Measuring mimicry in task-oriented conversations: degree of mimicry is related to task difficulty

Vijay Solanki\textsuperscript{1}, Alessandro Vinciarelli\textsuperscript{2}, Jane Stuart-Smith\textsuperscript{1} and Rachel Smith\textsuperscript{1}

\textsuperscript{1}Glasgow University Laboratory of Phonetics, University of Glasgow, Glasgow (UK)  
\textsuperscript{2}School of Computing Science, University of Glasgow, Glasgow (UK)  
V.Solanki.1@research.gla.ac.uk, Alessandro.Vinciarelli1@glasgow.ac.uk

Abstract

The tendency to unconsciously imitate others in conversations has been referred to as mimicry, accommodation, interpersonal adaptation, etc. During the last few years, the computing community has made significant efforts towards the automatic detection of the phenomenon, but a widely accepted approach is still missing. Given that mimicry is the unconscious tendency to imitate others, this article proposes the adoption of speaker verification methodologies that were originally conceived to spot people trying to forge the voice of others. Preliminary experiments suggest that mimicry can be detected using this methodology by measuring how much speakers converge or diverge with respect to one another in terms of acoustic evidence. As a validation of the approach, the experiments show that convergence (speakers becoming more similar in terms of acoustic properties) tends to appear more frequently when the Diapix\textsuperscript{UK} task requires more time to be completed and, therefore, is more difficult. This is interpreted as an attempt to improve communication through increased coherency.

Index Terms: mimicry, Hidden Markov Models, conversation analysis, Social Signal Processing

1. Introduction

During the last couple of decades, the computing community has made significant efforts towards automatic analysis and understanding of conversations, the “primary site of human sociality” [1]. Initially, the focus was on the automatic transcription of what people say. Such a task is particularly challenging in the case of spontaneous conversations because speech is punctuated by phenomena that are difficult to tackle for an Automatic Speech Recognition (ASR) system, namely disfluencies (hesitations, pauses, fillers, etc.), vocalizations (laughter, cough, etc.), paralanguage [2], etc. On the other hand, these phenomena have attracted significant attention in the last years because they can be interpreted as social signals, i.e., as the physical, machine detectable evidence of social and psychological phenomena that cannot be observed directly, but only inferred from the way people behave [3, 4].

This has allowed the development of computational approaches capable of analysing a wide spectrum of social and affective aspects of conversation, including emotions [5], social verticality (e.g., dominance [6] and roles [7]), conflict [8, 9], personality traits [10], etc. One phenomena that has attracted much attention is mimicry [11], i.e., the tendency of people involved in an interaction to converge towards common behavioural patterns, possibly including a similar way of speaking. The phenomenon has been named in different ways (see [12] for an extensive survey), including accommodation [13], interpersonal adaptation [14], synchrony (in particular when the convergence concerns temporal behavioural patterns) [15], etc. The different names account for different aspects of the phenomenon and different approaches to its investigation. However, what appears to be common to all points of view is that the phenomenon takes place when people tend to adopt similar behavioural patterns in an interaction.

Given that mimicry can be thought of as the unconscious tendency of people to imitate others, this paper proposes to address the problem of its detection by using speaker verification methodologies. The reason behind this choice is that these were originally developed to detect people imitating others for fraudulent purposes. In particular, the paper reports on preliminary experiments showing that a simple verification technique (see below) can detect conversation segments where speakers converge or diverge with respect to each other.

The key idea behind the approach is that the convergence towards a common way of speaking (possibly meaning that one of the speakers becomes more similar to the other) should not be measured locally, but over intervals of time long enough to let mimicry to emerge. For this reason, the approach includes two main steps (see Figure 1): The first is the application of speaker verification techniques at the level of individual words, measuring how similarly two speakers utter a given word. The second is the measurement of the correlation between the similarity at the level of the words and the time at which the words are pronounced. If the correlation is statistically significant and positive, it means that the two speakers tend to become, on average, more similar over time. To the best of our knowledge, this is the first work that adopts such an approach.

Preliminary experiments have been performed over a corpus of six dyadic conversations revolving around the Diapix\textsuperscript{UK} task (12 fully unacquainted participants in total) [16]. The results show that the approach detects statistically significant convergence or divergence between speakers a number of times larger than expected by chance. In particular, statistically significant effects are observed in 40\% of the analysis units considered in the experiments \(p\text{-value}=10^{-15}\). This seems to suggest that the observations are not the result of chance, but of actual mimicry phenomena involving the participants. As a validation, the experiments consider the relationship between the outcome of the detection process and the time needed to complete the Diapix\textsuperscript{UK} task. The results show that speakers tend to converge more when they need more time to complete a task, possibly indicating that mimicry is used as a means to improve collaboration when participants experience difficulties in addressing a task.

The rest of this paper is organised as follows: Section 2 describes the corpus used for the experiments, Section 3 outlines...
the approach, Section 4 reports on experiments and results and finally, Section 5 draws some conclusions.

2. The Data

The experiments comprising this work have been performed over 6 conversations between unacquainted individuals (12 participants in total). The data concerns the DiapixUK task [16]: individual members of participant pairs were presented with one of two slightly different versions of the same picture (e.g., a beach scene where the same person wears a black T-shirt in one version and a white T-shirt in the other one) and were required to spot all the differences in less than 15 minutes. Participants repeated the task for 12 different picture pairs where the order of the pictures presented was randomised for each participant pair. As a result, the corpus can be split into $6 \times 12 = 72$ non-overlapping intervals that were used as analysis units. On average, each of these intervals lasts for 8 minutes and 1 second for a total of 9 hours and 37 minutes.

All participants were females, born and raised in the Glasgow conurbation. The reasoning behind this condition was based around literature concerning gender and accent playing a major role in mimicry [17, 18]. The ages range between 19 and 65 with an average of 30.9. For three of the conversations participants were paired randomly whilst in the other three, they were paired based on personality similarity (all participants filled the BFI-44 questionnaire [19]) and attractiveness judgements made by the participants from the respective photographs of their potential experimental partners.

The experimental setting was designed to limit the role of non-verbal communication as much as possible. Participant pairs sat in a sound attenuated booth and were separated by a divider so that they could easily speak while being unable to see one another or each other’s presented picture. The participants were sitting roughly 30 cm from a computer screen presenting the DiapixUK task pictures, more or less at eye-level. The conversations were recorded with sampling rate 44.1 kH z using two AKG microphones (one per participant) designed to minimise background noise. The signals collected with the two microphones were combined into a single stereo recording, with individual speakers assigned to either the left or right channel.

The recordings were manually transcribed and automatically segmented into a total of 106, 466 words (9, 511 for training and 96, 955 for test).

3. The Approach

The key idea of the approach proposed in this work is that speaker verification techniques - originally conceived to detect fraudulent attempts to imitate others - are suitable to detect a phenomenon like mimicry that can be thought of as an unconscious attempt to imitate one’s interlocutor.

The main steps of the approach are depicted in Figure 1: The time interval corresponding to a particular pair of pictures (see Section 2) is segmented into words and these are split into two sets, namely those that have been uttered by speaker $A$ and those that have been uttered by speaker $B$ (the red and blue rectangles of the figure, respectively). Each word is then converted into a sequence of observation vectors (see Section 3.1) and, as a result, the process produces the sets: $\{X_i^{(A)}\}$ and $\{X_j^{(B)}\}$, where $X_i^{(A)} = \{x_i^{(A,1)}, x_i^{(A,2)} , \ldots , x_i^{(A,N_i)}\}$ and $N_i$ is the number of observations in word $i$ (a similar expression can be obtained for $B$ by simply changing the superscript).

A simple speaker verification technique is used to obtain a measure $d_i^{(A)}$ or $d_j^{(B)}$ of how much speaker $A$ is converging to $B$ or vice-versa (see Section 3.2). Once the measure has been obtained for all the words of the time interval corresponding to a specific pair of pictures, it is possible to detect mimicry by measuring the correlation between $t_i^{d_j}$, the time at which $A$ has uttered the $i$th word, and $d_i^{(A)}$ (similarly for the correlation between $t_j^{d_j}$ and $d_j^{(B)}$). If the correlation, is statistically significant, it is possible to say that one speaker is converging or diverging with respect to the other depending on the correlation’s sign (see Section 3.3). The correlation is calculated with the Spearman Coefficient, known to be less sensitive to possible outliers:

$$\rho(X,Y) = 1 - \frac{6 \sum_{i=1}^{N} d_i}{n(n^2 - 1)}$$

where $N$ is the number of pairs $(x_i, y_j)$ used to calculate the correlation between $X$ and $Y$, and $d_i$ is the difference between

- Picture 1
- Picture 2
- Picture 11
- Picture 12

Figure 1: Schematic of the approach. The words uttered during the conversation interval related to a specific picture were transcribed manually, automatically segmented and split into two groups, namely words uttered by $A$ (red rectangles) and words uttered by $B$ (blue rectangles). Each word is converted into a sequence of observation vectors (here, 12-dimensional MFCC vectors). For a given sequence of observation vectors, the distance measurement $d_i^{(A)}$ or $d_j^{(B)}$ are obtained using mixtures of Gaussians. The Spearman coefficient is used to measure the correlation between the distance measures and the time at which words have been uttered.
the rank of \( x_i \) and the rank of \( y_i \) in the sample data (after the \( x_i \)'s and \( y_i \)'s are arranged in ascending order).

### 3.1. Feature Extraction

In the experiments of this work, the words are converted into sequences of 12-dimensional MFCC vectors. Each vector is extracted from a 30 ms long analysis window and the delay between consecutive windows is of 10 ms. The MFCCs account for vocal tract properties and phonemes than for properties that a speaker can control (consciously or not) to imitate others (e.g. prosody). On the other hand, MFCCs have been shown to be successful in speaker verification and are a reasonable starting point for the preliminary experiments presented here.

### 3.2. Verification

Every conversation considered in the experiments of this work includes two speakers \( A \) and \( B \). The goal of this step is to estimate the following likelihood ratio:

\[
d_i^{(A)} = \frac{p(X_i^{(A)}|\Lambda_B)}{p(X_i^{(A)}|\Lambda_A)}
\]

where \( X_i^{(A)} \) is a sequence of observations extracted from a word uttered by \( A \), and \( \Lambda_A \) and \( \Lambda_B \) are models that account for speakers \( A \) and \( B \), respectively. The value of \( d_i^{(A)} \) is larger than 1 when word \( w_i \) is more likely to have been uttered by \( B \) than by \( A \) and smaller than 1 in the opposite case.

The expression above is the likelihood ratio typically applied in speaker verification problems [20]: if the value of \( d_i^{(A)} \) exceeds a threshold \( \theta \), it means that \( A \) and \( B \) can be considered to be the same person. However, the likelihood ratio is used differently in this work. Here \( A \) and \( B \) are known to be two different people and \( d_i^{(A)} \) can be thought of as a measure of how \( A \) is close to \( B \) in terms of acoustic evidence when uttering word \( w_i \). The expression of \( d_i^{(A)} \) can be obtained by simply switching \( A \) and \( B \) in the equation above. It is important to note that \( d_i^{(A)} \) can be estimated only using words uttered by \( A \) while \( d_i^{(B)} \) can be estimated using only words uttered by \( B \), hence the splitting of the words into two sets, one per speaker (see Figure 1).

The problem that remains open is how to estimate the probabilities involved in the equation above. For the work presented here, the models \( \Lambda_A \) and \( \Lambda_B \) are mixtures of Gaussians:

\[
p(X_i^{(A)}|\Lambda_A) = \prod_{k=1}^G \sum_{l=1}^{N_k} \pi_l \mathcal{N}(x_i^{(A)})|\mu_k, \Sigma_k|,
\]

where \( G \) is the number of Gaussians in the mixture, \( \pi_l \) is the coefficient of the \( l \)th Gaussian in the mixture (\( \sum_{l=1}^G \pi_l = 1 \)), and \( \mu_k \) and \( \Sigma_k \) are its mean and covariance matrix, respectively.

The use of the mixtures of Gaussians makes the approach word independent (the model is the same independently of the word being uttered). The main disadvantage is that the model does not take into account temporal aspects that might be important in the context of mimicry detection (e.g., when people imitate their respective intonations).

### 3.3. Mimicry Detection

Mimicry is not a deliberate attempt to imitate others, but a tendency to do so that typically results from other social and psychological processes such as mutual liking (people that like one another tend to imitate one another), social verticality (lower status people tend to imitate higher status ones), etc. For this reason, mimicry is more likely to be evident at the level of a conversation or, at least, at the level of a time interval long enough to let the phenomenon emerge.

For this reason, the mimicry detection approach proposed in this work consists of measuring the correlation between variables \( d_i^{(A)} \) and \( s_i^{(A)} \) (to see whether \( A \) converges to \( B \) or \( d_i^{(B)} \) and \( s_i^{(B)} \) (to see whether \( B \) converges to \( A \)). The rationale behind such a choice is that the correlation can measure whether the value of \( d_i^{(A)} \) and \( d_i^{(B)} \) tends to increase or decrease consistently as a conversation evolves. In particular, if the correlation is not statistically significant, it means that there is no actual tendency and, therefore, there is no mimicry. In contrast, if the correlation is statistically significant, then there is a tendency which corresponds to mimicry if the correlation is positive (meaning that \( d_i^{(A)} \) and \( d_i^{(B)} \) tend to increase) and to divergence otherwise.

### 4. Experiments and Results

The experiments presented here have been performed over the data described in Section 2. Since each conversation cannot be split into 12 segments corresponding to different pairs of pictures, the first segment of each conversation has been used for training the mixtures of Gaussians while the following 11 have been used for test purposes. Therefore, the test set includes 11 \( \times \) 6 = 66 segments that will be used as test units. The mimicry detection approach (see Section 3.3) requires one to consider the words uttered by the two speakers separately. Therefore, the total number of correlations to be estimated is 66 \( \times \) 2 = 132.

The approach requires one to set two main hyperparameters, namely the number \( G \) of Gaussians in the mixtures and the number \( D \) of MFCC coefficients in the observation vectors. In the experiments presented here, both parameters have been set arbitrarily (\( G = 10 \) and \( D = 12 \)) and no alternative values have been tested.

Given a set of pairs \( \{(d_i^{(A)}, s_i^{(A)})\} \) or \( \{(d_i^{(B)}, s_i^{(B)})\} \), it is possible to calculate the Spearman correlation coefficient. The main advantage of such a coefficient is that it is based on the ranking of the values and, therefore, it is more robust to possible outliers than the Pearson \( r \) typically adopted to estimate the correlation between variables. Figure 2 shows the correlation values for the 132 sets. The value of the correlation is statistically significant in 52 cases with confidence level less than or equal to 0.05 and the probability of getting such a result by chance is \( 10^{-15} \) according to a two-tailed binomial test. In particular, the correlation is statistically significant with confidence level 0.01 in 30 cases (\( p = 10^{-16} \) according to a two-tailed binomial test).

One of the main limitations of most approaches aimed at detecting mimicry is that they are difficult to validate, i.e. it is difficult to verify whether they are actually measuring the tendency of people to imitate one another or not. A possible validation approach is to ask a pool of observers to judge the conversation in terms of how much the interactants mimic each other. The agreement between the outcome of the measurement process and the judgements of the observers becomes then the criterion to validate the measurement process. Another approach is to verify whether there is a relationship between the outcome of the measurement process and some other, measurable aspects of the interactions under exam. This article adopts the latter approach and, in particular, analyses the relationship between the convergence of divergence of speakers and the amount of time
the speakers involved in task-oriented conversations converge speaker verification techniques allow one to measure whether This article has shown preliminary experiments where simple measurement approach proposed in this work, might be one of their counterparts. In other words, mimicry, as detected by the observation is that mimicry tends to take place when the partici- longer, on average, when at least one of the two speakers tends more frequently than would be expected by chance, but also that the convergence tends to be associated with tasks that re- when the speakers experience difficulties in addressing a task and, therefore, need a higher degree of coordination. As future work, the baseline approach adopted for the experiments can be improved in several ways. The first is to change the type of acoustic evidence used to represent the speech signals. MFCCs have been shown to be effective in speaker verification because they account for vocal tract prop- eries, but they do not account for a number of speech properties that people can adopt to mimic others (e.g., intonation, loud-ness, speaking rate, etc.). The second is to improve the models adopted for the speaker verification step. The experiments of this work are based on mixtures of Gaussians, one of the simplest forms of Hidden Markov Model (only one state) [21]. Pos-ible improvements can be achieved by using models that have more than one state (thus taking into account temporal aspects) or are word dependent. A possible further direction for future work is to study the relationship between the measurements obtained with the approach proposed in this work and social / psychological phe-nomena of interest in a conversation (e.g., the personality traits of the speakers, their interpersonal attraction, etc.). It would also be useful to evaluate perceptual judgements of mimicry made by participants listening to the speech produced in this study. Investigations of this sort might further validate the mimicry detection approach while providing additional insights about the use of the phenomenon.

5. Conclusions

This article has shown preliminary experiments where simple speaker verification techniques allow one to measure whether the speakers involved in task-oriented conversations converge or not towards their interlocutors in terms of acoustic evidence that can be extracted from speech recordings. As a validation of the methodology, the experiments show not only that converg-ence or divergence with respect to the interlocutors appear more than one state (thus taking into account temporal aspects) or are word dependent. A possible further direction for future work is to study the relationship between the measurements obtained with the approach proposed in this work and social / psychological phe-nomena of interest in a conversation (e.g., the personality traits of the speakers, their interpersonal attraction, etc.). It would also be useful to evaluate perceptual judgements of mimicry made by participants listening to the speech produced in this study. Investigations of this sort might further validate the mimicry detection approach while providing additional insights about the use of the phenomenon.

6. References


