicry, from the beginning of the interaction to point 9 ($p<0.05$). It also suggests a phase of mimicry from point 23 to the end, however this does not reach significance.

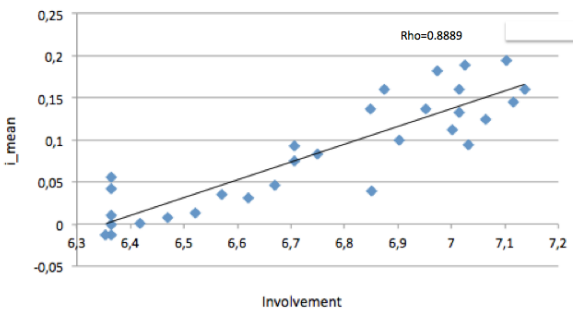
Figure 4: Representation of mean(I) obtained for each moving window for S1 and S2 interaction.



3.2. Mimicry, agreement and involvement

Results show that mimicry and level of agreement are neither correlated for the interaction S1/S2 (p=0.2398) nor for the interaction S1/S3 (p=0.2917). However, we found that involvement is strongly correlated to mimicry strength for the interaction S1/S2 (Rho=0.8889; p = 0.0013). **Figure 5** illustrates the correlation between the dynamics of involvement (smoothed at the scale of the window analysis of 200 seconds) and mimicry strength (mean(I)).

Figure 5: Correlation between the dynamics of involvement and mimicry strength (mean(I)).



4. DISCUSSION

Our results show that prosodic cues can be used to measure and detect mimicry in speech. They also support the assumption that mimicry is not a linear phenomenon but rather dynamic. In a cooperation task, such as described in Kousidis, et al [7], it can be hypothesized that speakers adapt each other's speech prosody linearly because they are cooperating throughout the interaction. In a spontaneous non-directed conversation such as in our data, we rather expect to find phases of mimicry and phases of non-mimicry. Investigating mimicry on a whole interaction therefore does not facilitate the measurement of temporal dynamics of mimicry.

Moreover, results show that levels of agreement are not correlated with mimicry strength. These results confirm Garrod's [4] assumption that "*presumably* [mimicry] *is not limited to cases where interlocutors are in agreement*". For a future study we plan on refining (and evaluating) the agreement annotation schema taking into account different degrees of agreement.

Finally, degrees of involvement are shown to be correlated with mimicry in speech; the more the speakers are involved in the interaction, the more they tend to mimic their speech prosody. We therefore argue that the absence or presence of mimicry in speech prosody can serve as a cue for the detection of degrees of involvement in spontaneous conversation. A point that we leave for future work is to determine whether speakers really tend to mimic (consciously or unconsciously) each other's speech, and if they do which speaker tends to mimic the other, or rather they tend to use the same prosodic patterns independently to convey the same functions (e.g. discourse).

5. CONCLUSION

This study has shown that the dynamics of mimicry in conversational speech can be measured by means of prosodic cues. We found that the higher the degree of involvement, the higher the strength of mimicry. Our study therefore supports the claim that mimicry in speech is part of social interaction and that it should be implemented into spoken communication systems.

6. ACKNOWLEDGEMENTS

The authors like to thank Prof. Petra Wagner and Dr. Fred Cummins for their assistance.

7. REFERENCES

[1] Boersma, P., Weenink, D. 2011. Praat: Doing phonetics by computer. *http://www.praat.org/*.

[2] Burgoon, J.K., Stern L.A., Dillman, L. 1995 *Interpersonal Adaptation: Dyadic Interaction Patterns.* Cambridge University Press.

[3] Edlund, J., Herldner, M., Hirschberg, J. 2009. Pause and gap length in face-to-face interaction, *Interspeech* 2779-2782.

[4] Garrod, S., Doherty, G. 1994. Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition* 53, 181-215.

[5] Giles, H., Coupland, N., Coupland, J. 1992. Accommodation theory: Communication, context and consequence. In Giles, H., Coupland, N., Coupland, J. (eds.), *Contexts of Accommodation*. Cambridge University Press, 1-68.

[6] Jaffe, J., Beebe, B., Feldstein, S., et al. 2001. *Rhythms of Dialogue in Infancy: Coordinated Timing in Development*. Blackwell Publishing.

[7] Kousidis, S., Dorran, D., McDonnell C., Coyle, E. 2008. Times series analysis of acoustic feature convergence in human dialogues. *Interspeech*.

[8] Oertel, C., Cummins, F., Campbell, N., Edlund, J., Wagner, P. 2010. D64: A corpus of richly recorded conversational interaction. *LREC*, 27-30.

[9] Pardo, J.S. 2006. On phonetic convergence during conversational interaction. *JASA* 119(4), 2382-2393.