Measuring speakers' similarity in speech by means of prosodic cues: methods and potential

Céline de Looze, Stéphane Rauzy

Measuring speakers' similarity in speech by means of prosodic cues: methods and potential

Céline De Looze[1], *Stéphane Rauzy*[2]

[1]Speech Communication Laboratory, Trinity College Dublin, Ireland
[2]Laboratoire Parole et Langage, Aix-Marseille Université, UMR 6057, Aix-en-Provence, France
deloozec@tcd.ie, stephane.rauzy@lpl-aix.fr

Abstract

This study presents a method for measuring speakers similarity (the tendency for speakers to exhibit similar speech patterns) by means of prosodic cues. It shows that similarity changes throughout social interaction and that its variations can inform about speakers' attitudes, similarity being more important when speakers are more involved in the interaction. It supports the assumption that similarity is part of social interaction and may be implemented into spoken dialogue systems.
Index Terms: synchrony, convergence, multi-modality, prosodic cues, involvement

1. Introduction

People interacting in a conversation have been observed to exhibit similar speech patterns in terms of speech sounds, syntax, lexicon and prosody (for a review see [1]). In most studies, speech similarity is described as a linear phenomenon. Burgoon et al [2], for instance, define it 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 similarity increases over time. However, it can be assumed, especially in spontaneous speech, that similarity tends to vary throughout the conversation, resulting in phases of similarity and phases of no-similarity and that its dynamic changes participate in making an interactive dialogue natural.

Relying on the assumption that similarity increases over time, most of the methodologies developed failed to identify the dynamics of similarity, with the exception of Edlund et al [3]. In this study, where the temporal variations of similarity are described in terms of synchrony and convergence and are measured in terms of pauses length (within-speaker silences) and gaps (between-speaker silences), it is shown that similarity between speakers is not indeed a global phenomenon and that it can be modeled dynamically.

In this paper we propose to measure similarity in speech for the whole interaction but also at various points of the conversation. Similarity is measured by means of prosodic cues (i.e. pitch level and span, voice intensity level and variation, mean pause duration and number of pauses). In fact, as [4], we assume herein that similarity is the result of different phenomena at different levels (prosodic, lexical, syntaxic, visual, etc.) and that only a full description of these different levels can capture its temporal dynamics. In this work, we propose a full description at the prosodic level where similarity is measured with regards to three prosodic parameters (i.e. f0, intensity and duration) and argue that the method developed can be applied to other levels.

Many terms have been used to describe the phenomenon of similarity such as accommodation, alignment, convergence, entrainment, synchrony, terms most of the time being bounded to a specific theory. In this work, we propose to define similarity not according to a particular framework but rather according to what will enable us to better describe, measure and as a long-term goal, model it. As in [3], we propose that synchrony and convergence are phenomena underlying similarity. Synchrony is defined as a situation when two speakers exhibit similar speech patterns, their speech variations resulting in two parallel channels and convergence as the situation when conversational partners' speech converge toward a common point (figure 1).

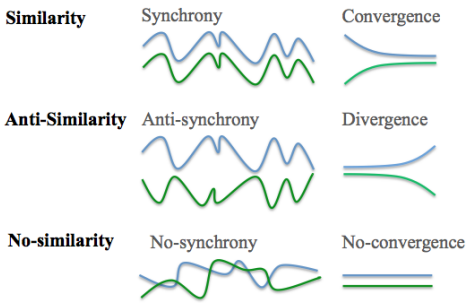

Figure 1: Phenomena underlying similarity

Going further, we also propose to consider anti-similarity, also divided into two underlying phenomena: Anti-synchrony is the tendency for speakers to differentiate their speech from the other's, resulting in mirror or anti-correlated patterns, and divergence, their tendency to move apart towards different directions. A last state considered is no-similarity, i.e. the situation when speakers neither exhibit synchrony, anti-synchrony, convergence nor divergence. Drawing an analogy with coupled oscillators model found in Physics, we assume that these underlying phenomena can be exhibited individually or in combination, resulting in 7 possible different states:

- 3 states of similarity (synchrony, convergence or both synchrony and convergence)
- 3 states of anti-similarity (anti-synchrony, divergence or both anti-synchrony and divergence)
- 1 state of no similarity (no synchrony and no convergence)

When we investigate the temporal dynamics of similarity, an underlying question that can also be raised is why similarity is interrupted or begins at certain points of the conversation. In this paper, we investigate whether the temporal variations of similarity are correlated with the speakers' degree of involve-

ment, assuming that similarity could be used as a cue to inform about speakers' attitudes. We argue that taking into account the temporal dynamics of mimicry improves the modeling of social interaction and hence spoken dialogue systems.

2. Experiment

2.1. Data

This study was based on the D64-corpus [5]. 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. Conversations were held in English.

2.2. Segmentation and measurements

The prosodic parameters under investigation are pitch level and span, voice intensity and number and mean duration of pauses. Acoustic measurements were obtained using the phonetic software Praat [6]. 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). The intensity of the voice was expressed as the root mean square (RMS) amplitude (rms-Int) and standard deviation Intensity (sd-Int). Silent pauses were detected automatically and corrected manually. Filled pauses, laughters and overlaps were excluded from the analyses.

2.3. Prosodic cues extraction

The difficulty encountered when measuring speech similarity is that it is not time-aligned. To resolve this, the TAMA (Time-aligned moving average) method, as proposed by Kousidis et al [7], was applied. Average values of prosodic cues were automatically extracted from a series of overlapping windows (frames) 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. Average values were calculated proportional to the utterances' length within the window, i.e. average values correspond to weighed means. Figure 2 shows the moving window along speakers interaction (represented by a conversation chart).

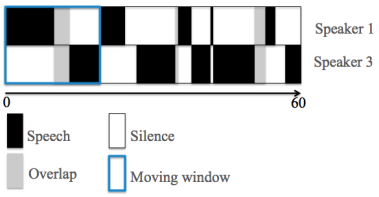

Figure 2: Parts of the conversation chart for speakers 1 and 3 interaction.

Average values were plotted to visually investigate whether speakers tend to exhibit similar speech patterns. Figure 3 represents the two TAMA times series of f0-median average values obtained for S1 and S3.

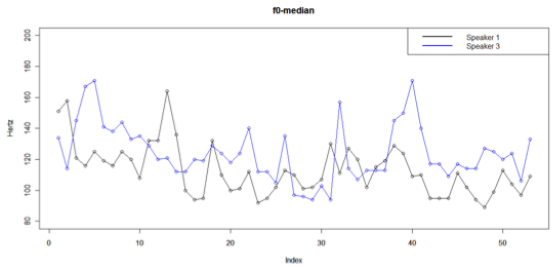

Figure 3: TAMA time series of f0-median average values (represented here in Hertz) obtained for S1 and S3 interaction (17 minutes).

2.4. Similarity measurement and significance

2.4.1. Analogy with coupled oscillators model

As mentioned in the introduction, similarity phenomena is envisaged herein by drawing an analogy with coupled oscillators model found in Physics (see for example [8]). This enables to measure the 7 states underlying similarity.

The model describes the dynamics of two oscillators (say two masses, each of them tied by a spring to a wall) coupled together by a spring. The spring plays here the role of a force coupling the respective oscillating trajectory x_1 and x_2 of the two masses.

The mathematical treatment of the problem let emerge two oscillating modes (normal modes). The symmetric mode associated with the sum of the mass positions,

$$\text{sum} = x_1 + x_2$$

describes the external oscillations of the two masses system seen as a whole. The asymmetric mode, associated with the difference of the mass positions,

$$\text{diff} = x_1 - x_2$$

describes the internal oscillations at works inside the two masses system. These oscillations, which take place at time scale T_c shorter than their external counterparts T_s (i.e. $T_c < T_s$), account for the energy exchanged between the two masses. The dynamics of the system is fully determined by a linear combination of the two normal modes.

Synchrony and anti-synchrony measurement

The model allows us to describe a continuum of behaviors, from pure synchrony, i.e. the coupling induces a synchronisation of the two trajectories (that is the case if the amplitude associated to the asymmetric mode is null, $A_{\texttt{diff}} = 0$), to pure anti-synchrony, i.e. the masses are forced to move in opposite direction (the amplitude of the symmetric mode is null, $A_{\texttt{sum}} = 0$). For an oscillating time series, the variance of the series is proportional to the square of the amplitude. The synchrony strength S can thus be measured by computing the ratio between the difference of variance of the two modes and the total variance of the system:

$$S = (\text{var(sum)} - \text{var(diff)}) / (\text{var(sum)} + \text{var(diff)})$$

where the variance is computed over a window time scale greater than the characteristic period T_s of the sum time series. The synchrony strength S is proportional[1] to the Pearson's

[1] It can be shown that the equality holds if $\text{var}(x_1) = \text{var}(x_2) = 1$.

correlation coefficient $\rho_{\texttt{pearson}}(x_1, x_2)$ of the two time series. It is expected to be significantly positive during synchrony phase, negative for anti-synchrony phase, and null if the amplitudes of the symmetric and asymmetric modes are equally distributed or if the oscillators are not coupled.

Convergence and divergence measurement

In addition to synchrony behavior, convergence effects will correspond to variations of the respective equilibrium position x_{10} and x_{20} of the two oscillators. Three cases must be considered, depending on the length l of the coupling string with respect to the distance d separating the two equilibrium positions of the masses in the absence of the spring. If the spring length l is lower than the distance d, the spring will act as an attractive force which will move closer the two equilibrium positions x_{10} and x_{20}. On the contrary, if l is greater than d, a repulsive force tends to push aside the two equilibrium positions, the third case being l equals d which let unchanged the equilibrium position of the two masses. By varying the spring length over time, one obtains a mechanism which depicts the possible transitions of the coupled system between the three following states, or phases,

- phase of convergence: $l < d$, x_{10} and x_{20} move closer,
- phase of stability: $l = d$, x_{10} and x_{20} remain unchanged,
- phase of divergence: $l > d$, x_{10} and x_{20} move aside.

A measure of the convergence strength C will thus consist in analysing the variation of the diff time series, averaged over a window time scale greater than the characteristic period T_s of the sum time series. Edlund et al. [3] suggest that the following quantity can be used:

$C = \rho_{\texttt{pearson}}(\text{abs(diff)}, t)$

where t are the times corresponding to x_1 and x_2 observations and the correlation coefficient is computed in window time scale greater than T_S. The convergence strength is varying from -1 (full convergence) to $+1$ (full divergence) and is expected to be null when stability phases are reached or if the oscillator are not coupled.

Here the analogy breaks off between coupled oscillators and speakers similarity. It has been helpful in identifying two time scales (T_S associated with the characteristic period of the sum time series and T_c associated with the diff time series), and at selecting the pertinent phases of similarity presented in the introduction. In particular, the analogy does not pretend to explain how the speakers dynamically perform phase transitions.

2.4.2. Application to the TAMA method

This in mind, we will consider hereafter that our two TAMA time series (described in 2.3), which represent the variation of a given prosodic parameter for both speakers, can be seen as the mass positions x_1 and x_2 of the coupled oscillators.

First, the synchrony strength S and the convergence strength C are computed for the whole interaction. Then, they are calculated at some parts of the conversation in order to detect transitions between phases of similarity and dissimilarity along the interaction. This is done by measuring S and C within moving windows of time scale greater than the characteristic period T_s of the sum time series.

The convergence strength C is directly computed from the value of the parameter. For the synchrony strength S however, the two times series are normalized (z-score transformation) in such a way that their mean equals zero and their variance is unit within each window:

$\text{mean}(x_1) = \text{mean}(x_2) = 0$; $\text{var}(x_1) = \text{var}(x_2) = 1$

After such a normalization step, the measure of the synchrony strength is equals to the Pearson's correlation coefficient of the two time series, i.e.$S = \rho_{\texttt{pearson}}(x_1, x_2)$.

We propose a confidence level interval for both the estimate of the S and C, by applying the Fisher's transformation to S, and similarly to C:

$F(S) = \text{arctanh}(S) = 1 / 2 * \ln((1+S) / (1-S))$

which approximately follows a normal distribution of mean $F(S0)$ ($S0$ is the value of S under the null hypothesis) and a standard deviation 1/SQRT(Neff-3), with Neff the effective number of independent data within the window for one of the series. Therefore, the z-score variable

$z = (F(S) - F(S0)) * \text{SQRT(Neff-3)}$

allows to decide whether the value of S and C is statistically significant or rather due to sampling fluctuations.

2.5. Annotation of involvement

The data was 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 [5]. 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. In this study, the correlation between degrees of involvement and similarity strength was investigated for interaction S1/S2.

3. Results

3.1. Synchrony, anti-synchrony, no-synchrony

3.1.1. For the whole interaction

A synchrony effect is detected for S1 and S3 on the whole interaction. S1 and S3 exhibit similar speech patterns in terms of voice intensity level and variation (rms-Intensity, p=0.00554 & sd-Intensity, p=2.81271e-06), mean pause duration (dpauses, p=0.01786) as well as the ceiling of their pitch range (f0-max, p=0.01426). For the whole S1/S2 interaction, the synchrony strength of each individual parameter does not reveal a statistically significant synchrony trend. This non detection can reflect that no synchrony is present in the interaction or alternatively that synchrony operates between S1 and S2, but only during a short phase of the whole interaction.

3.1.2. Temporal variations of synchrony

The temporal variations of synchrony is investigated by computing the synchrony strength within the time moving window of size 20 mentioned in the previous section. For the interaction S1/S3, speakers show similar variations in terms of f0-max, f0-span, rms-Intensity, sd-Int and dpauses and exhibit a long phase of synchrony ($p<0.05$). Figure 4 gives an example of S-series obtained for the parameters f0-max, f0-span, rms-Intensity, sd-Int and dpauses. In order to reduce error bars amplitude, mean synchrony strength (mean(S)), obtained from the set of the 3 prosodic parameters (f0, intensity and duration), was also calculated. Figure 5 shows the variation of mean(S) with time, which allows to define one long phase of synchrony, from point 9 to 32 ($p<0.05$).

Similar results are found for interaction S1/S2. S1 & S2

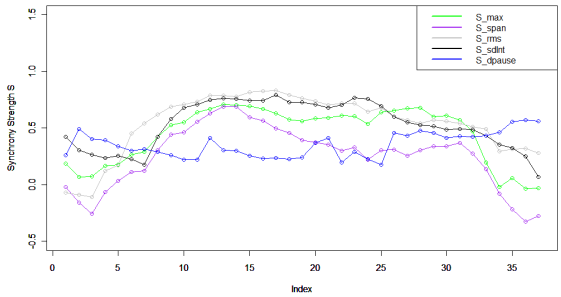

Figure 4: Representation of S for f0-max, f0-max/min, rms-Int, sd-Int and dpauses obtained for each moving window (Interaction S1/S3).

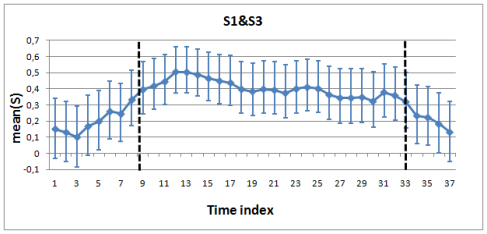

Figure 5: Representation of mean(S) obtained for each moving window for S1 and S3 interaction.

exhibit similar patterns in terms of voice intensity (rms-Int), during a shorter period roughly located at the middle of the interaction. The computation of mean(S) allows to locate more precisely this synchrony phase (from moving time window 16 to 25, not shown).

3.2. Convergence, divergence, no-convergence

No convergence was found for the whole conversations, for interaction S1/S2 as well as for S1/S3. The analyses on temporal variations neither allow to detect phases of convergence or divergence in interactions S1/S2 and S1/S3.

3.3. Synchrony and degrees of involvement

Since synchrony and not convergence was detected in our data, only the correlation between synchrony and involvement was investigated. This was done for interaction S1/S2. Results show that involvement is strongly correlated to synchrony strength S (Rho=0.9052; p = 3.55884e-05). More specifically, the higher the degree of involvement, the more speakers display similar speech patterns in terms of Rms-Intensity (r=0.89), dpause (r=0.89), npauses (r=0.59), f0-min (r=0.55), f0-span (r=0.51) and f0-median (r=0.42), S being stronger in terms of intensity of the voice.

4. Discussion

Our results show that prosodic cues can be used to measure and detect similarity in speech. They also support the assumption that similarity does not increase over time but is rather dynamic where the interaction is marked by phases of similarity and dissimilarity.

It is shown in this data that speakers exhibit parallel prosodic patterns rather than tending to converge towards common prosodic values. The fact that they do not display convergence may be explained either by physiological constraints or by the fact that the prosodic features of the speakers are intrinsically similar. For example, the fact that we do not find any convergence in interaction S1/S2 may be due to gender differences. Converging in this interaction (composed of a female and a male speaker) may require too much vocal effort. As for interaction S1/S3, there is no 'need' to converge since the two male speakers have a similar voice level. From these results, it can be assumed that depending on the features under investigation (i.e. speech sounds, prosody, syntax, lexicon, visual, etc.), speakers may exhibit either synchrony, or convergence or both. This justifies the 7 similarity states defined herein. It can also be hypothesized that speakers' synchrony may be linked to the informational structure and hierarchical organisation of discourse.

The fact that speakers' similarity is expressed by specific prosodic cues underlines the complexity of investigating speakers' strategies in communicating and suggests the use of different strategies. Investigating other levels and analysing more data will allow us to better understand how and what this phenomenon is used for in social interaction.

Finally, this study has shown that degrees of involvement are correlated with similarity in speech; the more the speakers are involved in the interaction, the more they tend to exhibit similar speech prosody. We therefore argue that the presence or absence of similarity in speech prosody can serve as a cue for the detection of degrees of involvement in spontaneous conversation.

5. Conclusions

This study has shown that the dynamics of similarity in conversational speech can be measured by means of prosodic cues. It reports that similarity does not increase over time but rather changes throughout social interaction. We also found that the higher the degree of involvement, the higher the strength of similarity. Our study therefore supports the claim that similarity in speech is part of social interaction and that it should be implemented into spoken communication systems.

6. References

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

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

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

[4] Guardiola, M. "Contribution multimodale l'tude de phnomnes de convergence dans l'interaction en face face", PhD Dissertation. In Progress.

[5] Oertel, C., Cummins, F., Campbell, N., Edlund, J. and Wagner, P., "D64: A Corpus of Richly Recorded Conversational Interaction", 27-30, LREC 2010.

[6] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer", 2011.

[7] Kousidis, S., Dorran, D., McDonnell C., Coyle, E. "Times Series Analysis of Acoustic Feature Convergence in Human Dialogues". Interspeech. 2008.

[8] Pain, H.J., "The Physics of Vibrations and Waves", 6th Edition, Wiley, SBN 978-0-470-01296-3, April 2005.