Measuring Mimicry in Task-Oriented Conversations: The More the Task is Difficult, The More we Mimick our Interlocutors

Vijay Solanki, Jane Stuart-Smith, Rachel Smith and Alessandro Vinciarelli

University of Glasgow, Glasgow (UK)

firstname.lastname@glasgow.ac.uk

Abstract

The tendency to unconsciously imitate others in conversations is referred to as mimicry, accommodation, interpersonal adaptation, etc. During the last 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 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 (the speakers become more similar in terms of acoustic properties) tends to appear more frequently when a task is difficult and, therefore, requires more time to be addressed.

Index Terms: mimicry, Hidden Markov Models, conversational technologies, 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 to analyze 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 of the phenomena that has attracted most 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 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 of the approach is that the convergence towards a common way of speaking (possibly meaning that one of the spekers 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 UK scenario (12 subjects fully unacquainted with one another in total) [16]. The results show that the approach detects statistically significant convergence or divergence between speakers a number of times larger than how expected by chance. In particular, statistically significant effects are observed in 40% of the analysis units considered in the experiments (p-value=10^{-15}). This seems to suggest that the observations are not the result of chance, but of actual mimicry phenomena involving the subjects. As a validation, the experiments consider the relationship between the outcome of the detection process and the time needed to complete the tasks of the Diapix scenario. The results show that speakers tend to converge more when they need more time to complete a task, possibly meaning that mimicry is used as a means to improve collaboration when the subjects experience difficulties in addressing a task.

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

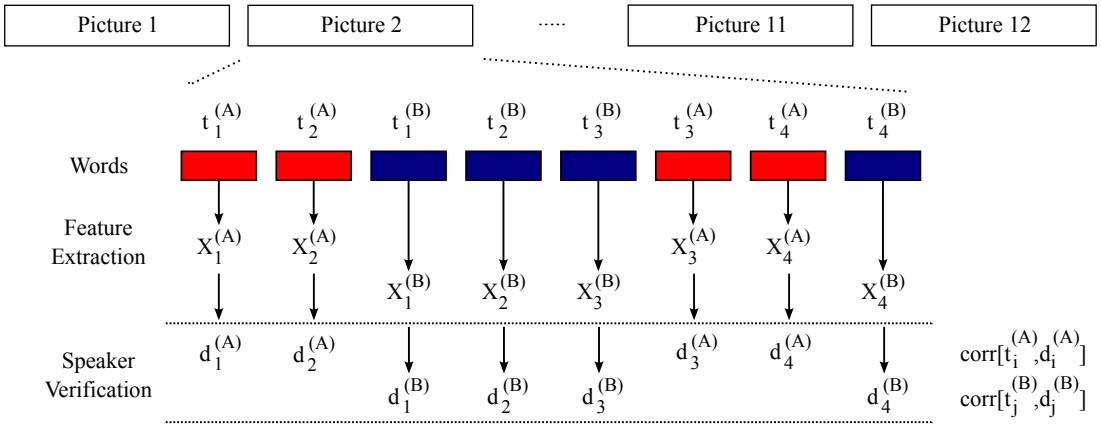

Figure 1: *Scheme of the approach. The words uttered during the conversation interval related to a specific picture are segmented manually and split into two groups, namely words uttered by A (red rectangles) and words uttered by B (blue rectangles). Each word is converted into a sequence of observation vectors (in the experiments of this work, 12-dimensional MFCC vectors). For a given sequence of observation vectors, the distance measurement $d_i^{(A)}$ or $d_i^{(B)}$ are obtained using mictures of Gaussians. The Spearman coefficient is used to measure the correlation between the distance measures and the time at which words have been uttered.*

2. The Data

The experiments of this work have been performed over 6 conversations between unacquainted individuals (12 subjects in total). The data revolves around the Diapix UK scenario [16]: two subjects watch 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 are expected to spot all the differences in less than 15 minutes. For each conversation, the subjects repeat the task 12 times for 12 different pairs of pictures. As a result, the corpus can be split into $6 \times 12 = 72$ non-overlapping intervals that will be used as analysis units. On average, each of these intervals lasts for 8 minutes and 1 seconds for a total of 9 hours and 37 minutes.

All subjects are female that were born and raised in Glasgow. The reason is that gender and accent play a major role in mimicry [17, 18]. The ages range between 19 and 65 for an average of 30.9. In three conversations, the subjects were paired randomly while in the other three they were paired based on personality similarity (all subjects filled the BFI-44 questionnaire [19]) and attractiveness judgments made over their respective pictures.

The experimental setting was designed to limit as much as possible the role of nonverbal communication. The two subjects sat in a sound attenuated booth and were separated by a divider so that they could easily speak while beeing unable to see one another. The subjects were sitting roughly 30 cm far from the computer screen showing the pictures of the Diapix task more or less at the height of their eyes. The conversations were recorded with sampling rate 44.1 kHz 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. The recordings have been manually 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 interlocutors.

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 $X_i^{(A)} = (\boldsymbol{x}_1^{(A,i)}, \boldsymbol{x}_2^{(A,i)}, \ldots, \boldsymbol{x}_{N_i}^{(A,i)})$ 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 viceversa (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 $t_i^{(A)}$, the time at which A has uttered the i^{th} word, and $d_i^{(A)}$ (similarly for the correlation between $t_k^{(B)}$ 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:

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

where N is the number of pairs (x_j, y_j) used to calculate the correlation between X and Y, and d_i is the difference between the rank of x_i and the rank of y_i in the sample data (after the x_is and y_is 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 more than for properties that a speaker can control to imitate, consciously or not, 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 in this work.

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:

$$d_i^{(A)} = \frac{p(X_i^{(A)}|\Lambda_B)}{p(X_i^{(A)}|\Lambda_A)} \quad (2)$$

where $X_i^{(A)}$ is a sequence of observations extracted from a word uttered by A, and Λ_A and Λ_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 θ, it means that A and B are the same person. However, the likelihood ratio is used differently in this work. Here A and B are known to be two different persons 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. In the case of this work, the models Λ_A and Λ_B are mixtures of Gaussians:

$$p(X_i^{(A)}|\Lambda_A) = \prod_{k=1}^{N_i} \sum_{l=1}^{G} \pi_l \mathcal{N}(\boldsymbol{x}_k^{(A,i)}|\boldsymbol{\mu}_l, \Sigma_l), \quad (3)$$

where G is the number of Gaussians in the mixture, π_l is the coefficient of the l^{th} Gaussian in the mixture ($\sum_{l=1}^{G} \pi_l = 1$), and $\boldsymbol{\mu}_l$ and Σ_l 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 like, e.g., 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 to emerge.

For this reason, the mimicry detection approach proposed in this work consists in measuring the correlation between variables $d_i^{(A)}$ and $t_i^{(A)}$ (to see whether A converges to B) or $d_j^{(B)}$ and $t_j^{(B)}$ (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 of this work have been performed over the data described in Section 2. Since each conversation can 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 analysis units. The mimicry detection approach (see Section 3.3) requires one to consider separately the words uttered by the two speakers. 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 of this work, both parameters have been set arbitrarily ($G = 10$ and $D = 12$) and no alternative values have been tested.

Given a set of pairs $\{(d_i^{(A)}, t_i^{(A)})\}$ or $\{(d_j^{(B)}, t_j^{(B)})\}$, it is possible to calculate the Spearman 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 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 mimick each other. The agreement between the outcome of the measurement process and the judgments 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, analyzes the relationship between the convergence (or the lack of it) of the speakers and the amount of time needed to spot all the differences between a pair of pictures (see Section 2 for the details of the Diapix UK task).

Table 1 shows the average amoumt of time needed to complete the task in the following three conditions:

- *Positive*: The correlation is positive and statistically significant with confidence level lower or equal to 0.05 for at least one of the speakers (left column).
- *Negative*: The correlation is negative and statistically

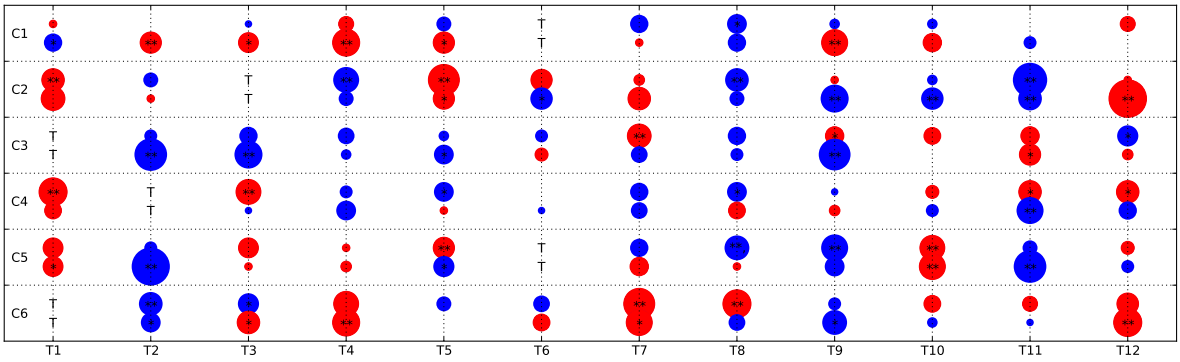

Figure 2: *For every conversations* Cn *(*$n = 1, \ldots, 6$*) and task* Tm *(*$m = 1, \ldots, 12$*), the bubble plot shows the correlation between* $d_i^{(A)}$ *and* $t_i^{(A)}$ *(upper row) and the correlation between* $d_i^{(B)}$ *and* $t_i^{(B)}$ *(lower row). The blue and red bubbles correspond to positive (convergence) and negative (divergence) correlations, respectively. Single and double asterisks correspond to statistical significance with* 0.05 *and* 0.01 *level of confidence respectively. The largest bubble corresponds to an absolute value of* 0.19 *while the missing bubbles correspond to null correlations. The letter "*T*" stands for* training*, meaning that that particular pair of pictures was the first to be used in the conversation and then was used to train the speaker models.*

Condition	Positive	Negative	Null
Avg. Length (s)	585 ± 51	427 ± 43	430 ± 33

Table 1: The table reports the average time ($\pm$ the standard error) required for completing a task in Positive, Negative and Null condition. The positive condition (at least one of the two spakers converges towards the other to a statistically significant extent) is associated to tasks that require longer time to be addressed.

significant with confidence level lower or equal to 0.05 for at least one of the speakers (central column).

- *Null*: The correlation is not statistically significant for both speakers (right colum).

The results show that the time required to complete the task is longer, on average, when at least one of the two speakers tends to converge to the other. A possible explanation of such an observation is that mimicry tends to take place when the subjects experience difficulties in completing the task (this is why it takes more time) and need to tighten the collaboration with their counterparts. In other words, mimicry, as per detected by the measurement approach proposed in this work, might be one of the means to make the interaction more effective when the task is more difficult.

5. Conclusions

This article has shown preliminary experiments where simple speaker verfication 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 convergence or divergence with respect to the interlocutors appear more frequently than how expected by chance, but also that the convergence tends to be associated with tasks that require more time to be addressed. The possible explanation of such an effect (see Section 4) is that mimicry tends to appear when the speakers experience difficulties in addressing a task and, therefore, need a higher degree of coordination.

As a 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. The MFCCs have been shown to be effective in speaker verification because they account for vocal tract properties, but they do not account for a number of speaking aspects that people can adopt to mimick others (e.g., intonation, loudness, 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]. Possible improvements can be achieved by using models that have more than one state (thus taking into account temporal aspects) or are word dependent.

Another possible direction for future work is to study the relationship between the measurements obtained with the approach proposed in this work and social / psychological phenomena of interest in a conversation (e.g., the personality traits of the speakers, their interpersonal attraction, etc.). This might further validate the mimicry detection approach while prividing further insights about the use of the phenomenon.

6. References

[1] E. Schegloff, "Analyzing single episodes of interaction: An exercise in conversation analysis," *Social Psychology Quarterly*, vol. 50, no. 2, pp. 101–114, 1987.

[2] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[3] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743–1759, 2009.

[5] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[6] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[7] H. Salamin and A. Vinciarelli, "Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody, and conditional random fields," *IEEE Transactions on Multimedia*, vol. 14, no. 2, pp. 338–345, 2012.

[8] S. Kim, M. Filippone, F. Valente, and A. Vinciarelli, "Predicting continuous conflict perception with Bayesian Gaussian Processes," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 187–200, 2014.

[9] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools," *Image and Vision Computing*, vol. 31, no. 2, pp. 203–221, 2013.

[10] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[11] J. Lakin, V. Jefferis, C. Cheng, and T. Chartrand, "The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry," *Journal of nonverbal behavior*, vol. 27, no. 3, pp. 145–162, 2003.

[12] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 349–365, 2012.

[13] H. Giles, J. Coupland, and N. Coupland, *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, 1991.

[14] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic interaction patterns.* Cambridge University Press, 1995.

[15] M. Chetouani, "Role of inter-personal synchrony in extracting social signatures: Some case studies," in *Proceedings of the Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges*, 2014, pp. 9–12.

[16] R. Baker and V. Hazan, "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods*, vol. 43, no. 3, pp. 761–770, 2011.

[17] D. Bulatov, "The effect of fundamental frequency on phonetic convergence," University of California at Berkeley, Phonology Lab Annual Report, 2009.

[18] M. Babel, "Dialect divergence and convergence in New Zealand English," *Language in Society*, vol. 39, no. 04, pp. 437–456, 2010.

[19] O. John, E. Donahue, and R. Kentle, "The Big Five Inventory - versions 4a and 54," *Berkeley: University of California, Berkeley, Institute of Personality and Social Research*, 1991.

[20] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.

[21] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.