Automatic Recognition of Attitudes in Video Blogs - Prosodic and Visual Feature Analysis

Noor Alhusna Madzlan[1 3], *JingGuang Han*[1], *Francesca Bonin*[1 2], *Nick Campbell*[1]

[1] CLCS, School of Linguistics, Speech and Communication Sciences, Trinity College Dublin
[2] SCSS, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3] ELLD, Faculty of Languages and Communication, UPSI, Malaysia

madzlann@tcd.ie, hanf@tcd.ie, boninf@tcd.ie, nick@tcd.ie

10.21437/Interspeech.2014-415

Abstract

This paper reports a study of attitude manifestations in video blogs. We describe the manual annotation of speaker attitudes in a corpus of over 130 video blogs and present an analysis of prosodic and visual cues in relation to attitude states. We use machine learning techniques for the automatic prediction of attitudes from prosodic and visual features in video blogs and compare the performance of prosodic and visual feature sets.

Index Terms: video blogs, prosody, face detection, attitude, automatic classification, statistical modelling

1. Introduction

Social media has become a popular form of personal expression. While the popularity of content based platforms, such as web blogs, Twitter and Facebook, confirms that written text is still the major form of online interaction, new forms of expression are evolving, and conversational video blogs (vlogs for short) are now a widespread phenomenon of online social media, which create a huge amount of user generated content.

Video blogs are a unidirectional form of communication where the user intends to convey a message, an emotion or a personal opinion. They are personal diaries made available to the larger public in the form of self-recorded videos, and combine the high quality of broadcast speech with the naturalness of spontaneous conversations. As such they provide ideal material for scientific research into communicative aspects of verbal and paraverbal behaviour.

Unlike face to face interaction, the speaker performs in a de-contextualized situation, addressing an imagined audience without being influenced by any listener reactions. Listener feedback can only be provided later, in the form of textual comments, and it does not influence the speaker's behaviour.

Unlike broadcast speech, video blogs can be considered semi-spontaneous; while a script might be drawn up before the recording, the video blogger is free to improvise and adapt the dialogue as in a natural conversation. Hence, video blogs display typical spontaneous speech phenomena such as disfluencies, hesitations, ungrammaticality, filled pauses, repetitions and false starts similar to natural face to face conversations.

Researchers have shown interest in this novel form of communication in the study of personality recognition [1, 2] and they have been widely studied with respect to non-verbal behavior and social attention, but to our knowledge, the pragmatic aspect of this new form of expression has been given little attention. Non-verbal cues in video blogs are interesting to explore as they provide indicators for the understanding of human behaviour, specifically attitudinal expression.

In this work we investigate the expression of *attitudes* in video blogs, and we define attitude as social affective states that the video bloggers intend to transmit. We are not interested in discovering any inner emotions of the video blogger, but in how he or she behaves and how that influences the audience at a pure pragmatic level. While in a previous work [3], we concentrated our analysis on the prosodic parameters, in this paper we present the analysis of both prosodic and visual features, exploring their interaction and predictive performances. To the best of our knowledge, no previous work has addressed presenter's attitudes in vlogs. In the present study, we compare the predictive values of different multimodal features to analyse vlogger attitudes.

We collected a corpus of 134 video blogs, and annotated them according to five attitudinal classes to explore the acoustic and visual characteristics of the different classes. This paper reports the results of a comparative analysis of the features and shows that they vary consistently in a way that allows us to interpret (and to predict) the attitude of the speaker from physical characteristics of the signal. This work will ultimately enable us to predict better indices for digital search, and controls for multimodal synthesis for ECAs in human computer interaction systems.

Main contributions of this work: *i)* We describe a novel corpus of video blogs and its annotation with respect to attitudes. *ii)* We analyse prosodic and visual features of attitude impressions in video blogs. *iii)* We report the results of a comparative analysis of acoustic and visual features in automatically classifying vloggers' attitudes. The paper is structured as follows: Section 2 summarises the literature on video blogs. Section 3 outlines the characteristics of the dataset. Section 4 describes the annotation process and adaptation of the annotation schema. Section 5 reports the prosodic and visual feature analyses, and a comparative analysis of the predictive performances of different sets of features. Conclusions are drawn in Section 6.

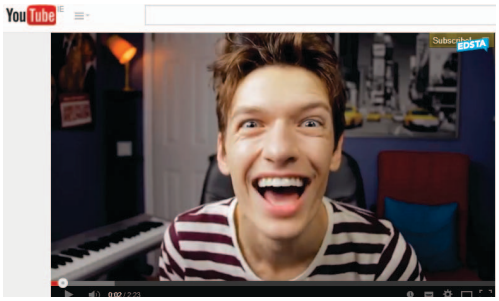

Figure 1: Showing a typical video blogger.

2. Related work

Video blogs have piqued the attention of many researchers from several different perspectives [1, 2, 4].

Studies have focused on the detection of emotion in vlogs. Biel et al [5] conducted a study on detecting emotion through facial expressions for the recognition of personality impressions. They examined facial activities of video bloggers using the Computer Expressed Recognition Toolbox (CERT) and extracted eight emotions: Anger, Joy, Contempt, Disgust, Surprise, Neutral, Sad and Fear.

De Mores et al [6] conducted a study on the prosodic viewpoint of attitude in Brazilian Portuguese. They referred to prosodic attitudes as the voluntary control of speakers' expression of social affects and described groups of attitudes, assertive, neutral, social and propositional attitudes that were expressed by two native Brazilian Portuguese speakers. A perception test was conducted for attitude recognition using audio, visual and combined audiovisual modalities. Results showed that audio signals obtained lower scores for social attitude recognition while both audio and visual modalities obtained high scores for propositional attitude recognition.

Henrichsen and Allwood [7] analysed the NOMCO speech corpus for detecting attitude states. They defined an annotation schema based on ten attitudes called the A10-based annotation, which includes Interested, Friendly, Casual, Bored, Thoughtful, Confident, Amused, Enthusiastic, Uninterested and Impatient, and they implemented a classifier for automatic attitude prediction.

Biel and Gatica Perez [8] conducted a study on automatic analysis of human behavior in video blogs, analysing prosody and gesture and their correlation with social attention. From their qualitative analysis the conversational nature of video blogs was confirmed, exemplified by natural spontaneous speech phenomena including hesitations and disfluencies. Speaking time, turns and length of utterance showed significant correlations with social attention. Speakers maintain proximity and eye contact with the camera. Speakers show higher frequency of looking at the camera when speaking. This type of behaviour is closely related to dominance characteristics, and the speaker is perceived to be receiving more attention from his social circle.

Weninger et al. analysed speaker trait characterization in web videos [9]. They adopt an approach based on multimodal features for recognition of character traits. They compared the classification performance of different feature sets (acoustic, linguistic and visual) to three classes (age, gender and race). Visual features showed higher performances for the classification of speaker age, and acoustic features performed best in detecting gender. The linguistic feature set performed well for recognising the race of the speaker.

3. Video Blog Dataset

A total of 134 video blogs were selected from YouTube. For ease of analysis, we selected a homogeneous data sample in terms of speaker characteristics. Eight American English speakers were selected, all male, between 18-25 years old.

The videos were downloaded using a free add-on tool for Mozilla Firefox[1]. The audio tracks were extracted from each video. Manual annotation and labeling was conducted using WaveSurfer [10]. Audio sample rate is 44,100 with mono sound. Total duration of the corpus is 429 minutes, and the mean duration of the selected videos is 3.26 minutes (sd=1.37).

All the videos present typical video blog characteristics. They are pre-recorded monologues, with the speaker facing the camera. This type of interaction allows for delayed feedback, and listener response is provided in comment boxes below each video. These video blogs are stored in reverse chronological order to present the most recent updates to viewers.

While in broadcast speech, the content and delivery are typically prepared and scripted with a fixed standard format that speakers need to follow, video blogs allow room for versatility in both content and delivery. Video bloggers' speech exhibits the spontaneity of face to face conversation specifically with relevance to disfluencies, hesitations, ungrammaticality, filled pauses, repetitions and false starts. These characteristics make vlogs a valuable source for the analysis of non-verbal signals, such as laughter [11, 12, 13], gazes [14], etc. The following is an example of video blog speech:

> *"Oh what is going on YouTube? Now I recently read a post about something quite disturbing. Once again at the institution of higher learning, we are faced with this again."*

Attitude	Description
Amusement	speaker laughs, chuckles
Impatience	speaker shouts, appears annoyed, harsh
Friendliness	speaker informally addresses the audience
Enthusiasm	speaker appears excited
Frustration	speaker sighs, appears 'defeated'

Table 1: showing the Attitude Annotation scheme we used

4. Annotation process

In order to understand the characteristics of speaker attitudes in video blogs we annotated the corpus to identify the attitudinal states. We used the annotation labels shown in Table 1, a subset of the A10 annotation from [7].

The manual annotation lasted six weeks, and we annotated in several stages[2]. We asked two annotators to label sections of the corpus according to the five attitudes. Inter-annotator agreement was calculated using Cohen's Kappa, and resulted in k=0.75 [15].

After labeling, the chunk start and end times were obtained using a TCL/TK script. We then used these start and end times extracted from the script to segment the visual data. Video frames were segmented using the split tool in Microsoft Windows Movie Maker.

5. Feature Analysis

In this section we present the feature extraction process and an analysis of the influence of several acoustic and visual features.

5.1. Prosodic features

The following acoustic parameters were extracted using a TCL/TK script:

- Fundamental Frequency (fmax, fmean, fmin) - f0 measured across each chunk

[1] https://addons.mozilla.org/en-US/firefox/addon/easy-youtube-video-download/

[2] The annotations of the corpus and links to the videos has been made public at the following link http://fastnet.netsoc.ie/noor/vlog-data.zip

Type	fmean	fmax	fmin	fpct	fvcd	pmean	pmax	pmin	ppct	h1-h2	h1-a3	h1	a3	dn
Unit	*Hz*	*Hz*	*Hz*	*%*	*%*	*dB*	*dB*	*dB*	*%*	*dB*	*dB*	*dB*	*dB*	*sec*
Amu	168.4 *(35.42)*	231.5 *(57.23)*	108.14 *(32.58)*	45 *(25)*	58 *(20)*	59.28 *(6.18)*	74.46 *(5.74)*	35.92 *(11.10)*	45 *(28)*	6.70 *(6.08)*	30.43 *(12.61)*	-24.19 *(9.47)*	-54.61 *(6.79)*	1.04 *(0.30)*
Ent	220.21 *(49.10)*	298.00 *(59.23)*	131.48 *(57.20)*	35 *(28)*	70 *(27)*	63.62 *(6.68)*	78.03 *(4.21)*	36.56 *(13.18)*	34 *(23)*	6.25 *(5.65)*	27.91 *(11.74)*	-27.32 *(0.75)*	-55.24 *(5.21)*	1.25 *(0.73)*
Frn	197.6 *(32.22)*	248.1 *(45.50)*	148.63 *(40.72)*	26 *(14)*	74 *(12)*	64.71 *(4.10)*	76.72 *(2.38)*	37.78 *(16.43)*	24 *(20)*	4.08 *(4.80)*	20.69 *(16.11)*	-32.32 *(18.21)*	-54.52 *(5.66)*	0.52 *(0.12)*
Fru	116.53 *(28.81)*	155.4 *(42.21)*	84.86 *(27.91)*	32 *(25)*	48 *(23)*	51.87 *(8.26)*	67.71 *(7.72)*	29.22 *(13.09)*	43 *(28)*	3.57 *(5.07)*	29.27 *(10.74)*	-26.53 *(8.89)*	-55.80 *(5.24)*	1.22 *(2.35)*
Imp	234.4 *(40.61)*	312.1 *(47.05)*	142.4 *(49.39)*	39 *(27)*	55 *(23)*	63.77 *(4.24)*	78.96 *(2.69)*	35.84 *(12.43)*	35 *(25)*	5.35 *(5.97)*	28.55 *(14.81)*	-26.12 *(11.61)*	-54.66 *(6.27)*	1.37 *(0.65)*

Table 2: Prosodic Mean values for each attitude category with standard deviation in brackets. (Amu: amusement, Ent: enthusiasm, Frn: friendliness, Fru: frustration, Imp: impatience). *These values represent the position over the duration of the chunk

- Approximate shape of the pitch contour (fpct) - position of the peak as fraction of utterance length
- Voicing (fvcd) - percentage of voicing (vocal fold vibration) across each chunk
- Power/Intensity (pmax, pmean, pmin) - amplitude of the waveform in decibels
- Power/Intensity movement (ppct) - peak position of power (rising/falling)
- Voice Quality (H1-a3) - tenseness of the voice [h1-h2,h1,a3] [16]
- Duration (dn) - length of the utterance

In addition to the traditional prosodic parameters; pitch, intensity, and duration of the segments, voice quality is considered to be a relevant acoustic parameter for communicative speech analysis as it has significant correlates with the interlocutor, speaking style and speech act [17]. In Table 2, we report the average and standard deviation of the extracted parameters for each category.

From ANOVA examination of speaker dependent acoustic features, we found no significant difference among speakers for the majority of features except Speaker A who differed in Voice Quality and Speaker E who differed in Intensity. These speakers were excluded from the analysis of those features.

The prosodic analysis of the attitudes showed 'Impatience' having the highest pitch and 'Frustration' the lowest. From ANOVA examination of attitudes, we found that 'Frustration' and 'Amusement' differ significantly in terms of pitch from the other categories (lower pitch, $p<0.005$). 'Frustration' is also represented by speech with low intensity (significantly differing from the others, $p<0.005$), while, as expected, Impatience is the attitude characterized by a higher intensity.

'Impatience' and 'Enthusiasm' show similar distributions in terms of pitch, with a similar high average pitch. To gain a better understanding of the influence of the different prosodic parameters, we conducted a principal component analysis (PCA) which revealed component 1 (PC1-) to be mainly influenced by variations in pitch. We confirmed the strong influence of voice quality which accounted for component 2 (PC2) and 3 (PC3), see Fig. 2.

5.2. Visual features

Attitudes and emotions are also expressed through rich facial expressions [18]. We used an image processing approach, the Active Appearance Model (AAM), for analysing and measuring the dynamics of the vlogger facial movements. Fig. 3 shows an

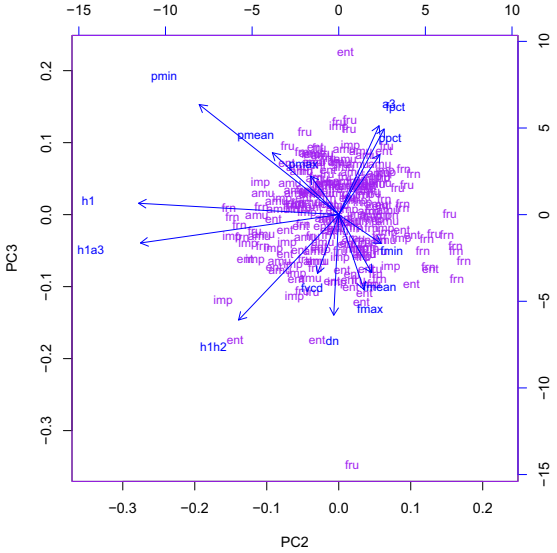

Figure 2: Principal Component Analysis for Audio features, showing voice quality features (h1-a3, lower left quadrant) clearly distinguished from other prosodic parameters in the PC2 dimension.

example of face detection visualization. We extracted 65 landmarks on the face of the speaker (see Fig. 3) and tracked the movement of each landmark to determine the average movement of each. Note that we do not consider the direction or the shape of the movement, but for this work we only consider the amount of movement in a given attitudinal segment for a specific landmark. Although the direction and the shape of the movement could give interesting information, this processing is

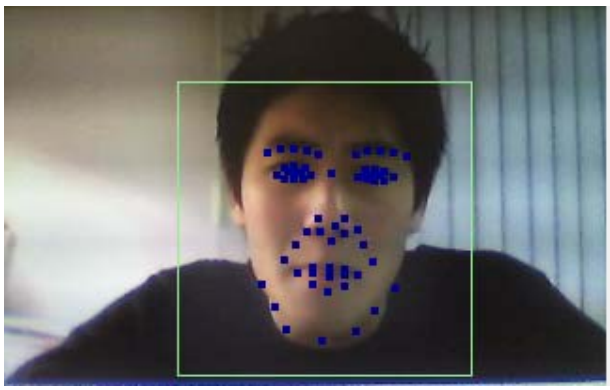

Figure 3: Active Appearance Model

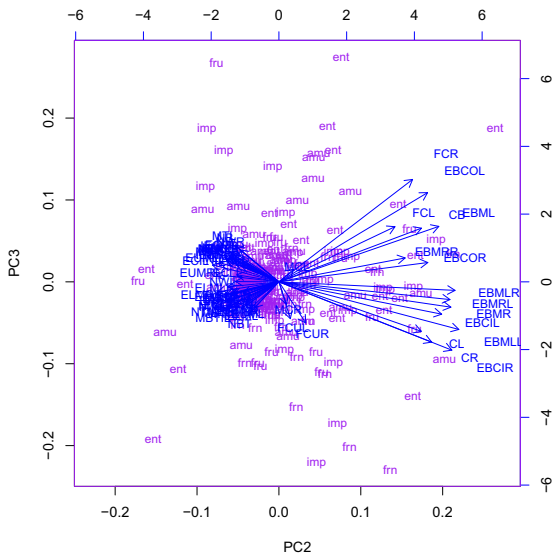

Figure 4: Principal Component Analysis for Visual features. We can clearly see the eyebrows (right cluster) separated from all the other facial features (left clusters) by PC2. Eyebrows are indicated with the code name EB*.

reserved for future work.

We consider the amount of movement of the entire head as represented by the point at the top of the nose, and subtract this movement from every other point's value to consider the movement of each landmark point independently from the general head movement.

Fig. 4 shows the PCA visual features analysis. PC1 is heavily influenced by overall head movement, so we show the second and third components. We noted that the second and third components show a distinct separation of eyebrow features from the rest of the facial features.

We used an interactive 3Dplot to explore the relation of the features in the various coordinate spaces with the attitudinal states, but did not find clear separation of any individual attitude according to visual features. We consider that the simple subtraction of the nose movement may be insensitive to any rotation about that point and may not be the best normalization for overall head movement.

5.3. Contribution of the features

We trained a Support Vector Machine (SVM) to better understand the contribution of the individual feature sets. A one vs one classification model was applied on the data collection described in Section 3, and evaluated against the ground truth annotation described in Section 4. The goal of this work is not to develop an automatic classifier, but rather to exploit the classification model in order to explore the feature space. Table 3 shows precision recall and F-scores for each group of audio and video predictors. For ease of comparison, F-score's results are graphed in Fig. 5.

In the case of audio, we find that pitch is, not surprisingly, the strongest correlate of different attitude states, but note that intensity and voice quality have also strong independent contributions. For the video, no individual component shows a particularly strong correlation with attitudinal state, but all features contribute more or less equally. We note that the jaw parameter shows the greatest correlation (probably a better indicator of speech activity) as well as the nose, that represents the overall head movement.

Audio			
Features	Precision	Recall	Fscore
Pitch	0.63	0.57	0.57
Intensity	0.46	0.40	0.41
Voice quality	0.25	0.27	0.22
Audio all	0.71	0.67	0.69
Video			
Features	Precision	Recall	Fscore
Nose	0.31	0.27	0.27
Mouth	0.25	0.26	0.25
Eyes	0.24	0.21	0.21
Eyebrow	0.21	0.21	0.19
Jaw	0.32	0.29	0.30
Video all	0.21	0.23	0.20

Table 3: Prediction performance of different features sets.

From a finer analysis of the confusion matrices, we noted that the 'Amusement' and 'Enthusiasm' classes are not well distinguished by the model, but that 'Friendliness' is detected with high accuracy. Specifically, mouth movements appear to be discriminative in the detection of friendliness.

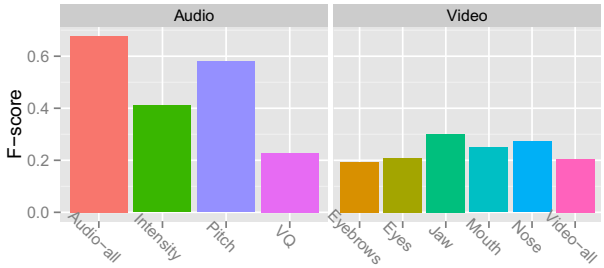

Figure 5: Comparative analysis of the feature sets.

6. Conclusions

In this paper, we described a corpus of video blogs and presented it as a rich source of attitude manifestations. We performed a multimodal feature analysis using audio and visual signals and studied their influence on the various attitudinal states. We compared the features within audio and visual subsets using both Principal Component Analysis and statistical modeling with SVMs to gain an understanding of the individual feature contributions. The analysis of the principal components highlighted the contribution of voice quality among the audio features, and of the eyebrow movements among the visual features. We found that audio features were good predictors of attitudinal states and the overall combined audio prediction was better than that of any individual feature alone. However, for visual features the overall prediction power is still low and we believe that this is due to the inadequate detail from normalised measures of absolute movement. In future work we will look at directionality of facial movements and at the shapes of individual facial features.

7. Acknowledgements

This work is supported by the English Language and Literature Department, UPSI, Ministry of Education Malaysia, the Innovation Bursary of Trinity College Dublin, the Speech Communication Lab at TCD, and by the SFI FastNet project 09/IN.1/1263.

8. References

[1] J.-I. Biel, O. Aran, and D. Gatica-Perez, "You are known by how you vlog: Personality impressions and nonverbal behavior in youtube." in *ICWSM*, 2011.

[2] J.-I. Biel and D. Gatica-Perez, "The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers," *Ethnicity*, vol. 16, no. 4.8, pp. 0–7, 2012.

[3] N. Madzlan, J. Han, F. Bonin, and N. Campbell, "Towards automatic recognition of attitudes: Prosodic analysis of video blogs," in *Speech Prosody, Dublin, Ireland*, 2014, pp. 91–94.

[4] J.-I. Biel, D. Gatica-Perez *et al.*, "Voices of vlogging." in *ICWSM*, 2010.

[5] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 53–56.

[6] J. A. de Moraes, A. Rilliard, B. A. de Oliveira Mota, and T. Shochi, "Multimodal perception and production of attitudinal meaning in brazilian portuguese," in *Proc. of Speech Prosody*, 2010.

[7] P. J. Henrichsen and J. Allwood, "Predicting the attitude flow in dialogue based on multi-modal speech cues," *NEALT PROCEEDINGS SERIES*, 2012.

[8] J.-I. Biel and D. Gatica-Perez, "Vlogsense: Conversational behavior and social attention in youtube," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 7, no. 1, p. 33, 2011.

[9] F. Weninger, C. Wagner, M. Wollmer, B. Schuller, and L.-P. Morency, "Speaker trait characterization in web videos: Uniting speech, language, and facial features," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3647–3651.

[10] K. Sjőlander and J. Beskow, *Wavesurfer [Computer program] (Version 1.8.5)*, 2009. [Online]. Available: http://www.speech.kth.se/wavesurfer

[11] E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell, "Laughter and topic transition in multiparty conversation," in *Proceedings of the Fourteenth SigDial Meeting on Discourse and Dialogue. Metz, France*, 2013, pp. 304–308.

[12] F. Bonin, N. Campbell, and C. Vogel, "Laughter and topic changes: Temporal distribution and information flow," in *3rd IEEE Conference on Cognitive Infocommunications*, 2012, pp. 53–58.

[13] F. Bonin, N. Campbell, and C. Vogel,"Time for laughter," *Knowledge-Based Systems*, 2014, in Press.

[14] K. Jokinen, "Gaze and gesture activity in communication," in *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, ser. Lecture Notes in Computer Science, C. Stephanidis, Ed. Springer Berlin / Heidelberg, 2009, vol. 5615, pp. 537–546.

[15] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[16] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, p. 466, 1997.

[17] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15th ICPhS*, 2003, pp. 2417–2420.

[18] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 300–305.