Represent the Degree of Mimicry between Prosodic Behaviour of Speech Between Two or More People

Problem presented by

Joseph Anthony Connor

ExpertoCrede

ESGI107 was jointly hosted by
The University of Manchester
Smith Institute for Industrial Mathematics and System Engineering

with additional financial support from
KTN Ltd
Natural Environment Research Council
Manchester Institute for Mathematical Sciences

Report author

Matthew Arran (University of Cambridge), Graham Benham (University of Oxford), Liam Dempsey (Imperial College London), Elizaveta Dubrovina (Imperial College London), Roxana Feier (University of Oxford), John Fozard (University of Nottingham), Anna Lambert (UCL), Joseph Maestri (Imperial College London), Naoko Miyajima (Durham University), Tijana Radivojević (Basque Center for Applied Mathematics), Emily Riley (University of Cambridge)

Executive Summary

ExpertoCrede want people to be better understood and aims to improve the way we communicate with each other. The study group focused on finding a way to analyse a conversation in order to extract information on the level of rapport between the two speakers. Since such data may contain sensitive personal data, the study group put an emphasis on only using methods which could be run locally on a smartphone and avoided techniques that were computationally expensive or needed large datasets.

Firstly, the group considered the nature of turn taking within conversations where there was rapport between the participants. This was done by manually labelling conversations from the BBC Listening Project [1]. The data suggested an element of memorylessness in turn taking within friendly conversations, and therefore Markov chains were used to model conversations. Each person speaking was considered to be a state in the Markov chain, and the probability of the speaker switching was estimated from the data. The possibility of these probabilities changing within longer conversations was also considered, and some preliminary ODE models for this were suggested.

The group sought a way to distinguish between the two speakers and looked for a simple speaker identification algorithm which would be easy to implement on a mobile phone. Such a tool is necessary to enable all subsequent analysis of the sound data. Two algorithms were considered which provided reasonable level of information. Firstly, a low-frequency classification approach was used that took advantage of the natural difference in pitch of two speakers (especially in the case of a male-female conversation). The second approach utilised a Gaussian mixture model on the extracted Mel Frequency Cepstrum Coefficients.

Based on existing literature, conversational rapport was expected to correspond to mimicry in prosodic features of speech. To analyse this rapport, tools were developed to extract pitch, volume and speech rate from audio files, using Praat, a standard tool in academia, and custom-written Matlab codes. These tools were found to be broadly successful in extraction of these features. However, in the time available no consistently significant correlations or trends were found in natural high-rapport conversations from the BBC Listening Project, either over the course of a conversation or between the last few seconds of one speaker's speech fragment and the first few seconds of the next speaker's. Further work would involve following up on some potential correlations in such conversations, and in particular a comparison with low-rapport conversations.

Version 1.0
May 1, 2015
v+35 pages

Contributors

Matthew Arran (University of Cambridge)
Raphael Assier (University of Manchester)
Graham Benham (University of Oxford)
Bozena Deka (Polish Academy of Science)
Liam Dempsey (Imperial College London)
Elizaveta Dubrovina (Imperial College London)
Nabil Fadai (University of Oxford)
Roxana Feier (University of Oxford)
Thomas House (University of Manchester)
Anna Lambert (UCL)
Jane Lee (University of Oxford)
Joseph Maestri (Imperial College London)
Naoko Miyajima (Durham University)
Doireann O'Kiely (University of Oxford)
Colin Please (University of Oxford)
Tijana Radivojević (Basque Center for Applied Mathematics)
Emily Riley (University of Cambridge)
William Rowley (University of Manchester)
Zachary Wilmott (University of Oxford)

Contents

1 Problem Description

(1.1) ExpertoCrede is a company that wants people to be better understood and aims to improve the way we communicate with each other. The study group was tasked with identifying and quantifying the degree of rapport between two people in a conversation. Whilst methods have previously been developed which extract this from textual information, we were primarily interested in using just features of the sound of the two speakers' voices, such as pitch, rate, tone and turn-taking. Information about the level of rapport between two people is highly sensitive and so to avoid privacy concerns the study group focused on methods which could be run locally on a smartphone rather than any computationally expensive methods that require large datasets or use cloud computing.

(1.2) If we have a method to identify people who have rapport, we can use this in a tool to help people reach out to friends that they may not know they have. The group took two different approaches, one more local, and one more global. Both approaches rely on being able to separate the two speakers in a conversation. The group divided into three subgroups to address these issues; global turn taking analysis, speaker identification and local mimicry analysis.

(1.3) The focus of the global turn taking analysis was to examine the distribution of speech length at each 'turn' of the conversation. Extracts from the BBC Listening Project [1] were manually separated into speakers for the data to be analysed.

(1.4) Speaker identification focused on identifying characteristic frequencies in each persons speech in order to distinguish between two speakers. This is a challenging problem due to the prosodic nature of speech as people naturally alter the frequency of their speech as they talk, and so may become almost inseperable in frequency at times. However, several methods were explored and progress was made in each method.

(1.5) In order to develop some measure to mimicry, the local analysis group considered the change in certain aspects of speech (pitch, intensity, speed etc), and more importantly their change over time. It is hypothesised that as people get to know each other over the course of a conversation these aspects will converge to some extent.

2 Turn Taking Analysis

2.1 Exploratory Data Analysis

(2.1.1) The group wished to explore the nature of turn taking within a conversation where there was rapport. In particular, we were interested in the durations of each person's turns in the conversation, and whether these evolved throughout the conversation as they developed a rapport.

(2.1.2) The main data set we used to investigate this was the BBC Listening Project [1] which is a partnership between the BBC and the British Library. It consists of freely available recorded conversations between two people, typically family members or close friends. The topics of conversation are very broad, ranging from the secrets of a long happy marriage, to the diagnosis of a life threatening disease. However, they are almost all friendly conversations between two people who are very close, and therefore we expect rapport to be present.

(2.1.3) The audio files of these conversations were not labelled with the times of each person speaking, and initially we did not have a working speaker identification algorithm. Therefore the group manually labelled conversations with the times at which the speaker changed. Sixty conversation fragments between one and five minutes in length were labelled.

(2.1.4) An example conversation is shown in figure 1. The group noticed that it was very common for one person to be the dominant speaker and the other to make very short interjections frequently. These interjections were recorded only when they were at least a second in length.

(2.1.5) From this data we consider conversations as pairs of data

$$(a_i, b_i) \tag{1}$$

where a_i and b_i are the lengths of the ith speech fragments by person A and B respectively. We hypothesised that when a conversation has rapport the duration of adjacent speech fragments may show a positive correlation.

(2.1.6) However, figure 2 shows no obvious correlation between the length of adjacent speech fragments, contradicting our initial intuition. However, we notice an ‘L’ shape within the heat map, suggesting different levels of verbosity within the population.

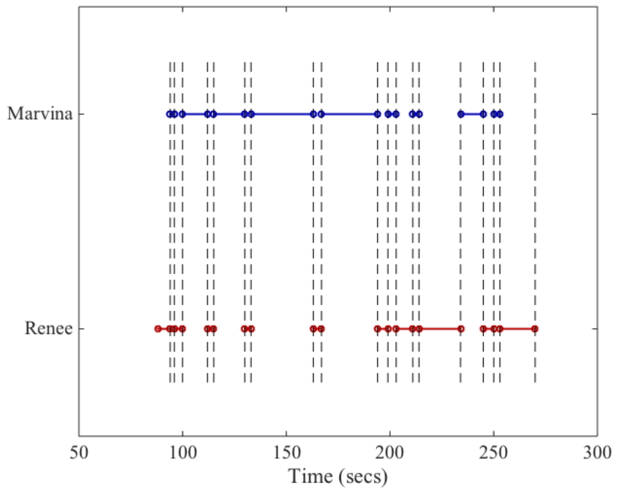


Figure 1: Turn taking within an example conversation from the Listening Project.

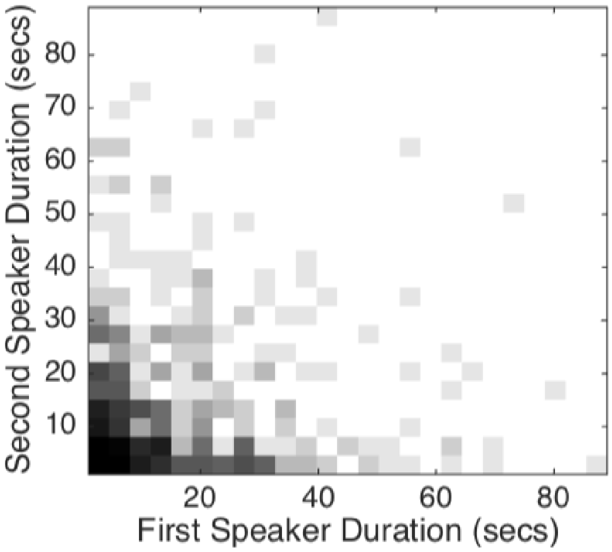


Figure 2: A heat map of pairs (a_i, b_i) from all sixty listening project conversations. A darker square corresponds to more data points within that square.

(2.1.7) To explore the levels of verbosity, we consider the probability distribution of the rate of stopping talking, ie. the reciprocal of the mean length of speaking. We assume it follows a gamma distribution, and fit it to the data from the Listening Project. This is illustrated in figure 3.

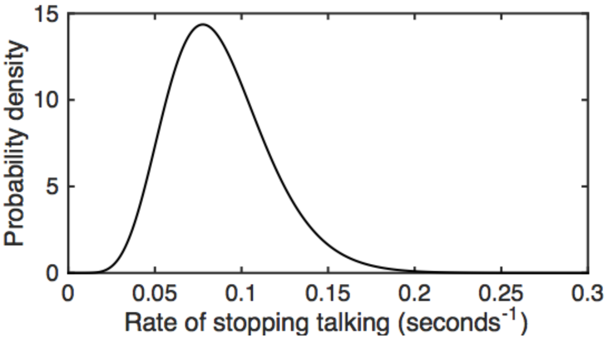


Figure 3: A gamma distribution for the rate of stopping talking fitted from the conversations from the Listening Project.

2.2 Motivation for Markov chain

(2.2.1) Using conversations from the Listening Project we estimated the cumulative probability distribution of the duration of speaking times and compared with exponential and gamma distributions. The very good fit with an exponential distribution led us to believe that there is most likely an element of memorylessness in turn-taking (figure 4 shows such a fit for an empirical cumulative probability distribution for a 40-min long conversation from the Listening Project). This suggests that modelling using Markov chains is a reasonable strategy.

(2.2.2) The mathematical assumption of memorylessness does not contradict our intuition on how mimicry would manifest as speaking times of two people modelled by a Markov chain still can converge/diverge as the conversation evolves. It actually means that the amount of time that one person takes in their conversation is not dependent on the former speaking times of their partner or themselves, but only on the current state of the conversation.

(2.2.3) Mathematically, a sequence of random variables $X_1, X_2, ...$ is called a Markov chain if

$$\mathbb{P}(X_{i+1} = x | X_1 = x_1, X_2 = x_2, ..., X_i = x_i) = \mathbb{P}(X_{i+1} = x | X_i = x_i),$$

i.e. if the conditional probability of the future state given the past depends only on the present.

(2.2.4) To further explore these findings we derived two mathematical models for the length of turn times in a conversation. The first includes a level of randomness with a stochastic differential equations approach, and the second considers the sequence of turn times as a Markov chain.

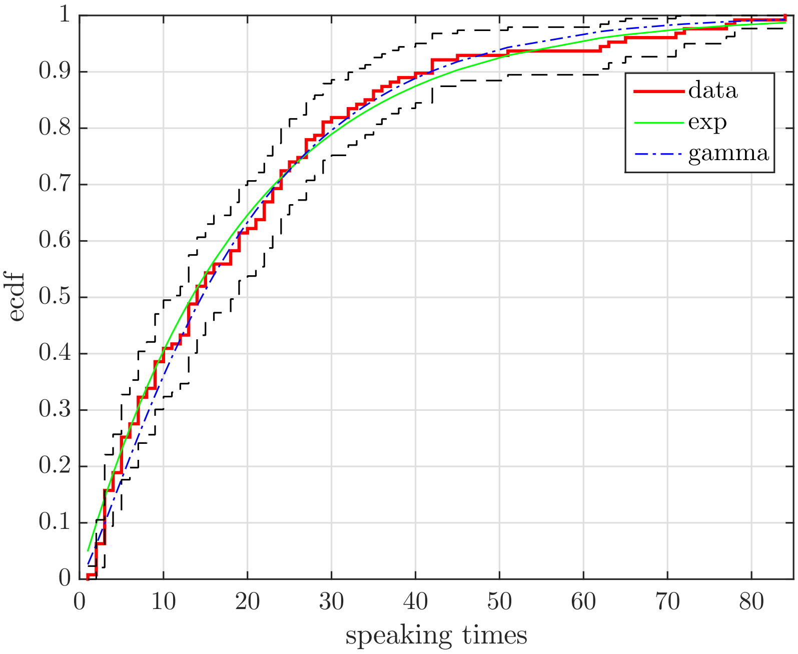


Figure 4: Empirical cumulative distribution function of speaking times with 95% lower and upper confidence bounds compared with exponential and gamma distributions.

2.3 Stochastic Differential Equations

(2.3.1) Let us consider a conversation which is made up of a series of exchanges between two participants. The turn taking in the conversation can then be thought of as a sequence of pairs,

$$(X_1, Y_1), (X_2, Y_2), \ldots$$

where (X_i, Y_i) are the durations of each turn of person X and Y respectively. The SDE approach was to model the evolution of these pairs so that they were governed partly by some mimicry and partly by a random process.

(2.3.2) The SDEs which we considered were of the form

$$\begin{aligned} dX_t &= \beta_1 \left(Y_t - X_t\right) dt + \sigma_1 f\left(X_t\right) dW_{1t}, \\ dY_t &= \beta_2 \left(X_t - Y_t\right) dt + \sigma_2 f\left(Y_t\right) dW_{2t}, \end{aligned}$$

where β_i are the mimicry coefficients, σ_i are the volatilities of the Wiener processes, W_i, and $f\left(X_t\right)$ is a function which controls the randomness such that we never jump to a negative conversation time. Therefore we require $f\left(X_t\right)$decreases with the size of X_t. A naive apprach would be to take it as linear but then we get the problem that at large conversation times, randomness grows unnaturally. A better model is to take it as a function which grows with X_t but has a cap as X_t gets very large. A suitable function would be,

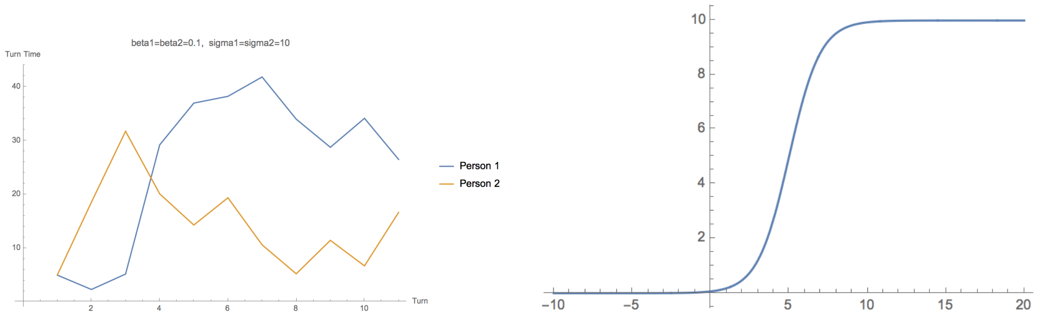
$$f\left(X_t\right) = \frac{A}{1 + e^{-X_t + \alpha}}, \tag{2}$$

where A and α are scaling and translating coefficients to be chosen suitably.

(2.3.3) Note, the best way to ensure that the conversations do not take negative times is to rewrite them as,

$$\begin{aligned} d\left(\log X_t\right) &= \frac{\beta_1}{X_t}\left(Y_t - X_t\right) dt + \frac{\sigma_1}{X_T} f\left(X_t\right) dW_{1t} \\ d\left(\log Y_t\right) &= \frac{\beta_2}{Y_t}\left(X_t - Y_t\right) dt + \frac{\sigma_2}{Y_t} f\left(Y_t\right) dW_{2t} \end{aligned}$$

(2.3.4) To explore these ideas further, one could attempt to characterise a conversation by simulating data many times and trying to fit the coefficients β_i and σ_i to the match the real data. This would then give us an insight into the levels of mimicry in a conversation, and thus perhaps also rapport.



(a) (b)

Figure 5: A simulated conversation using the stochastic differential equations model, in (a) 10 exchanges from the model and (b) decaying volatility function (2) with $A = 10$ and $\alpha = 5$.

2.4 Markov Chain Model

(2.4.1) Let us now consider a conversation between two individuals A and B. Assuming the sequence of turns (events of turn-taking) is a Poisson process[1], let us try to model the conversation such that the speaking time at time i depends only on the time at time $i-1$. Let us imagine that when person A is speaking, there is a probability of conversation switching to person B, $P_{AB} = a$ and likewise, when person B is speaking, the probability of switching to person A is $P_{BA} = b$. Considering there are only two possible states (only one person can be talking) we must conclude that the probabilities of each person continuing to speak are $1-a$ and $1-b$ respectively, see figure (6). The probabilities a and b can be interpreted as levels of willingness to let the other person speak or levels of indifference of person A and person B, respectively, depending on the type of the conversation. We would therefore expect a friendly conversation to have higher values of a and b than an unfriendly conversation.

[1] Poisson process counts the number of events and the time points at which these events occur in a given time interval. The sequence of inter-arrival times of consecutive events are independent and identically distributed exponential random variables.

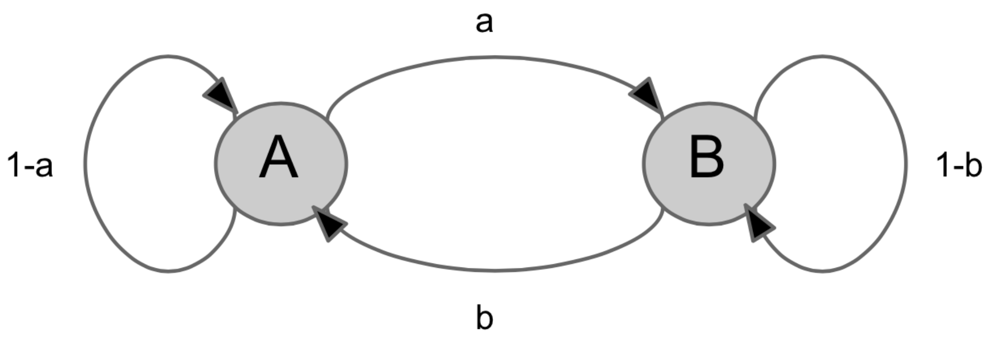


Figure 6: Conversation viewed as a Markov chain where person A or B is speaking with probability of switching, a or b, respectively.

(2.4.2) This way we can construct a Markov chain $X_1, X_2, ...$ on a state space $\mathcal{S} = \{A, B\}$ with the transition probability matrix

$$P = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}. \tag{3}$$

(2.4.3) Note, we could easily extend this model to a three or four state system, where the other states represent interruption or silence, i.e. a Markov chain with state space $\mathcal{S} = \{A, B, AB, O\}$ and a 4-dimensional square transition matrix P.

(2.4.4) The data we had available for analysis is measured in time intervals of no less than 1 second, hence the use of a discrete time Markov process, with discrete time intervals of one second, is natural. Thus after each second the person either continues to speak, or the conversation switches to the other person, with the probabilities described above.

(2.4.5) Let us for now consider a time-homogeneous Markov chain - whose transition matrix does not depend on time and is given by (3). Since the state space is finite, the chain is irreducible and aperiodic[2], from the theory of Markov chains we know that the equilibrium (stationary or steady state) probability distribution $\pi = (\pi_A, \pi_B)$ that the person A or person B is speaking is given by the fixed point equation

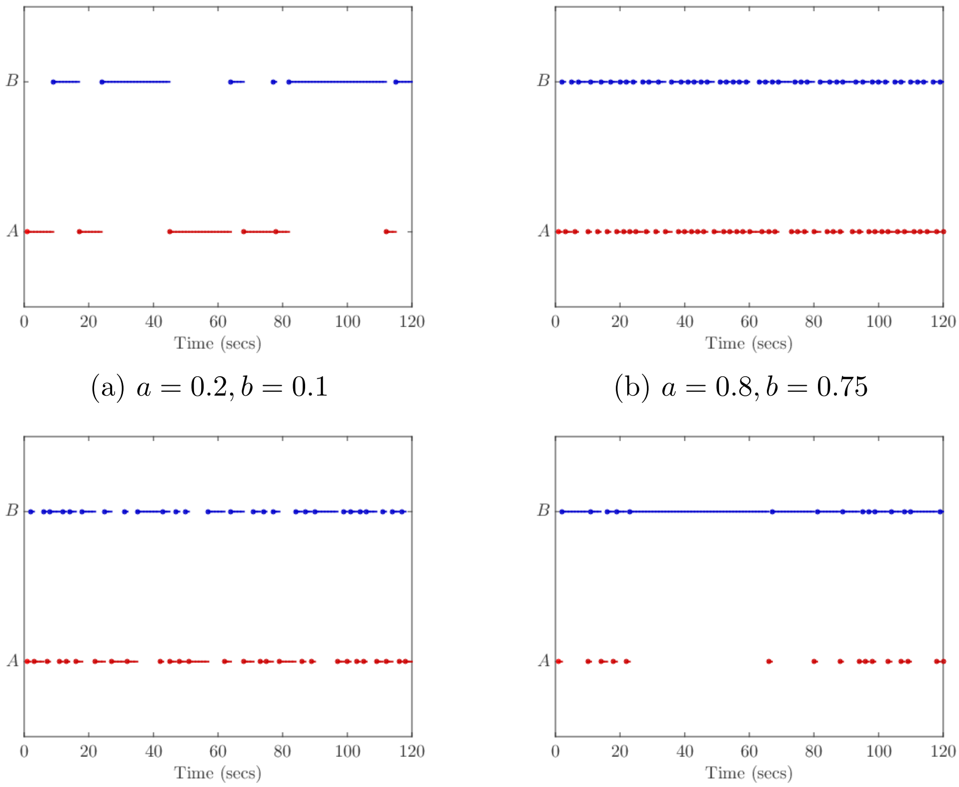
$$\pi = \pi P.$$

[2] for more details on Markov chains see e.g. [8]

(2.4.6) In our model for turn-taking these unconditional probabilities are

$$\pi_A = \frac{b}{a+b} \text{ and } \pi_B = \frac{a}{a+b}.$$

(2.4.7) Figure 7 represents different scenarios for conversations obtained by simulating Markov chain with different choices of probabilities a and b. The model resembles the conversations from the Listening Project (see figure 1), capturing a range of situations, e.g. in which both person A and person B manifest long speaking times (a), the conversation is very frequently switching from one person to the other (b), moderate levels of willingness from both sides to let the other person speak (c) and a case where one person (B) tends to speak during more time with occasional switches to the other person (A) whose turns are short (d).



(a) $a = 0.2, b = 0.1$

(b) $a = 0.8, b = 0.75$

(c) $a = 0.5, b = 0.4$

(d) $a = 0.5, b = 0.4$

Figure 7: Speaking times in conversations obtained by simulating Markov chains for four different choices of transition probabilities a and b. Model was able to capture different scenarios appearing in real conversations.

(2.4.8) The next question is therefore how to establish the rates/probabilities inherent in the Markov chains. The first approach we considered is by estimation of the probabilities directly from data. The second approach deals with models for the probabilities a and b that allow its evolution over time and is described in the following sections.

(2.4.9) The simplest approach is to assume that both probabilities are constant and thus represent the willingness of each partner to stop speaking during the whole course of the conversation. Given a sequence of durations of turns in conversations $\tau_i, i = 1, ..., N$, we first construct a Markov chain of length T (total time of the conversation) as

$$X = (\underbrace{A, \dots, A}_{\tau_1 \text{ times}}, \underbrace{B, \dots, B}_{\tau_2 \text{ times}}, \underbrace{A, \dots, A}_{\tau_3 \text{ times}}, \dots)$$

and calculate the number of transitions from A to B, number of transitions from B to A, number of times the chain is in state A (number of seconds person A is speaking) and number of times the chain is in state B, noted as n_{AB}, n_{BA}, n_A and n_B, respectively.

(2.4.10) Then we estimate the transition probabilities by the following expressions (see [9])

$$\hat{a} = \frac{n_{AB}}{n_A} \text{ and } \hat{b} = \frac{n_{BA}}{n_B}.$$

Note that $n_A + n_B = T$ and $n_A = n_{AA} + n_{AB}$ (analogously, $n_B = n_{BB} + n_{BA}$).

(2.4.11) The drawback of this approach is that it cannot show us how mimicry might evolve over time. In order to see if there is any evidence that rates could change over time we divide the total time period $[0, T]$ in m parts and estimate probabilities $a(t_i)$, $b(t_i)$, $i = 1, \dots, m$ for each of those parts. If one opts for equal lengths, i.e. $t_1 = t_2 = ...$, care must be taken with the last interval, which might not coincide with the desirable interval length.

(2.4.12) Figures 8 and 9 show estimated constant and time dependent probabilities for two longer conversations. An interesting structure can be observed. The conversation corresponding to the figure 8 was an interview of Nigel Farage by Andrew Marr and while listening and manually labelling the speaking times the study group noticed a level of competitiveness and a lack of rapport. Estimated probabilities do not seem to converge, moreover, the difference between them is significant. On the other hand, the conversation from the figure 9 was between a husband and wife (from the BBC Listening Project) in which a notable level of rapport was present. The corresponding difference between the transition probabilities is greatly reduced.

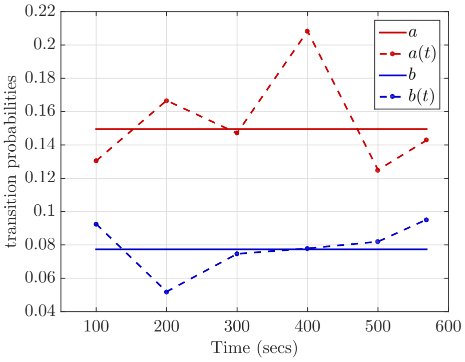


Figure 8: Estimated single (solid line) and time dependent (dashed lines) transition probabilities for a conversation in which no rapport was evident.

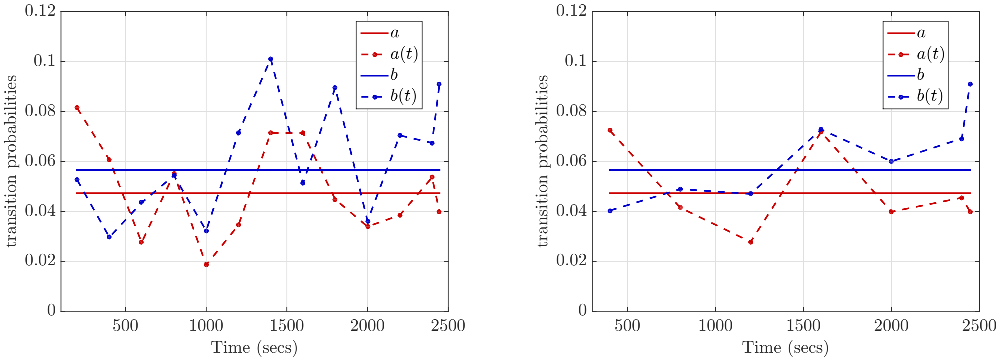


Figure 9: Estimated transition probabilities within 200 seconds (left) and 400 seconds window (right) (dashed lines) with probabilities estimated for the whole conversation (solid lines) from the Listening Project in which rapport was evident.

(2.4.13) The calculations performed above are not computationally demanding and can be easily carried out on a smartphone device. One might come up with a measure of the distance between the two series of estimated probabilities as a measure of rapport, for example root mean square distance between the two transition probabilities

$$\mathbf{d} = \Big(\sum_{i=1}^{m}(a(t_i) - b(t_i))^2\Big)^{\frac{1}{2}}.$$

Such a measure could be combined with other important features of conversations and tracked in time for all interlocutors.

(2.4.14) We next consider modelling the probabilities such that they are governed by a system of coupled ordinary differential equations. Because the Markov chain itself is a random object, this deterministic model for the probabilities still fits with our observations of randomness in turn-taking.

2.5 Comfortability Model

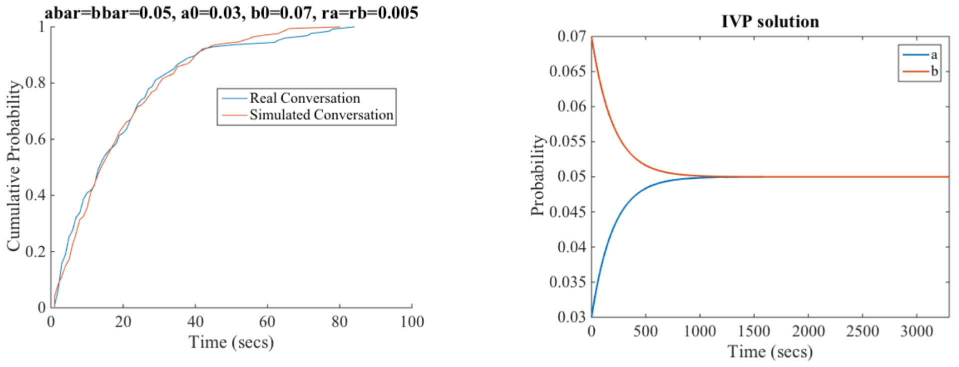
(2.5.1) There was much discussion over the sorts of ODEs we might consider to govern the evolution of the probabilities over time. A simple model, which was dubbed the 'Comfortability Model', represents how rapport might be revealed by the convergence of the probabilities to a 'comfortable' state. The ODEs we used were,

$$\begin{aligned} \dot{a} &= r_a \left(\bar{a} - a\right), \\ \dot{b} &= r_b \left(\bar{b} - b\right), \end{aligned}$$

where $\bar{a}$, $\bar{b}$ are the comfortable states, and r_a, r_b are the rates of convergence (or divergence if negative) to the comfortable states. We also have the initial conditions which represent the levels of willingness to let the other person speak at the beginning of the conversation.

$$\begin{aligned} a(0) &= a_0, \\ b(0) &= b_0. \end{aligned}$$

(2.5.2) We took this model and compared it to some long conversations which were taken from the BBC Listening Project. The aim was to plot the cumulative probability distribution of the simulated conversation and try to fit the six parameters of the model, $\bar{a}$, $\bar{b}$, a_0, b_0, r_a, r_b to the plot of the real converation data. Assuming the model is valid, this would reveal something about the nature of the conversation. It should be noted that we never attempted to fit the data computationally but if one were to try, it could be done using non-linear optimisation software. Of course the data from a Markov process simulation will be different each time, so we would have to take the cumulative data from many simulated conversations.



(a) (b)

Figure 10: Comparison of comfortability model and the BBC Listening Project data. In (a) is fitted simulated data from a 55 minute conversation and (b) is the evolution of the probabilities over time with the same fitted parameters in (a).

2.6 Talking-Listening Model

(2.6.1) We sought to identify whether allowing the coefficients in the differential equations governing the transition rates a and b to depend on the state of the system (A, B) was capable of more complex dynamical behaviours. In particular, we were interested in whether it could show periods where one speaker was the dominant party in the conversation, and whether different behaviours could be observed when there was a mismatch between a given speakers desire to converse and their actual engagement in the conversation.

(2.6.2) Our idea is that there are two different values for the speaker switching rates that a speaker wishes to adopt, depending on when they are currently the speaker or the listener. The speaker tends towards making longer utterances (a approaches a_{talk}, similarly for b), whilst the person currently listening tends towards making shorter utterances (a approaches $a_{\text{listen}} > a_{\text{talk}}$). This corresponds to

$$A : \begin{cases} \dot{a} = r_{a_A}\left(a_{\text{talk}} - a\right) \\ \dot{b} = r_{b_A}\left(b_{\text{listen}} - b\right) \end{cases}$$

$$B : \begin{cases} \dot{a} = r_{a_B}\left(a_{\text{listen}} - a\right) \\ \dot{b} = r_{b_B}\left(b_{\text{talk}} - b\right) \end{cases}$$

(2.6.3) The solutions to this system tend to have large excursions in which one speaker dominates the conversation. This is illustrated in figure 11.

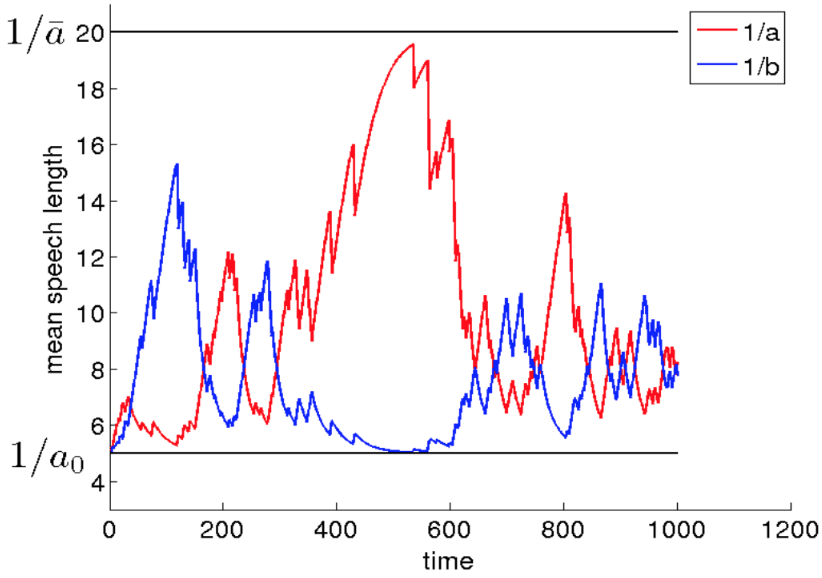


Figure 11: Evolution of the probabilities a and b using the Talking-Listening model. Here we have taken $a_{\text{talk}} = b_{\text{talk}} = \bar{a} = 0.05$, $a_{\text{listen}} = b_{\text{listen}} = a_0 = b_0 = 0.2$ and $r_{aA} = r_{aB} = r_{bA} = r_{bB} = 0.3$.

2.7 Conclusions

(2.7.1) From the manually labelled conversations from the BBC Listening Project, we do not find a correlation between lengths of pairwise speech fragments as we initially expected. Instead it seems that there is an element of memorylessness in turn taking. Therefore, modelling conversations using Markov chains seems a reasonable strategy, and indeed simulated conversations using this approach show very similar structure to real conversations.

(2.7.2) The probabilities of each speaker stopping talking were estimated from the data. We began by considering these as constant through time, although there was some evidence that these changed throughout longer conversations. We hypothesised that these probabilities would evolve as rapport developed between participants in the conversation. Therefore, we developed a number of ODE models for these probabilities.

(2.7.3) To continue this line of research, a larger quantity of labelled data would be very valuable, particularly from longer conversations where rapport occurs. This would allow us to fit our ODE models to data and test their predictive power. Additionally, it would be interesting to compare these results with conversations where rapport does not develop, to determine whether different patterns of turn taking are present.

3 Speaker Identification

(3.0.1) Speaker identification is the process of determining which part of a speech stream is uttered by which speaker. The problem of identifying speakers in a two person conversation when there are two microphones (i.e. a stereo recording) is called the “cocktail party problem”. This problem can be solved using principle component analysis which can be implemented trivially using a software package such as MATLAB. Due to technology constraints, the problem of when there is only one microphone (i.e. a mono recording) is considered which motivates approaches based on acoustic characterisations of speech.

(3.0.2) There exists a variety of open source (e.g. CMU Sphinx, MSR Identity Toolbox, ALIZE) speech recognition software; however, these are computationally expensive and require a large set of training data. Instead, we attempted to extract speech characteristics by analysing the spectrum and energy of the audio signal. Two approaches were used, both of them providing a reasonable estimate of the time when a transition between two different speakers occurs. The results were not very accurate and have been tested on too few conversation samples. However we believe this is a promising route and with more time could be further developed into a more robust algorithm.

3.1 A Low Frequency Classification Approach

(3.1.1) The natural difference in the pitch of male and female voices motivates using pitch as a characteristic to identify different speakers in a male-female conversation. This idea has been tried before (see MFCC in section 3.2) but can be computationally expensive or requires a large dataset. The idea here is to see if we can develop a robust method which is computationally inexpensive. The following method is just a starting point since it uses only one characteristic of speech. However, if the work was to be continued then a larger number of characteristics could be used, such as volume, speed and frequency range, at which point the problem becomes a classification problem by (i) detecting when a sufficient change in characteristics is observed and (ii) identification of the two speakers by their individual characteristics.

(3.1.2) Detecting when a speaker changes is not well defined due to a multitude of reasons including when a speaker interrupts the other, interluding silences and tailing off. Further, it was found that when listening to conversations from the BBC Listening Project that there would be a large number of pauses in a single persons speech occurring naturally. Therefore, using pauses in conversation as a means of detecting a change in speaker would not be effective.

(3.1.3) As a starting point we consider a part of a conversation where we are given the information that person 1 talks for a while, and then person 2 takes over - the important simplification being that we know that there is only change in speaker. The aim is to then identify the time at which the person 1 stops talking and person 2 starts talking[3]. The characteristic we consider here is the amount of energy in the low frequency part of the spectrum, although the general idea described could be applied to any characteristic.

3.1.1 Outline of the Method

(3.1.4) The method developed is outlined below. We assume that we have an audio signal and we take a discrete sample $x(t)$ for $t = 1, 2, ..., T$. The first part of the method is to split the audio signal into a number of sections and for each section we analyse the energy of the low frequency modes.

1. Split the conversation into N audio samples.
   $\tau = \text{floor}(T/N)$
   $X_j = x(1 + (j-1)\tau : j\tau)$
2. Take Fast Fourier Transform of each audio sample X_j.
   $\hat{X}_j = \text{fft}(X_j)$.

[3] The problem when we do not know the number of switches does contain an extra layer of complexity which would have to be dealt with statistically - perhaps with a weighting towards a sensible number of switchovers in a given time interval.

3. For each audio sample $\hat{X}_j$ we split the frequency space into K groups and measure the energy of the lowest frequency group.
   $\kappa = \tau/(2K)$
   $E_j = \sum_{m=1}^{\kappa} |\hat{X}_j(m)|^2$

The next part of our method is to analyse if there are any changes in the energy of the lowest frequency modes. We postulate that the change we are interested in is the change in the amount of energy rather than the variation of the energy. Therefore we calculate the envelope of the local maxima of the energy curve and then take a moving average. Taking a moving average smooths out the natural variation so we can assign characteristics to a particular interval of time.

4. Take the average of local maxima.
   $E_j \rightarrow F_j = \{E_j : E_j > E_{j-1}, E_{j+1}\}$
   $\bar{F} = \sum F_j/\text{size(F)}$

5. Remove small elements which are less than 10% of this average (i.e. points where a lot of energy is located in other modes or local maxima caused by background noise).
   $F_j \rightarrow G_j = \{F_j : F_j > \bar{F}/10\}$

6. Take a $2P$ moving average for some value of P, say $P = 5$.
   $v = \text{ones}(1, 2P)/2P \rightarrow H = G \star v \; (\text{conv}(G, v)) \rightarrow H(l) = \sum_{j=l-(P-1)}^{l+P} G(j)$.

Now the problem is to detect the change in speaker from the moving average. Given that there is a change at a certain point p_c in $H(p)$, and we want to find the point p_c. We assume that the change is in the mean value of $H(p)$, assume it occurs at the point $p = q$ and postulate that for $p < q$, $H \sim N(\mu_1, \sigma_1^2)$, and for $p > q$, $H \sim N(\mu_2, \sigma_2^2)$. From this we can calculate maximum likelihood estimates of $\mu_{1,2}$ and $\sigma_{1,2}$. We can do this for every point q, and then we say that we have the best fit when the variance is at a minimum, or the sum of the variances is at a minimum.

7. Find where the variance is at a minimum.
   Given q, $\mu_1 = \sum_{p<q} H(p)/q$, $\mu_2 = \sum_{p>q} H(p)/(\text{size}(H) - q)$.
   Identify change point p_c, which is the value of q which minimises $(\sigma_1^2 + \sigma_2^2)$.

8. Identify the index p_c from the filtered data, and associate this index with the actual change over time, i.e. its corresponding index in the original audio data.

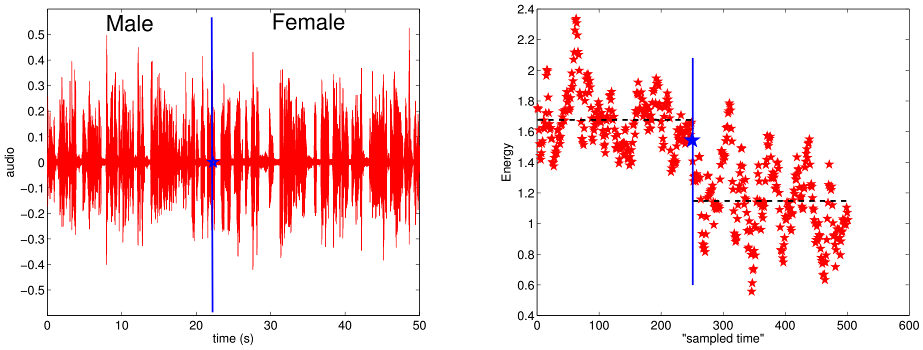


Figure 12: Male-Female conversation results. Left: Raw audio data with changeover point marked in blue. Observed time: 22-23s. Right: Processed data (moving point average of low frequency energy) over "sampled time". Calculated time 22.3s.

3.1.2 Preliminary Results

(3.1.5) The speaker characteristics will vary from person to person, with large differences between some people, and fairly small differences between others. So as a first test, we would like to check this algorithm works when we would perceive there to be a larger than average difference between the speakers - i.e. in a male-female conversation. To do this we took a 50 second segment of a conversation from the BBC listening project, where we have checked there is one change over point which was observed to be around 22-23s. Figure (12) shows the results of the algorithm on this conversation.

(3.1.6) On the left we have the raw audio data plotted over time, and in blue we have marked on the change over point. It is not obvious from this raw audio data where the changeover point is. On the right, we have plotted out the processed data using the algorithm described above. It is clear from this graph that this algorithm identifies a characteristic of the speakers which we can use to distinguish the speakers in this case. Listening to the conversation we observe a change in speaker around the 22-23s mark, and this analysis predicts a changeover point of 22.3s in agreement with the observed changeover time. The testing of this algorithm is by now means comprehensive but repeating this analysis on several parts of the same conversation produced similar results.

(3.1.7) If we pick a conversation between two males for example, where their voices are not so distinct then we observe that the change in this characteristic is not so obvious and although it appears to be observable in the data, it was difficult to pick up mathematically (using maximum likelihood). However, it appears that the variation of the characteristic described above, between the two speakers may provide be enough to distinguish the speakers. There was not enough time to test this idea.

3.2 Mel-Frequency Cepstral Coefficients (MFCC)

(3.2.1) A more sophisticated approach to voice recognition can be achieved through using the mel-frequency cepstral coefficients (MFCC) as a characteristic of a particular speaker; this appears to be the standard pre-processing step within existing tools for speaker recognition [11].

(3.2.2) These coefficients are calculated though a somewhat complex sequence of transformations. A windowed Fourier transformation is taken of the input signal, and the log power spectrum calculated from the logarithm of the square of the absolute values of the Fourier transform. Motivated by the human auditory system's response, a bank of (often 22) overlapping triangular filters, with central frequencies spaced on the mel scale, are applied to the log-power spectrum. The output of this filter is a low-dimensional vector; partly to approximately decorrelate the components of this vector, and partly as an analogue to the cepstrum of a signal, a discrete cosine transform is applied to these vectors. The lowest-order coefficient simply corresponds to the total power of the signal, and so is discarded, and usually the next 12 modes are retained.

(3.2.3) The MFCC provide a representation of the sound within each of the sampling windows. For some pairs of speakers, the average of the MFCC vectors over a short phrase seems to be distinct, but in many cases the average MFCC levels differ little between different speakers.

3.2.1 Outline of the Method

(3.2.4) Gaussian mixture models (GMM) are statistical models for data, which represent data points as being generated by a collection of multivariate gaussian distributions.

(3.2.5) Such models are commonly used for the unsupervised clustering of data, in which each point is assigned to the Gaussian distribution that is most likely to have generated it. In the context of speaker recognition, however, a GMM is trained on the set of MFCC coefficients generated by a single speaker. The first (and sometimes second) order temporal derivatives of the MFCC coefficients are often included in the state vector for each timewindow. Such models are a common approach to speaker identification [13, 12]

(3.2.6) Many of the open source programs for performing speaker diarization are able to operate in an unsupervised mode, where they are not explicitly trained on the voice of each of the speakers, or even the number of speakers. One approach is to split the speech into a number of different sections, and then merge and split the individual GMM models describing each section until the segmentation of the audio stream satisfies certain statistical criteria. We applied some of the open-source software available, but had mixed results; natural conversations are perhaps more challenging than meetings and television broadcasts, which are the applications that current software was optimized for.

(3.2.7) To understand if the MFCC coefficients supplied any useful information for our purposes, we applied a very simple supervised method (somewhat similar to that of [13]) to two of the conversations from the listening project. Two small sections of audio (a few seconds in length), each containing the voice of a single speaker and as little silence of possible, were selected from near the start of each conversation. We applied the MFCC transformation to each of these and used them, along with their first-order temporal derivatives, to train a pair of 10-component GMMs, one for each speaker, using functions included in the open-source VOICEBOX package for MATLAB [10].

(3.2.8) We then calculated the MFCC coefficients for the remainder of the audio. For each sampling window, we calculated the (logarithms of the) probability densities, d_1 and d_2, for each of the trained GMMs. We calculated $d = d_1 - d_2$, which takes positive values when the sample is most likely to be generated by speaker 1, and negative values when the sample is most likely to be generated by speaker 2. We then took the moving average of d over 100 samples. Regions of positive and negative d were identified (from the zero-crossings of d), and sections of length greater than some threshold (here chosen to be 1 second) were identified with the appropriate speaker.

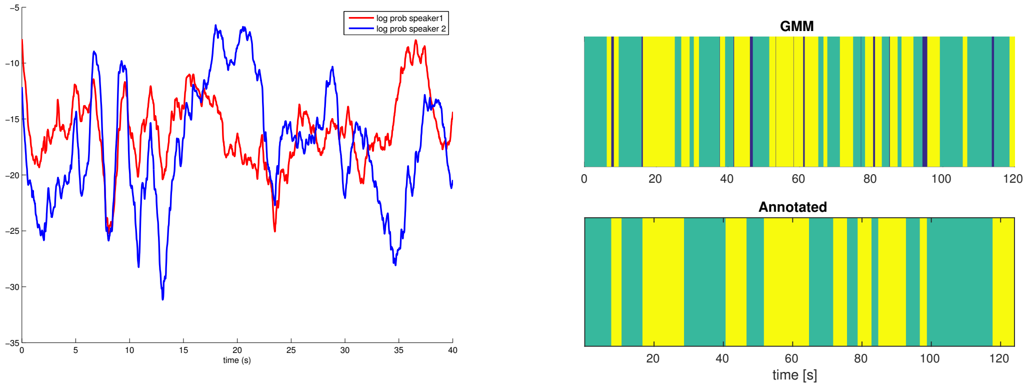


Figure 13: Speaker identification using MFCCs and GMM. Left-hand plot shows the (smoothed) probability densities for the pair of GMM trained on each speaker; this indicates which speaker is most likely to be talking at each time-point. Right-hand plots shows the periods of speech assigned to each speaker, and a comparison with the manual classification of this conversation.

3.2.2 Preliminary Results

(3.2.9) Results from the GMM/MFCC approach are shown in Figure 13. This gave results that were roughly in agreement with those manually annotated by a listener. However, we found that the training data needed to be selected carefully, as our first attempt for this particular data set gave very poor results.

3.3 Conclusions

(3.3.1) Let us now summarise the work we have completed on speaker recognition. We have developed two different methods of speaker identification that are computationally inexpensive and both methods suggest that speaker identification can be achieved using methods based on characteristics of speech rather than taking an approach that relies on either computationally expensive methods or the use of a large set of training data.

(3.3.2) Our first method used a moving point average of the low frequency energy to distinguish between two speakers. Our results suggest that it is possible to accurately distinguish between two speakers (with sufficiently different characteristics) using this method. However, if speech characteristics are not sufficiently distinguishable then our work suggests that another measure such as time variation in the low frequency energy may be a suitable metric.

(3.3.3) The second method confirms that the MFCC contain sufficient information to distinguish between two speakers. The method is reasonably computationally efficient, but requires training on carefully selected samples of the voices of the two speakers. This may be difficult to achieve for our current application.

4 Features of mimicry

(4.0.1) There are a number of non-verbal features of speech that one may expect to correlate to the level of rapport that exists in a conversation: pitch, volume, speech rate, pause durations between words, latencies of responses when the speaker changes, frequency of interruption, intonation and accent. These are all further referenced in [7]; we chose to focus on the first three. The main question we are addressing is whether there is a matching of these features between the speakers as they build rapport and thus start to mimic each other.

4.1 Pitch

(4.1.1) Tools were developed to extract the pitch of a speaker at a given time from an audio file, using Praat and some purpose-written Matlab code. Praat is freely available speech-analysis software, used as standard in academia [2]. The Matlab code was written to be more computationally efficient, and so viable for implementation on a smartphone once rewritten in a suitable language.

(4.1.2) A description of the Matlab code's algorithm, together with the full code, is in appendices, section A.1.

(4.1.3) The pitch-identification tools were checked using samples of music and a number of conversations. Provided the recording was of good quality, such as those from the BBC Listening project and from studio-recorded music rather than from laptop microphones, the tools proved effective and robust. Pitch varied considerably with different vowel sounds (range approx 100Hz), and was undefinable for unvoiced phonemes, but suitable time-averaging (measurements every 0.1 seconds over approx 5 seconds) gave good results.

(4.1.4) We investigated trends in pitch over entire conversations from the BBC Listening Project, averaging pitch over each period when a speaker was talking. We expected in this natural conversation to observe mimicry, with the pitches of two different speakers moving closer together over time, and tracking each other throughout a conversation. However, no significant trends or correlations were observed, implying no long-time-scale pitch mimicry in these natural conversations.

(4.1.5) We then tested whether there is mimicry at the pitch level when the speaker changes. More precisely, if there is mimicry in the conversation we might expect one speaker to match their pitch level when they take over from their conversation partner. The pitch for this part of the analysis was extracted using Praat. The conversation being analysed is taken from the BBC Listening Project episode aired on 21 September 2014, and is between a mother Sarah and her daughter Natalie discussing Natalie's progressive blindness.

(4.1.6) Figure 14 shows a scatter plot of the average pitch of the last six seconds of one speech fragment and the first six seconds of the next speech fragment, when conversation switches to the other person. The data is separated into speaker 1 taking over from speaker 2 and speaker 2 taking over from speaker 1; that way, if one person tends to mimic the other more, that will be made clear by the data.

(4.1.7) There seems to be some pitch matching when conversation shifts from speaker 1 to speaker 2 (correlation of 0.48), but the data is very noisy; there appears to be no pattern when speaker 2 is followed by speaker 1.

(4.1.8) When the averaging is done over intervals smaller than six seconds, the correlations are even weaker.

(4.1.9) The averaging was done excluding any pauses, to ensure that the average pitch is taken when the person is actually talking. However, this will not exclude sounds such as laughing or other non-verbal interjections, which may have very different pitches from the person's normal speaking voice. If a speech fragment is less than six seconds in length, the averaging is done over the whole time interval.

(4.1.10) Because different people will have different baseline pitches (in particular male/female voices), the data shown in Figure 14 is normalised by the

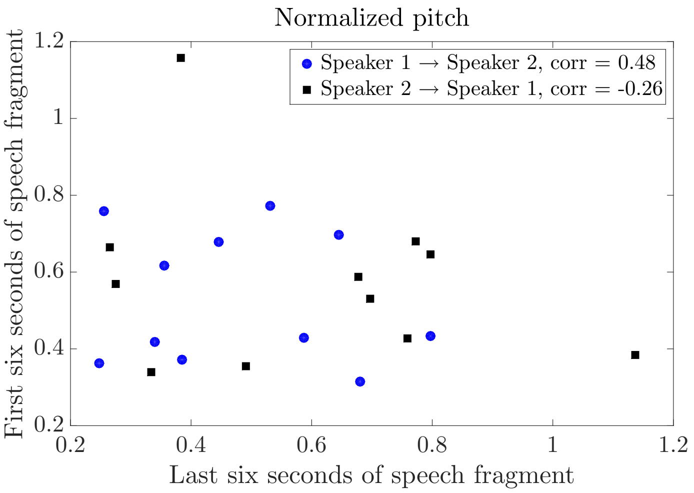


Figure 14: Correlation between average pitch in the last six seconds of a speech fragment and the first six seconds of the next speech fragment

average pitch of that person throughout the entire conversation.

4.2 Volume

(4.2.1) We similarly tested whether there is mimicry in the volume at which people talk to one another. We attempted to analyse the data in a variety of ways. Both overall changes in volume over the whole conversation as well as the range of volume at which people speak over a speech fragment were considered.

(4.2.2) First there are some issues we need to address with volume analysis. Absolute volume is difficult to measure without specialist equipment and environments. In a phone conversation issues with distance from microphones is unlikely to be important as most people hold their phones at fixed distance. However, there are other issues likely to be found with volume analysis over the phone – in particular noise but also phone connection issues. The data set studied here is all from the listening project where we do not know the position of the microphone (unsure whether equidistant from the speakers) and whether the speakers' position changes – but we expect fidgeting will occur.

(4.2.3) We first study the overall volume to see if there is any convergence to a particular volume over time. Figure 15 shows that there is little correlation with time. The volume here is extracted via calculating the root

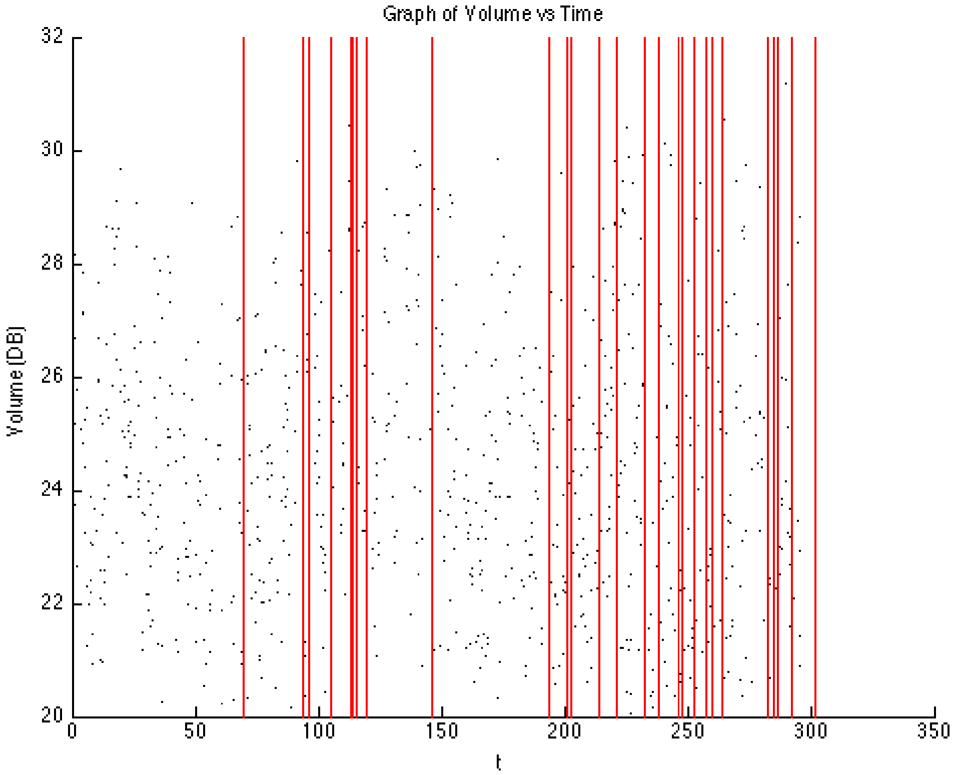


Figure 15: In this figure the black dots represent the RMS volume per syllable, this is plotted against time to see if there are any global volume trends over the conversation. The red lines show the point at which the speaker changes. This shows no overall trends nor is there any clear difference between the two speakers.

mean squared (RMS) amplitude - we exclude all pauses in the conversation by cutting out volumes below ≈ 20DB. The RMS is calculated over the average syllable length [4].

(4.2.4) Thanks to the manual extraction of speaker times we were able to study the volume over a speech fragment to see if there were any correlations or patterns in the volume, between speakers. Figure 16 shows how the RMS amplitude and the volume difference between the two speakers changes over the conversation. The RMS amplitude seems to increase over the conversation for both parties, however the difference between the RMS volume of the two speakers does not converge to zero. We observe a range of different RMS difference showing that there is not just one very loud or very quiet speaker. The RMS volume was calculated over an entire

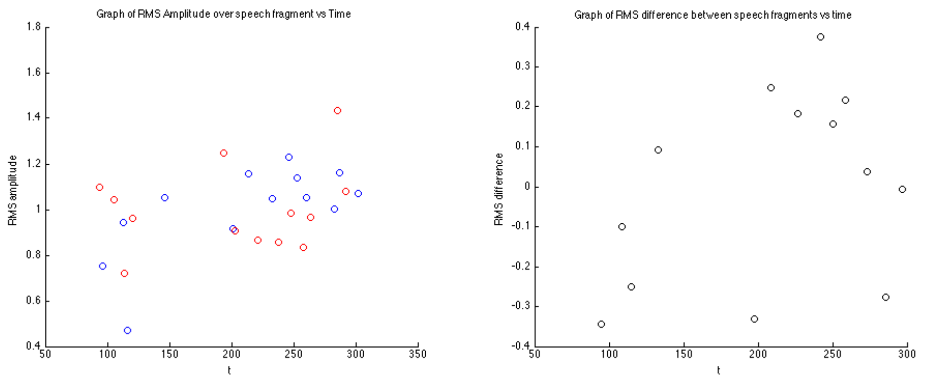


Figure 16: Left: RMS amplitude over each speech fragment - where red represent the first speaker and blue represents the second speaker (with each speaker non-dimensionalised by their average RMS amplitude over the entire conversation). Right: the RMS amplitude difference between the two speakers - there is no convergence to zero.

speech fragment, which can be quite long, hence we look at the maximum volume of a speaker over their speech fragment as we expect this will be an easier feature for the ear to pick out.

(4.2.5) Figure 17 shows both how the maximum amplitude and the maximum volume difference between the two speakers changes over the conversation. Again we see volume increase over the conversation, but no correlation between the two speakers.

(4.2.6) We then attempted to study the correlation between the loudest and the quietest syllable in a person's speech fragment to see if there was any correlation between the spread of volume between the first and second speaker. We extract this information from the waveform by first filtering the data to remove high frequencies, using a low pass filter, then the peaks in amplitude are found and the spread of volume is found by comparing the lowest and highest peaks. Figure 18 shows the difference of the spread of volume between the two speakers over time, in this plot we see that the speakers move to speak at similar volume ranges. However, when we take a longer (and different) conversation we see no long range convergence over time. This is shown in Figure 19, where there appear to be short range correlations at the start of the conversation.

(4.2.7) To try and extract more volume features of the conversation we look at the separated audio of the the two conversations. Here we look for patterns in volume changes to see if this is mimicked by the listener.

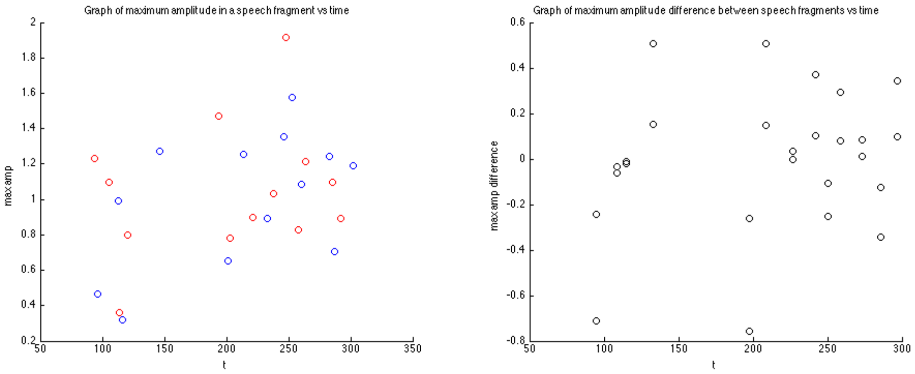


Figure 17: Left: maximum amplitude over each speech fragment - where red represent the first speaker and blue represents the second speaker (with each speaker non-dimensionalised by their average maximum amplitude over the entire conversation). Right: the maximum amplitude difference between the two speakers - there is no convergence to zero.

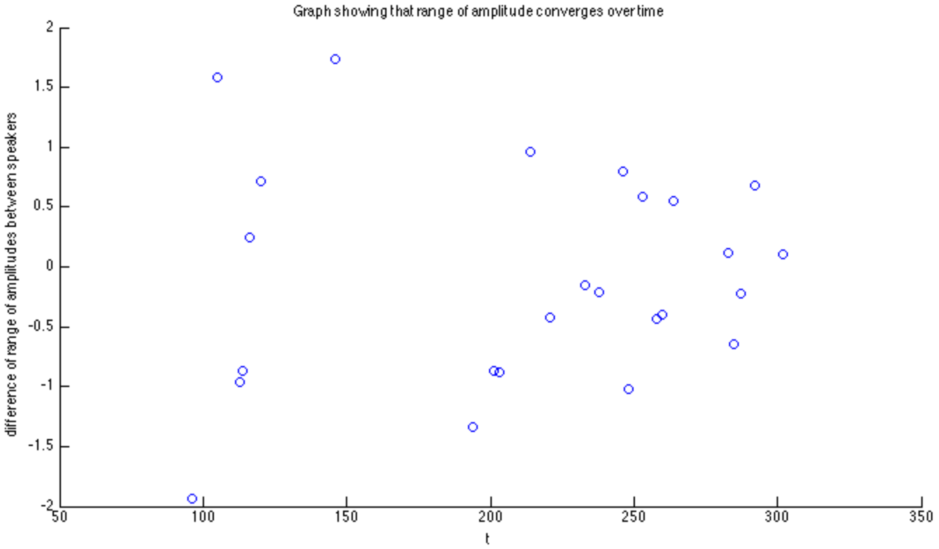


Figure 18: The difference between the maximum and minimum volume peak is calculated for each speech fragment, the average over each person is taken. Then the difference between the two speakers is plotted against time. The graph shows convergence with time.

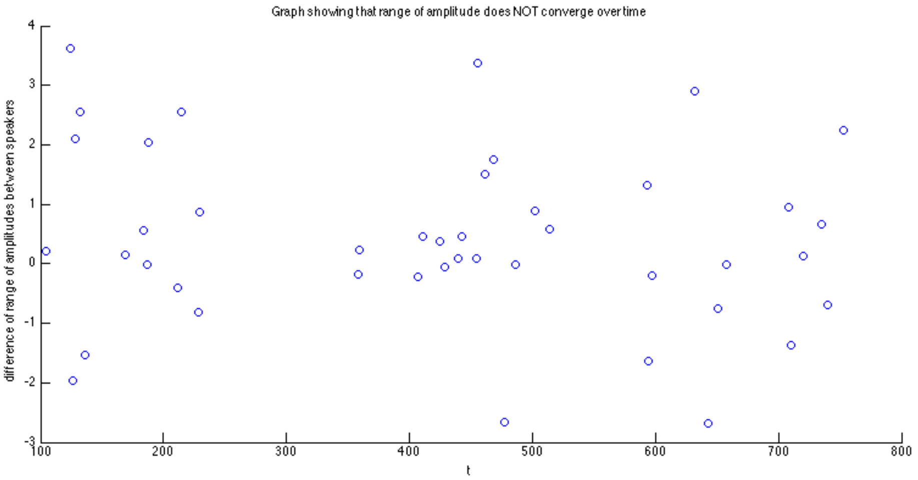


Figure 19: The same is done as in Figure 18 but for a longer conversation. There is correlation over parts of the conversation but no overall correlation.

The audio files for each speaker are analysed separately and the change in volume over one second is calculated. Again pauses are removed and we expect to capture the change in volume over about 6 syllables per data point. As shown in Figure 20 there is little to be said about the mimicry of volume changes over one speech fragment, however we are able to observe differences between the two speakers of the course of the conversation, the first speaker appears to change their volume much more over one second sections than the second speaker.

(4.2.8) Finally, we repeated the analysis done for pitch to see if there is any mimicry when the speaker changes. We used Praat to extract volume information and ran the analysis on a conversation from the 21 September 2014 episode of the BBC Listening Project (conversation between Sarah and Natalie).

(4.2.9) Figure 21 shows a scatter plot of the average volume of the last six seconds of one speech fragment and the first six seconds of the next speech fragment. There seems to be some degree of mimicry when speaker 1 follows speaker 2 (correlation 0.71), although it is difficult to generalise

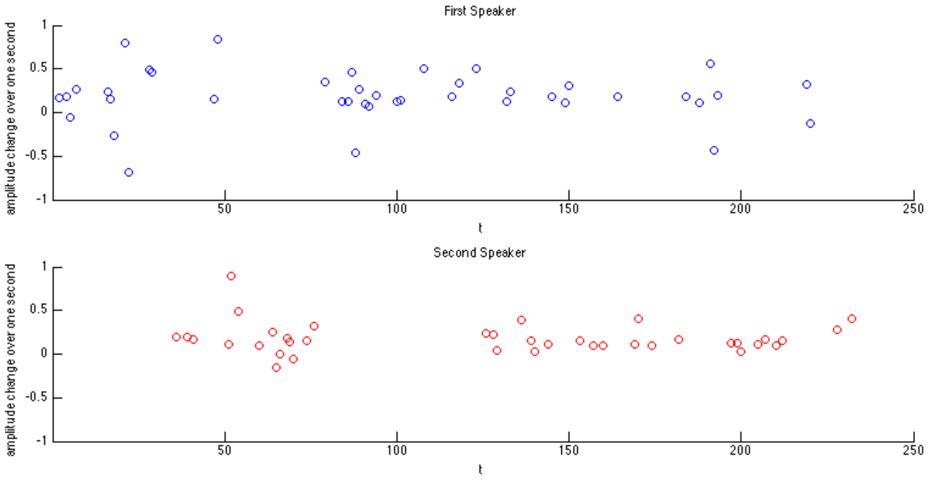


Figure 20: The volume changes over one second for each speaker are plotted against time to see if there is any mimicry in the pattern at which different people speak - there do not appear to be any obvious trends through similar sections.

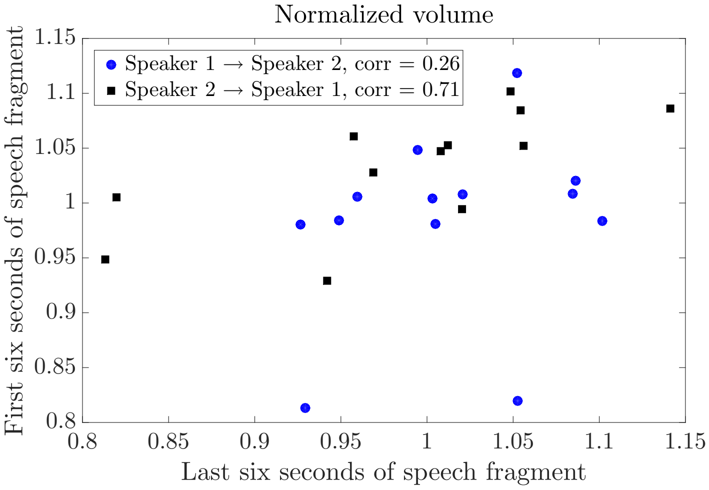


Figure 21: Correlation between average volume in the last six seconds of a speech fragment and the first six seconds of the next speech fragment

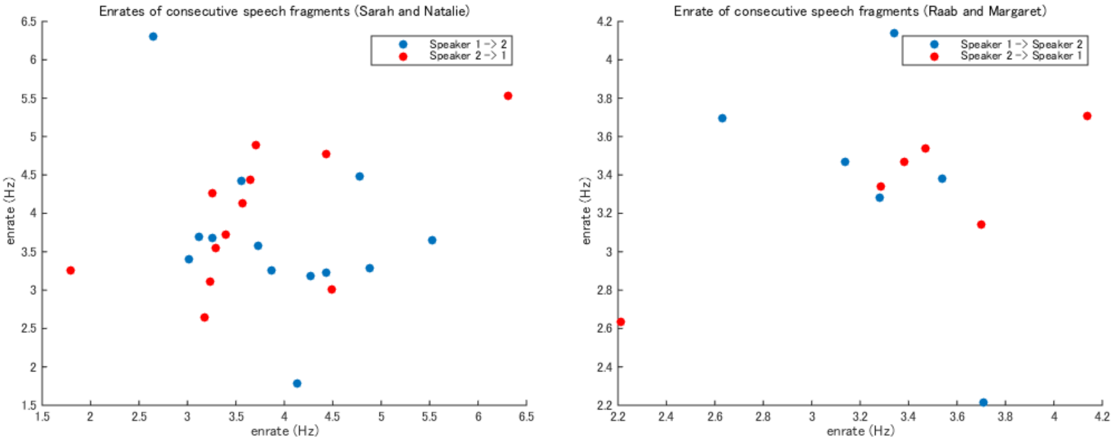


Figure 22: Correlation of speech rates of consecutive speech fragments (conversations taken from the BBC Listening project; 2 different conversations between different pairs of people are shown here)

from one conversation.

4.3 Speech rate

(4.3.1) Finally for our mimicry analysis we focused on the speech rates of the two participants, hoping to see them converge over time between individuals with rapport. We referred to [5] for possible ways to estimate speech rate without using manual methods and decided to implement the enrate method [6]. This is due to several reasons: though [5] showed that enrate was the least accurate out of the 8 methods they tested, it was the easiest and still fairly reliable.

(4.3.2) There are some issues with this method however, namely that it uses a low-pass filter (we used a Butterworth filter) and discrete Fourier transform. Whether this is realistic enough to do on a smartphone app may be a concern.

(4.3.3) The below figures are both of speech rates of conversations between people who have a strong rapport between (the left is a conversation between a mother and daughter, the right is between a long-married couple). From them, we cannot draw a conclusion regarding the relationship between speech rates and how much of a rapport people in a conversation feel. There is no strong correlation between speech rates for any of the speakers.

4.4 Further work

4.4.1 Pitch

(4.4.1) More work is required to find which trends and correlations in pitch are associated with rapport, since none of the trends or correlations investigated proved convincing.

(4.4.2) To do this, a larger number and variety of conversations are required, including a number of conversations with poor rapport for comparison.

4.4.2 Volume

(4.4.3) To understand how the volume changes over a conversation a better understanding of how humans hear different sounds is necessary. If the volume is perceived to be less by the human ear even though our data shows they are the same we need to be able to correct for this.

(4.4.4) Furthermore, we need to reduce the errors in our results from the data set – we need to be able to measure volume where the speakers are equal distance and at fixed positions away from the microphone. Additionally, we would like to be able to take into account that during phone conversations the speakers may be in different environments with different noise levels. There is an application which based on the noise level calculates the appropriate volume which the user should speak at [3] – this may be possible to use over the phone to take into account the different noise levels in each person's background environment.

(4.4.5) Also, experiments may be useful for tracking the volume level of a conversation by getting one speaker to increase/decrease their volume slowly to see if the other speaker follows suit. Or by slowly increasing the background noise.

4.4.3 Mimicry metric

(4.4.6) Our original plan was to find a list of features that correlate with the degree of vocal mimicry in a conversation, and then combine them in a single metric, which would make it easy to compare the level of mimicry in different conversations. Because finding features that predict mimicry proved difficult in its own right, we spent very little time on designing this metric.

(4.4.7) Some initial ideas to go from the features discussed earlier (pitch, volume, speech rate) to measures of mimicry are discussed in [7]. It is suggested that as mimicry develops during a conversation, two things may happen: (1) the autocorrelation of speech patterns for the speakers considered individually decreases, as they move away from their own speech styles to match the style of the other person; (2) the correlation of speech patterns between the two speakers increases over time as the two participants start to mimic each other. As a caveat, the conversations in [7] have a special format where one of the participants takes an active role of a presenter at the beginning of the discussion; there is a question of whether the patterns uncovered here appear in other types of conversations.

A Appendices

A.1 Matlab pitch calculation

(A.1.1) In the Matlab code, pitch at time t_0 was found by calculating the time delay $\tau_0(t_0)$ that maximises autocorrelation $\langle s(t), s(t+\tau)\rangle_{t_0<t<t_0+\Delta t}$ of a .wav audio signal $s(t)$. If the autocorrelation is calculated over a time interval less than one syllable (achieved in practice by pre-filtering the signal with a 0.05 second long Hanning window), this autocorrelation will be maximised by the period of the fundamental frequency. The pitch at that time $f_0 = 1/\tau_0$. Autocorrelation as a function of $f = 1/\tau$ for a 180Hz sine wave is plotted in figure 23.

(A.1.2) To reduce noise, all values of $f_0 > 265$Hz are rejected, since these are above the frequency of human speech, and correspond to fricatives, pauses between speech, or background noise. Full code is below:

```
function pitch=pitch_detection(t0,input)
%PITCH Calculates pitch of input waveform at times t0
%From series of sound pressure levels input, taken with frequency f_input,
%calculates pitch over time interval T_window after time t0

% Frequency of input signal
f_input=44100;
```

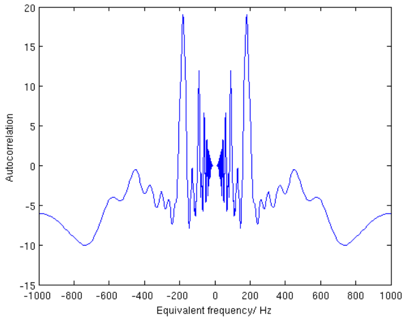


Figure 23: Autocorrelation of a Hanning-filtered 180Hz sine wave, as a function of $f = 1/\tau$

```
% Time over which desired pitch is measured
T_window=0.05;

% Vector lengths of data
n0=floor(t0*f_input);
n_window=floor(f_input*T_window);

% Hanning window function
window=-0.5*(1-cos(2*pi*(0:1/(n_window-1):1)));

pitch=NaN(length(n0),1);
for i=1:length(n0)
    % Autocorrelation calculation and plotting
    [corr,lag]=xcorr(window.*input((n0(i)+1):(n0(i)+n_window)));

    %figure;
    %plot(f_input./lag,corr)
    %set(gca,'XLim',[-1000,1000])
    %xlabel('Frequency (Hz)');
    %ylabel('Autocorrelation');

    % Finds autocorrelation peaks in terms of index lag, frequency
    [~,peaks]=findpeaks(corr);
    f_peaks=f_input./lag(peaks);

    % Selects largest peak, excluding that with zero lag
    if(length(f_peaks)>1)
        idnoninf=~isinf(f_peaks);
        [~,peak0]=max(corr(peaks(idnoninf)));
        f_peaks=f_peaks(idnoninf);
        pitch(i)=abs(f_peaks(peak0));
    end
end
idhigh=isinf(1./(sign(pitch-265)-1));
pitch(idhigh)=NaN;
end
```

A.2 Praat software for feature extraction

(A.2.1) Extracting pitch and intensity (i.e. volume) information using Praat can be done as follows: (1) split the audio into two separate .wav files for the two speakers[4]; (2) read the separate files into Praat and generate Pitch and Intensity objects by going to Analyse periodicity → To Pitch . . . and To Intensity . . . ; (3) save these objects to text files from the Save menu; (4) read the information in the text files into Matlab in a way that is useful for analysis.

(A.2.2) One advantage of Praat over a more simple function for pitch extraction is that it generates pitch contours that are relatively smooth. In contrast, implementations which find the most likely pitch independently over each time interval are more “jumpy", with frequent octave switches. Because human speech has few of these sudden changes in pitch, Praat penalizes pitch values which are very different from those at previous time steps, resulting in a smoother profile.

(A.2.3) If analysing pitch and intensity at once, it is useful to match the time intervals at which the two are sampled. By default, Praat sets the time step to 75 Hz/minimum threshold in Hz for pitch and 80 Hz/minimum threshold in Hz for intensity. Thus, one easy way to match the time steps is to set the minimum threshold to 75 Hz when generating the Pitch object and 80 Hz when generating the Intensity object.

(A.2.4) The Matlab scripts written to aid in this feature extraction process are located in the Dropbox folder under Code/PraatFeaturesAndAnalysis. In particular, separate_audio.m can be used to separate input audio into two audio files and extract_features_Praat.m is useful for reading in

[4] This was done using the speaker times determined manually by someone listening to the audio; ideally, speaker identification can be done systematically using an algorithm.

data from .Intensity and .Pitch text files generated by Praat.

References

[1] The BBC Listening Project http://www.bbc.co.uk/radio4/features/the-listening-project

[2] Boersma, Paul & Weenink, David (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.08, retrieved 24 March 2015 from http://www.praat.org/

[3] *Voice-O-Meter* , Use Your Noodle, LLC

[4] *Slow Down! Why Some Languages Sound So Fast* , Time Magazine

[5] Dekens T., Demol M., Verhelst W, Verhoeve P. (2007): "‘A comparative study of speech rate estimation techniques"’, *Interspeech* **2007**.

[6] Morgan N., Fosler E., Mirghafori N. (1997): “Speech recognition using on-line estimation of speaking rate”, *Eurospeech* **97**.

[7] Sun X., Truong K., Pantic M., Nijholt A. (2011): “Towards visual and vocal mimicry recognition in human-human interactions", *IEEE*

[8] J. R. Norris, *Markov Chains*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press (1998)

[9] Bruce A. Craig and Peter P. Sendi, Estimation of the transition matrix of a discrete-time Markov chain, Health Econ. 11: 33-42 (2002)

[10] Brookes, Mike. VOICEBOX [Computer program] Retrieved 24 March 2015 from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[11] Davis, S., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech and Signal Processing, IEEE Transactions on, 28(4), 357-366.

[12] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1), 19-41.

[13] Reynolds, D. A., and Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. Speech and Audio Processing, IEEE Transactions on, 3(1), 72-83.