UTILISING HIDDEN MARKOV MODELLING FOR THE ASSESSMENT OF ACCOMMODATION IN CONVERSATIONAL SPEECH

Vijay Solanki[1], Alessandro Vinciarelli[2], Jane Stuart-Smith[1] & Rachel Smith[1]

[1]Glasgow University Laboratory of Phonetics, University of Glasgow, Glasgow (UK)

[2]School of Computing Science, University of Glasgow, Glasgow (UK)
V.Solanki.1@research.gla.ac.uk, Alessandro.Vinciarelli@glasgow.ac.uk

ABSTRACT

The work presented here suggests a method for assessing speech accommodation in a holistic acoustic manner by utilising Hidden Markov Models (HMMs). The rationale for implementation of this method is presented along with an explanation of how HMMs work. Here, a heavily simplified HMM is used (single state; mixture of gaussians) in order to assess the applicability of more sophisticated HMMs. Results are presented from a small-scale study of six pairs of female Scottish-English speakers, showing measurement of significant trends and changes in holistic acoustic features of speakers during conversational interaction. Our findings suggest that methods integrating HMMs with current holistic acoustic measures of speech may be a useful tool in accounting for acoustic change due to speaker interaction.

Keywords: speech accommodation, conversation, acoustic phonetics, MFCCs, HMMs

1. INTRODUCTION

Studies of the phonetic features that contribute to speech accommodation [12] have made robust findings demonstrating its dependency on both segmental and non-segmental acoustic features along with its persistence across both long-term and short-term interactions [18, 17, 9]. However, previous research generally doesn't attempt to pull these streams of evidence together to assess accommodation holistically. The concept of listeners making accommodative gestures based on holistic interpretations of the speech signal is not new and has found empirical support in recent studies [3].

Traditional holistic methods of analysing speech accommodation are often perceptual, which can be hard to quantify. Other methods include segmental approaches, relying on manual phonetic transcription, which can be a painstaking and time-consuming process. In addition to these methods, it is also possible to use a holistic acoustic method, whereby the properties of the entire speech signal are captured and compared. However, assessing speech accommodation with holistic acoustics is not trivial. This is due to the subtle and contextually dependent nature of speech being at odds with the global approach of holistic acoustic measures. This makes assessing the realisation and relative direction (convergence/divergence) of speech accommodation difficult to account for with holistic acoustics. A holistic measure of accommodation must capture subtle features of human interaction whilst consisting of multiple spectral properties. This paper presents a small-scale study from a larger PhD project, part of which is attempting to resolve this issue through the application of Hidden Markov Modelling.

1.1. Holistic Acoustic Measures of Speech Accommodation

The acoustic features reflecting speech accommodation have tended to be considered as static, rather than as interacting and interdependent elements of the same signal. Acoustic measures such as the formant frequencies of vowels [1], speech rate [8] and fundamental frequency [7] have all proven to be excellent in assessing relative change in these acoustic features but taken separately, they cannot represent the interplay between acoustic features. In this respect, holistic acoustic measures offer a potential alternative. A good example of a holistic acoustic measure can be found in the literature on accent recognition. [13]'s ACCDIST measure utilises Mel-Frequency Cepstral Coefficients (MFCCs) in tandem with the ACCDIST accent tables. When tested against other measures of the spectral envelope, measurements with MFCCs resulted in an accent recognition rate of up to 92.3%. This provides good justification for the use of MFCCs as a holistic acoustic measure of speech. Further justification can be found in the artificial speech recognition community where they have long been in use (see [15]).

Whilst there are other holistic measures of speech (see [4]), MFCCs provide a good tool for a first pass attempt at holistically accounting for the speech signal.

1.2. Hidden Markov Models

Hidden Markov Models (HMMs) are able to characterise the general form of a continuous signal. When implemented for linguistic purposes, a HMM can be used to estimate the probability of a given speech sound having been uttered by a particular speaker.

More specifically, HMMs are probability distributions defined over joint sequences of symbols (eg. phoneme categories) and observations (eg. MFCC coefficients). For speech signals, they can be characterised when considering a sequence $S = (s_1, s_2, \ldots, s_N)$ of states, where every s_i belongs to a predefined set of symbols $V = \{v_1, \ldots, v_D\}$, and a sequence $X = (\vec{x}_1, \ldots, \vec{x}_N)$ of observations, where $\vec{x}_i$ is a vector of physical measurements extracted from a speech signal at time t_i ($t_j > t_i$ if $j > i$). A HMM is the joint probability $p(X, S|\Lambda)$ of observing X and S to occur together, where Λ is the set of the parameters. The parameters' set Λ can be characterized by taking into account the actual expression of the probability:

$$p(X,S|\Lambda) = \pi_{s_1} b_{s_1}(\vec{x}_1) \cdot a_{s_1 s_2} b_{s_2}(\vec{x}_2) \ldots a_{s_{N-1} s_N} b_{s_N}(\vec{x}_N)$$

where π_{s_1} is the probability of the sequence S starting with state s_1 (there are D parameters π_{v_i}, one per element of V), the $a_{s_i s_j}$ are the probabilities of a transition between state s_i and state s_j (there are $D \times D$ parameters arranged in a matrix A where element i, j corresponds to the probability of a transition between v_i and v_j), and $b_{s_i}(\vec{x})$ is the emission probability density function, i.e. the probability of observing $\vec{x}$ when the state is s_i (there are D distributions, one for each element of V, and each of them has parameters that are included in Λ).

In general, HMMs are used as follows: first a vector of physical measurements is extracted at regular time steps from a speech signal, resulting in a sequence X. Then, the sequence of states S^*, most likely to underlie the sequence of observations is found by:

$$S^* = \arg \max_{S \in \mathscr{S}_N^{(V)}} p(X,S|\Lambda)$$

Where $\mathscr{S}_N^{(V)}$ is the set of all possible sequences of N symbols each belonging to V. When V contains D symbols, there are D^N possible sequences. This provides the probability distribution of any given speech sound being uttered by the speaker who produced sequence X.

In this study, where we assess accommodation in 6 pairs of speakers, HMMs allow for utterances produced by a speaker A – whilst in interaction with a speech partner B – to be tested against A's general speech characteristics to determine if A's speech changes holistically during interaction. This then provides a measure of speech accommodation which accounts for multiple acoustic features. The HMMs used here are a deliberate over-simplification, their use as meant as justification for the further development of this approach.

2. METHOD

2.1. Participants

We assessed the evidence for speech accommodation in 12 participants, organised into 6 pairs. Ages ranged from 19 to 65 (mean 30.92 yrs). All participants gave English as their native language and were screened for normal hearing and eyesight.

Gender and dialect have both been shown to impact phonetic accommodation [2, 7]. For this reason, female-female speaker pairs were used and participants were born and raised in the City of Glasgow.

Similarity attraction has also been theorised to impact accommodation. A protocol for participant pair self selection was designed around literature detailing similarity judgements based on facial features (eg.[19, 16]). This provided two groups, Self-Selected and Randomly Paired.

Measures of personality and interpersonal attraction were taken but are not reported here.

2.2. Task Materials & Experimental Task

The DiapixUK task [5] (an empirically validated spot-the-difference task) was used to elicit free flowing conversation. It consists of twelve images, each with a counterpart that is the same apart from twelve slight differences. Participants had to find the differences between the images, using verbal communication only.

Participants sat in opposite corners of a sound attenuated booth, with a divider between them. They could not see each other but could still hear one another. Each participant had an AKG mono microphone, designed to minimise background speech/noise, recording them. These were fed into separate channels and combined into a stereo signal, with one channel assigned to each participant. Speech was recorded at a sampling rate of 44100Hz. Participants were seated $\sim$30cm away

from a flat screen monitor, adjusted to eye level. DiapixUK images were presented on these monitors in four blocks, within each block participants completed three DiapixUK tasks. Each pair completed twelve DiapixUK tasks in total and here, speech recorded in each task is referred to as an 'Interaction'. The order in which the images were presented were randomised. Stimuli were presented using PsychToolbox [14] in MATLAB®. Data was orthographically transcribed in Praat [6] and force-aligned and segmented in LaBB-CAT [10].

2.3. Hidden Markov Model Analysis

The aim here is to provide a basic proof of concept. As such, analysis has been limited to the most simple form of HMM, a single state HMM. This is simply a mixture of gaussians. Here, a mixture of 10 gaussians is used. If the application of HMMs at their most basic level elicits results, justification will be provided to develop the technique using fully realised HMMs.

2.3.1. Step 1: convert acoustic signal to MFCC

This step provides a form that can account for acoustic properties across instances of the same word. Signal segments with similar acoustic properties are represented by similar vectors. For our purposes, the MFCCs of our data were calculated in the Hidden Markov Model Toolkit (HTK) [20] using the HCopy function. More specifically, the MFCC is derived from discrete Fourier transform based log spectra, [11] provides the specific transformations used by HTK.

2.3.2. Step 2: train speaker models

The first Interaction between two speakers, A & B, is used to *train* the speaker models, i.e. to set the value of the parameters in Λ_A and Λ_B so that the probabilities $p(X_A, S_A|\Lambda_A)$ and $p(X_B, S_B|\Lambda_B)$ are maximized, where X_A is the sequence of all observation vectors extracted from all words uttered by A in the first Interaction (same for B). The training is performed through a mathematical model (the Baum-Welch algorithm) implemented in HTK. In this work, there is one HMM for each speaker. The HMM corresponding to speaker A has one state ($D = 1$) and $p(X, S|\Lambda_A)$ is the probability of speaker A having uttered the words from which the sequence of observations X has been extracted (Λ_A is the parameter set of the HMM corresponding to speaker A).

2.3.3. Step 3: compute likelihood ratio

After models have been trained, it is possible to estimate the probability that a given word w has been uttered by a given speaker: if X_w is the sequence of observation vectors extracted from the speech signal segment corresponding to word w, then $p(X_w, S_A|\Lambda_A)$ is the probability of that word having been uttered by A and $p(X_w, S_B|\Lambda_B)$ is the same probability for B. The right hand side of the following expression:

$$\theta = \frac{p(X_w, S_A|\Lambda_A)}{p(X_w, S_B|\Lambda_B)}$$

is called *likelihood ratio* θ. When $\theta > 1$, it is more likely that the word has been uttered by A than by B and vice versa when $\theta < 1$.

2.3.4. Step 4: correlate time with changes in the speech spectrum

An Interaction can be thought of as a sequence of words uttered either by A or by B. If $w_i^{(A)}$ is the i^{th} word uttered by A, then the following likelihood ratio can be considered a measure of how speaker A becomes more similar to speaker B:

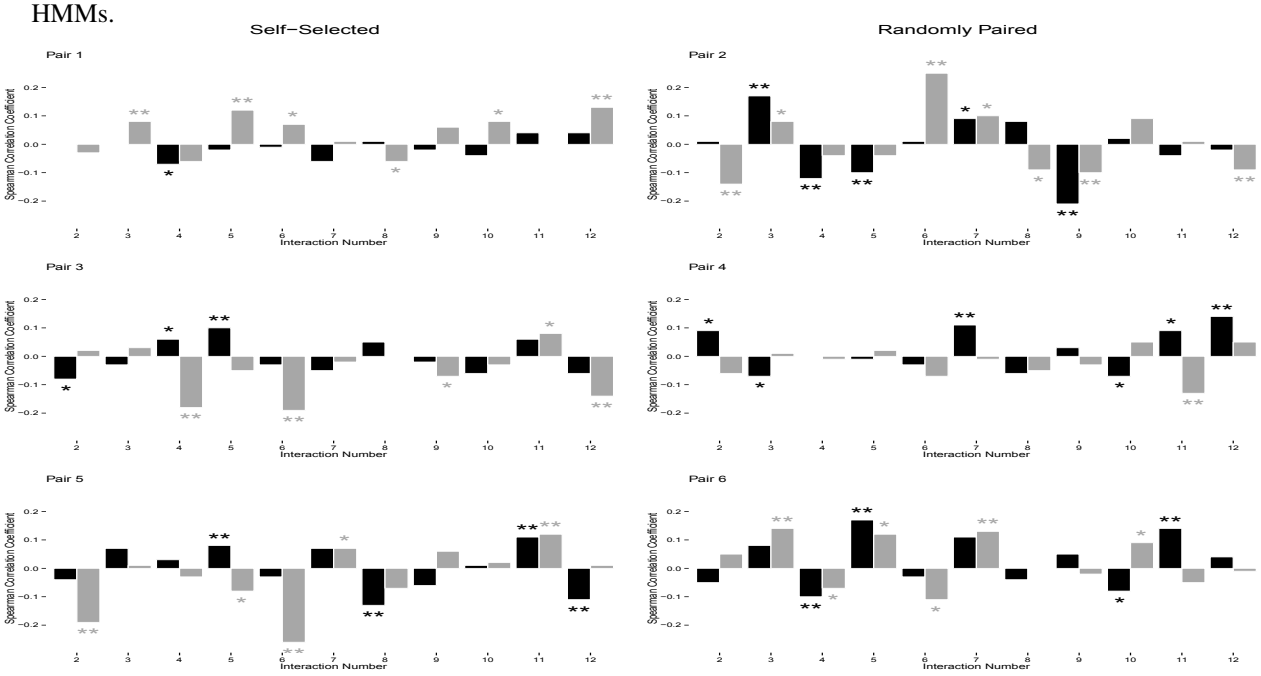
$$\theta_i = \frac{p(X_{w_i^{(A)}}, S_A|\Lambda_A)}{p(X_{w_i^{(A)}}, S_B|\Lambda_B)}.$$

The ratio θ_i can be measured for each word uttered by A resulting in a sequence of pairs (θ_i, t_i), where t_i is the time when word w_i starts. If the correlation between the θ_i's and the t_i's is negative to a statistically significant extent, then A tends to converge to B and vice versa, if the correlation is positive to a statistically significant extent. If the correlation is not statistically significant, then there is no evidence for change. Switching A and B in the expression of the likelihood ratio demonstrates how B shifts with respect to A. Here, we report the results of correlations between pairs of speakers within each Interaction, for the 11 Interactions that they had with each other (the first acts as the training model).

3. RESULTS

Figure 1 shows the results of the analysis for each of the six pairs tested. A significant negative correlation coefficient demonstrates that, over the course of an Interaction, the vectors characterising the speech of a given speaker become more similar to that of their conversational partner. A significant positive correlation coefficient demonstrates that the vectors

Figure 1: Results for the correlations between time and likelihood ratios for each speaker. Dark bars and light bars represent speakers *A* & *B*, respectively, from a given pair. Single stars show statistical significance at the 5% level and double stars at the 1% level. Interaction numbers begin at 2 because the first Interaction is used to train HMMs.



become less similar. For example, interpreting Interactions 2, 3 & 4 of Pair 1 (top left panel) demonstrates that in Interaction 2, speaker *A* did not change whilst speaker *B* had a non significant shift towards their partner. Interaction 3 shows speaker *A* remaining static and speaker *B* significantly shifting away. Interaction 4 shows both speakers shifting towards each other but only speaker *A* shifts significantly.

Results show all pairs demonstrating at least seven instances of a statistically significant shift for at least the 5% level. Pair 2 shows the highest amount of shift with twelve statistically significant effects. Although correlations are small (mean R^2 = ± 0.11, sd = 0.04), this is something that was expected when assessing a subtle phenomenon with holistic measures. However, they remain statistically significant and a binomial test puts the probability of obtaining this result by chance at $\sim 10^{-12}$. This suggests that the results reflect an actual convergence or divergence in this holistic acoustic measure across the course of the interaction.

The measured similarity of participants in a given pair did not impact the results. However, this might be due to the self selection protocol lacking construct validity or because of the small sample size of the groups ($n = 6$).

4. DISCUSSION

The findings of this study demonstrate that using even the most basic of HMMs (i.e. single state; mixture of gaussians) in conjunction with holistic acoustic measures (here, MFCCs) can identify and measure shifts in speech production relative to a conversational partner. It is important to highlight that what is being evaluated here are statistical models of the entire speech spectrum. A comparative analysis using traditional phonetic methods is needed to confirm that the result obtained is a true measure of phonetic accommodation. Heavily simplified HMMs have been implemented here and work is under-way to address this. Developing this approach with more sophisticated HMMs will at least confirm (if not improve) effectiveness. Additionally, the degree to which function words contribute to accommodation was not assessed here. Excluding them may help in training the models as a good deal of social information is lost in function words due to their high frequency.

The method presented here shows promise in detecting accommodation in conversation but development and modifications are clearly needed to confirm its applicability. If this approach proves capable of detecting accommodation, it could be a valuable tool for uncovering the processes contributing to speech accommodation during interaction.

5. REFERENCES

[1] Babel, M. 2009. Selective vowel imitation in spontaneous phonetic accommodation. *UC Berkeley Phonology Lab Annual Report (2009)* 163–194.

[2] Babel, M. 2010. Dialect divergence and convergence in new zealand english. *Language in Society* 39(04), 437–456.

[3] Babel, M., Bulatov, D. 2012. The role of fundamental frequency in phonetic accommodation. *Language and speech* 55(2), 231–248.

[4] Babel, M., McAuliffe, M., McGuire, G. 2014. Spectral similarity and listener judgments of phonetic accommodation. Proceedings of the 10th International Seminar on Speech Production, Cologne, Germany.

[5] Baker, R., Hazan, V. 2011. Diapixuk: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods* 43(3), 761–770.

[6] Boersma, P., Weenink, D. Dec. 2012. Praat: doing phonetics by computer [computer program].

[7] Bulatov, D. 2009. The effect of fundamental frequency on phonetic convergence. *Berkeley Phonology Lab Annual Report* 2009, 404–434.

[8] Casasanto, L. S., Jasmin, K., Casasanto, D. 2010. Virtually accommodating: Speech rate accommodation to a virtual interlocutor. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* 127–132.

[9] Finlayson, I., Lickley, R., Corley, M. 2012. Convergence of speech rate: Interactive alignment beyond representation'. *Twenty-Fifth Annual CUNY Conference on Human Sentence Processing, CUNY Graduate School and University Center, New York, USA* 24.

[10] Fromont, R., Hay, J. 2008. Onze miner: the development of a browser-based research tool. *Corpora* 3(2), 173–193.

[11] Ganchev, T., Fakotakis, N., Kokkinakis, G. 2005. Comparative evaluation of various mfcc implementations on the speaker verification task. *Proceedings of the SPECOM* volume 1 191–194.

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

[13] Huckvale, M. 2007. ACCDIST: An accent similarity metric for accent recognition and diagnosis. In: Müller, C., (ed), *Speaker Classification II* Springer Lecture Notes in Computer Science. Springer Berlin Heidelberg.

[14] Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C. 2007. What's new in psychtoolbox-3. *Perception* 36(14), 1–1.

[15] Mistry, D. S., Kulkarni, A. 2013. Overview: Speech recognition technology, mel-frequency cepstral coefficients (mfcc), artificial neural network (ann). *International Journal of Engineering Research and Technology* volume 2. ESRSA Publications.

[16] Olivola, C. Y., Funk, F., Todorov, A. 2014. Social attributions from faces bias human choices. *Trends in Cognitive Sciences* 18(11), 566 – 570.

[17] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119, 2382.

[18] Pardo, J. S., Gibbons, R., Suppes, A., Krauss, R. M. 2011. Phonetic convergence in college roommates. *Journal of Phonetics*.

[19] Todorov, A., Olivola, C. Y., Dotsch, R., Mende-Siedlecki, P. 2015. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology* 66(1), 519–545. PMID: 25196277.

[20] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. C. 2006. *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department.

6. ACKNOWLEDGEMENTS

This work was supported by the University of Glasgow's Kelvin-Smith PhD Scholarship.