Speaker discriminability for visual speech modes

Jeesun Kim 1, Chris Davis1, Christian Kroos1, Harold Hill2

1 MARCS Auditory Laboratories, University of Western Sydney, Australia
2 School of Psychology, University of Wollongong, Australia

j.kim@uws.edu.au, chris.davis@uws.edu.au, c.Kroos@uws.edu.au, harry@uow.edu.au

Abstract

Does speech mode affect recognizing people from their visual speech? We examined 3D motion data from 4 talkers saying 10 sentences (twice). Speech was in noise, in quiet or whispered. Principal Component Analyses (PCAs) were conducted and speaker classification was determined by Linear Discriminant Analysis (LDA). The first five PCs for the rigid motion and the first 10 PCs each for the non-rigid motion and the combined motion were input to a series of LDAs for all possible combinations of PCs that could be constructed using the retained PCs. The discriminant functions and classification coefficients were determined on the training data to predict the talker of the test data. Classification performance for both the in-noise and whispered speech modes were superior to the in-quiet one. Superiority of classification was found even if only the first PC (jaw motion) was used, i.e., measures of jaw motion when speaking in noise or whispering hold promise for bimodal person recognition or verification.

Index Terms: Visual speech, speaker recognition, speech modes.

1. Introduction

Biometric person recognition or verification involves using an intrinsic or behavioral trait of a specified user to allow him/her to access to a closed system without the need for a password, physical key, identification card, etc. Some sources of biometric information offer the potential for unobtrusive automatic person recognition and have enormous potential for security applications. To be effective, a biometric marker needs to have universality, distinctiveness, and permanence along with being collectible and socially acceptable.

Typical biometric markers involve measuring an aspect of a person’s physiology (e.g., fingerprint). Such physiological markers, although reasonably precise, typically have negative properties, such as being intrusive and requiring compliance on behalf of the user. Other behavioural-based biometrics (e.g., voice, gait recognition) can be cheaper, less intrusive and more covert, but characteristically are less robust [1].

Identification based upon behavioural information obtained via a single modality tends to be fragile since the signal can be distorted (e.g., background noise, signal distortion, uneven lighting, etc.). To mitigate this problem data from multiple modalities can be used. Here, the limitations in unimodal based measures can be offset by using information from a modality providing complementary information.

Two signals have attracted the bulk of attention with regards to person recognition: the face and the voice. Both auditory and visual speech provide signals regarding what is said and who said it and an audio-visual combination has the potential to provide complimentary information. Although the ways in which multimodal signals might be combined have been well studied, with a taxonomy developed categorizing signal, feature, model, score and decision levels and combination rules [2], less research has been conducted on what should be measured and under what circumstances this measurement should take place (especially for visual speech).

With regards to audio, there is a general convergence on methods for deriving auditory features for speaker recognition, e.g., cepstral features extracted over short time spans combined with the use of complex classifiers, normalization procedures and the use of adaptation and these are the subject of on-going research and development [3]. However, the situation is more complicated for the visual speech signal, since not only a feature extraction method but also the specific source(s) of visual signals to analyze need to be decided. That is, some studies have used the whole head or head profile [4], whereas other have used aspects of the face [5] or face features [6], still others have just concentrated on the mouth and lips [7]. The majority of studies have used static information (although some studies have investigated dynamic information, particularly as a defense against replay attacks [8]) and the extraction of visual speech features has either been through shape/model based geometrical features (following [9]) or via appearance (pixel value) based approaches using a region of interest (ROI).

Rather than using image features or appearance information, the current study employed 3D motion data as input since we are interested in the dynamic movements associated with visual speech. The dynamics of visual speech provide an important biometric measure because they combine the advantages of behavioural and physiological data [10]. It has been demonstrated that head and face movements play a role in person identification by humans. For instance, facial motion affects face identification as shown by morphing one person’s face into another’s and superimposing non-rigid facial movement [11]. Furthermore there is a tight coupling between non-rigid and rigid head motion [12], suggesting that rigid head motion might provide information about a person’s identity. Indeed, it has been shown that rigid motion is useful for distinguishing between individuals [13]. However, it might be that some speech modes provide more distinctive head motion information than others. For example, it has been shown that people are much better at correctly matching auditory speech to head motion for expressive sentences that have greater head motion than non-expressive ones [14].

To date, there have been no studies that have explicitly examined the distinctiveness of visible speech dynamics as a function of speaking mode. Speech mode (e.g., Lombard speech or whispering) is an important variable to investigate in regard to the potential of visual speech information to be a useful biometric since it has been shown that the dynamics of visual speech are exaggerated in Lombard speech [15] or
when whispering, compared to talking in quiet [16]. Exaggeration of biological movements such as facial expression and arm movements has been shown to facilitate person recognition judgments and emotion identification [17]. These effects are thought to only occur when movements critical to task performance are exaggerated. Thus, our approach is to measure visual speech movements in conditions that have been previously shown to produce exaggerated head and face movements (natural exaggeration). That is, we measured speech-related head and face movements when the talker was whispering, talking over noise or talking in quiet. Based upon the above, we predicted that these movements would provide better discrimination between individual talkers for the whispered and talking in noise speech modes.

This study is an exploratory one, aimed at examining whether different speaking styles affect the degree to which talkers can be discriminated by the dynamics of their visual speech (and whether one style does better than the rest). To this end, we concentrated on using a small, well-controlled corpus of sentences and a small number of talkers, with the dynamic visual speech features to be used for separating talkers defined by the data (using Principle Components Analysis). It is important to note that the current study is not aimed at proposing a specific implementation of a speaker identification system based on visual speech features; the aim is to make systematic 3D measures in a controlled fashion in order to investigate whether speech mode has an effect on visual speech speaker discriminability.

2. Experiment

2.1. Method

Talkers. Visual speech was obtained from four people (3 males, 1 female). All were native speakers of English (one British, two Australian and one American); ages ranged from 32 to 54 years.

Materials. Ten sentences were selected from the 1965 revised list of phonetically balanced sentences (Harvard Sentences, 1965).

Noise. To induce Lombard speech, two types of background noise were employed, multi-talker babble and white noise. A commercial babble track (Auditec, St. Louis, MO) was used; the flat white noise was generated at the same average RMS power as the babble track. The noise was presented to participants either through ear plugs or through two loud speakers (Yamaha MS 101-II). The conditions will be referred to as BabbleP and BabbleLS for the babble noise ear plug and loud speaker conditions and WhiteP and WhiteLS for the white noise plug and loud speaker conditions.

Apparatus for data capture. Two Northern Digital Optotak machines were used to record the movement data. Sound was captured from both a head-mounted (Share SM12A) and a floor microphone (Semhheiser MKH416P48U-3). Video was captured using a digital camera (Sony HDCAM HKDW-702).

Procedure. Each session began with the placement of the movement sensors during which time participants were asked to memorize the ten sentences to be spoken. Each talker was recorded individually in a session that lasted approximately 90 minutes. Talkers were seated in an adjustable dentist chair in a quiet room and were asked to say aloud ten sentences (one at a time) to a person who was directly facing them at a distance of approximately 2.5 metres. The talker then repeated the ten sentences. This basic procedure was repeated five times more, once for each speech mode condition. These conditions consisted of the talker speaking while hearing multi-talker babble through a set of ear plugs (at approximately 80 dB SPL); hearing the same babble through two Loud Speakers; hearing white noise through ear plugs (at approximately the same intensity); hearing white noise through the Loud Speakers; and finally whispering the sentences (at a level judged loud enough for the conversational partner to hear).

Data processing. Non-rigid facial and rigid head movements were extracted from the raw marker positions and movements of the head rig. The data were recorded at a sampling rate of 60Hz. Each frame was represented in terms of its displacement from the first frame and Principle Component Analysis (PCA) was used to reduce the dimensionality of the data (PCAs were calculated for rigid and non-rigid motion and both combined). The auditory analysis was carried out using the Praat program [18] on the in-quiet and ear plug conditions (the latter were uncontrolled by noise). In order to characterize the contribution of the Principle Components (PCs), absolute values of each PC were summed to represent the amount that each PC contributed to head and face movements over time. These data (“PC strength”) were used as the dependant measure in a series of ANOVAs to determine whether there were differences in the amount of movements across speech modes, sentences and talkers.

In the analysis that follows we chose to concentrate on the first 10 PCs (see below). The combined rigid and non-rigid PCs motions were visualized using a Tcl/Tk GUI in which the 3D motion of each PC could be viewed. The following gloss of what each PC represents are of course approximate verbal descriptions: PC1 could be described as jaw motion (and mouth opening); PC2 as lip rounding (without jaw motion) and eyebrow raising; PC3 as head translation (towards and away from the interlocutor); PC4 as lip protrusion; PC5 as mouth opening and eyebrow closure (e.g., PC2); PC6 as pitch; PC7 as roll; PC8 as yaw with negative tilt; PC9 as yaw with positive tilt; PC10 as yaw.

Linear Discriminant Analysis. To ensure generalisibility of the speaker classification results a jackknife procedure was employed. The whole data set containing the motion data of 10 sentences (with two repetition of each sentence, although in three cases data was only available for one repetition and the single value available was used) spoken by 4 talkers was divided into a training set (9 sentences, all talkers and instances) and a test set (1 sentence, all talkers and instances). The training data were subjected to PCA and the scores of the components (PC scores) computed using Matlab (The MathWorks, Inc). The PCA direction cosines were then used to project the test data into the PC space, i.e., their PC scores were computed based on the PCA model of the training data. This was repeated 10 times with a different sentence constituting the test set every time. Separate PCAs were conducted for the rigid head motion (6 parameters: 3 translational, 3 rotational), the non-rigid face motion (24 markers with x-, y-, and z-coordinates after removal of the impact of the rigid head motion) and both motion types combined. Note that for the combined data set rigid and non-rigid motion parameters were concatenated and not simply the raw data used in order to allow a better interpretability of the PCs.

The PC scores were summed per sentence instance and divided by sentence duration to normalize for different sentence lengths. The converted scores made up the independent variables in the LDA while the grouping variable was the talker ID. After investigating the amount of recovered
variance by each PC it was decided that the first 5 PCs for the rigid motion (99.38% of overall variance recovered averaged over the ten training sets) and the first 10 PCs for both other motion types (96.55% for the non-rigid motion, 95.19% for the combined) should be retained as input for the LDA. For the rigid motion there is only one PC left that can be considered as the ‘garbage’ collector, but for the non-rigid and combined data sets higher numbered components might still contain useful information for talker separation.

Since we wanted to get a broad picture of the influence of speaking mode on speaker separation, it was decided to apply the LDA to all possible combinations of PCs that could be constructed using the retained PCs, e.g., for the rigid head motion \{1\}, \{2\}, \ldots \{1,2\}, \{1,3\} \ldots \{2,3\}, \{2,4\} \ldots \{1,2,3,4,5\}. This resulted in 124,620 separate LDAs: (31 rigid + 1023 non-rigid + 1023 combined) x 6 speaking modes x 10 cross-validation data sets. The results were averaged across the ten training/test sets yielding 12,462 data points in the end.

As for the LDA itself, the discriminant functions and classification coefficients were determined on the training data and were then used to predict group membership of the test data (the number of correct classifications of each of the 8 test items (1 sentence x 2 instances x 4 talkers) in each LDA were converted to percentages). The R statistical computing package (R Foundation for Statistical Computing) and in particular the LDA function from the MASS extension package were used for all LDA related computations.

### 2.2. Results

Analysis of the auditory data indicated that there was a Lombard effect for speech produced in the noise conditions, i.e., this speech was louder than that produced without background noise \(F(1,78) = 365.41, p < 0.05\). On average the size of this effect was 11 dB. The length of the renditions also increased for the in-noise conditions compared to no-noise “in quiet” speech \(F(1,78) = 33.4\) with an average increased production time of 290 ms (the average in-quiet utterance duration was 2.1 seconds). Whispered speech was also longer than in-quiet speech (by an average of 310 ms).

Discriminant and ANOVA were conducted to determine whether better discrimination based on the visual speech of the talkers could be obtained for speech produced in-noise or whispered compared to that produced in-quiet. Discriminant analysis for the combined motion PCs showed that over all comparisons, classification performance for both the in-noise (except for the WhiteLS condition) and whispered speech modes were superior to the in-quiet one.

An ANOVA was conducted with the factors of Motion Type (Rigid, Non-rigid, Combined) and Speaking Mode (in-quiet, BabbleLS, BabbleP, WhiteLS, WhiteP and Whispered) with percent correct discrimination as the dependent measure. The main effect of Speaking Mode was significant \(F(5, 10370) = 155.06, p < 0.001\) as was Motion Type \(F(2, 2074) = 21.70, p < 0.001\) and the interaction of these variables \(F(10, 10370) = 124.86, p < 0.0001\). Each of the comparisons against the in-quiet percent correct score showed a significant benefit [BabbleLS vs. in-quiet, \(F(1,2076) = 734.5, p < 0.00001\); BabbleP vs. in-quiet, \(F(1,2076) = 818.6, p < 0.00001\); WhiteP vs. in-quiet, \(F(1,2076) = 2330.9, p < 0.0001\); Whispered vs. in-quiet, \(F(1,2076) = 127.6, p < 0.001\), however, speech in the WhiteLS condition was actually slightly worse than in-quiet \(F(1,2076) = 241.6, p < 0.001\).

The better discrimination for exaggerated head and face movements shows that these movements were not just different (e.g., bigger) compared to those produced when talking in quiet, but that they were more characteristic of each talker (this was particularly the case for the WhiteP condition). Figure 1 shows the percent correct discrimination for the in-noise and whispered conditions minus the in-quiet performance as a function of adding subsequent PCs (up to PC10) to PC1. As can be seen, there was better performance even PC1 alone and the more PCs added after the first seven led to a decrement in performance relative to in-quiet speech (this effect was quite dramatic for the WhiteLS condition).

![Figure 1](image1.png)

**Figure 1.** Shown is the percent increase in correct discrimination compared to the in-quiet condition for each of the in-noise conditions and the whispered condition as a function of the linear combination of (combined) PCs (from PC1 leftmost column in each condition to PC1+PC2, PC1+PC2+PC3 \ldots for each subsequent column).

Figure 2 shows the correct classification scores (relative to the in-quiet condition) broken down by Movement Type. As can be seen, for rigid motion the only conditions that showed increased discrimination were the Whispered and WhiteP ones. For Non-rigid motion all the in-noise conditions (except WhiteLS) showed an increase, but whispered speech did not. Finally, for the combined movement condition, all the exaggerated conditions produced better discrimination than did the in-quiet one.

![Figure 2](image2.png)

**Figure 2.** Increase in percent correct LDA performance (relative to the in-quiet scores) for all LDAs as a function of Speaking Mode and Movement Type.
To investigate whether the lack of enhanced discrimination in the WhiteLS condition (compared to the in-quiet one) was because this condition did not produce exaggerated motion, the magnitudes of the combined movement PCs for all the in-noise conditions (as well as the whispering condition) were compared to the in-quiet condition. The difference in motion between the in-noise conditions (and also the whispered one) when compared to the in-quiet one can be characterized as consisting of a significant increase in all 10 PCs; with very prominent changes in jaw (PC1) and mouth motion and face expansion (PC2), increases both in lip protrusion (PC4), mouth and eyebrow closure (PC5) and head rotation (PC6). This was confirmed in the following analyses. There was a main effect of speech mode (in-noise vs. in-quiet) for each contrast: BabbleP, \(F(1,215) = 100.48, p < 0.0001\); BabbleLS, \(F(1,157) = 83.07, p < 0.0001\); WhiteP, \(F(1,157) = 72.57, p < 0.001\) (and for whispered speech movements vs. in-quiet speech \(F(1,157) = 57.50, p < 0.005\)). This was also the case for the WhiteLS condition vs. in-quiet speech, \(F(1,160) = 65.52, p < 0.005\). Speech in-noise also produced a different pattern of PCs compared to in-quiet speech as indicated by significant interactions between the main effects of noise vs. in-quiet and the difference PCs. This interaction occurred for BabbleP vs. In-quiet x PCs, \(F(9,1933) = 27.39, p < 0.05\); BabbleLS vs. In-quiet x PCs, \(F(9,1415) = 14.33, p < 0.005\); WhiteP vs. In-quiet x PCs, \(F(9,1413) = 15.23, p < 0.005\) (and between the whispered speech mode vs. in-quiet x PCs \(F(9,1422) = 11.12, p < 0.05\)). Once again, the WhiteLS condition showed a similar pattern to the other condition: WhitelS vs. In-quiet x PCs, \(F(9,1440) = 12.02, p < 0.005\). Thus in terms of the exaggeration of motion (cf., the in-quiet condition), the WhiteLS condition appeared to be similar to the other noise conditions.

However, unlike the other conditions, the WhiteLS condition involved exposing both the talker and listener to white noise. Under this challenging condition it is possible that the talker rushed their renditions and thus the way his/her speech was exaggerated was not natural. Support for this proposition comes from an examination of the time taken to utter each sentence in the WhiteLS condition, as this was on average (and for almost every sentence) shorter than those uttered in quiet.

3. CONCLUSION

We examined the first 10 PCs of head and face movements for talkers speaking in different types of background noise, in quiet and whispering. In general, it was found that the largest changes from the in-quiet to noise conditions were in jaw and mouth movements, face expansion and contraction and head rotation. The LDA showed that classification performance for the combined motion PCs for both the in-noise and whispered speech modes were superior to the in-quiet one. This benefit in talker classification performance was found even if only the first PC was used and suggests that measuring jaw motion for talkers speaking in noise or whispering holds promise for one component of bimodal person recognition or verification. In sum, natural exaggerations of facial speech enhance talker identity specific information.

4. Acknowledgements

This work was supported by the Australian Research Council (DP0666857 & TS0669874).

5. REFERENCES