A Machine-learning based Technique to Analyze the Dynamic Information for Visual Perception of Consonants

Wai Chee Yau[1], Dinesh Kant Kumar[1] and Hans Weghorn[2]

[1] School of Electrical and Computer Engineering, RMIT University
GPO Box 2476V, Melbourne, Victoria 3001,Australia

[2] Information Technology, BA-University of Cooperative Education
Stuttgart, Germany

Abstract. This paper proposes a machine-learning based technique to investigate the significance of the dynamic information for visual perception of consonants. The visual speech information can be described using static (facial appearance) or dynamic (movement) features. The aim of this research is to determine the saliency of dynamic information represented by the lower facial movement for visual speech perception. The experimental results indicate that the facial movement is distinguishable for nine English consonants with a success rate of 85% using the proposed approach. The results suggest that time-varying information of visual speech contained in lower facial movements is useful for machine recognition of consonants and may be an essential cue for human perception of visual speech.

1 Introduction

The advancements in computer-based speech recognition models in the past decades have provided new insights into the understanding of human speech perception. Human speech perception is bimodal and consists of the acoustic and visual modality [5]. The bimodal nature of human speech perception is clearly proven by the McGurk effect, which demonstrates that when a person is presented with conflicting visual and audio speech information, the perception of the sound maybe different from both modalities [12]. An example is when a person hears a sound of /ba/ but sees a lip movement of /ga/, the sound /da/ is perceived.

The acoustic domain is characterized by speech sounds whereas the visual component is described using visual speech signals. The visual speech data refers to the movements of the speech articulators such as lips, facial muscles, tongue and teeth. The visual information from a speaker's face is long known to influence the perception and understanding of spoken language by humans with normal hearing [19, 20]. The ability of people with hearing impairment to comprehend speech by looking at the face of the speaker is yet another clear demonstration of the significance of the visual information in speech perception.

Chee Yau W., Kant Kumar D. and Weghorn H. (2007).
A Machine-learning based Technique to Analyze the Dynamic Information for Visual Perception of Consonants.
In *Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science*, pages 119-128
DOI: 10.5220/0002424801190128
Copyright © SciTePress

The visual analysis of speech by computers are useful in improving the understanding of human speechreading skills. The insights gained regarding the nature of visual speech signals may be beneficial for understanding humans' cognitive abilities in speech perception which encompasses modalities with different temporal, spatial and sensing characteristics. Such machine-based analysis might be able to suggest which visual aspects of speech events are significant in classification of utterances [4]. The results of machine-vision analysis of speech are also useful for applications such as automatic speech recognition in noisy environments.

The visual speech information can be dichotomized into the static and the dynamic components. Campbell [3] reports that the static features are important for visual speech perception where observers are able to recognize phonemes from pictures of faces. Some of the machine-based visual speech recognition approaches proposed in the literature [13, 18] are based on static visual speech features extracted from mouth images. Nevertheless, the dynamic information is demonstrated to be important in human visual speech perception by Rosenblum et. al. [17] (through experiments using point-light display where dots are placed on the lips, cheeks, chin , teeth and tongue tip of the speaker). This paper provides a different view point to the time-varying aspect of visual speech perception through machine analysis. This paper investigates the significance of dynamic, visual speech information - lower facial movements for perceiving consonants. This paper uses only the lower facial movement and not the head movement of the speaker because the most prominent visual speech information lies within the lower face region [15]. A video processing technique is adopted to analyze the mouth video and extract the lower facial movement for machine classification of the consonants. The lower facial movements comprise of the movements of the jaw, lips and teeth. The goal of this research is to use a computer-vision based technique to evaluate the significance of dynamic information encoded in the visible facial movements for visual perception of consonants.

2 An Overview of The Proposed Technique

Spoken language consists of successions of sounds produced by the movements of the speech articulators such as tongue, teeth, lips, velum and glottis in altering the shape of the vocal tract. Figure 1 shows the organs in human speech production system [24]. The smallest units of speech sounds are known as phonemes. Phonemes can be categorized into vowels or consonants depending on the relative sonority of the sounds [7]. The articulation of each phoneme is associated with particular movements of the speech articulators. Nonetheless, the movements of certain speech organs such as velum and glottis are not visible from the frontal view of the speaker [6]. The speech articulators' movements that can be modelled using vision-based system are limited to mostly lips, jaw and teeth motions.

This paper focuses on recognition of consonants due to the fact that consonants are easier to 'see' and harder to 'hear' than vowels [8]. The articulations of vowels are produced with an open vocal tract whereas the productions of consonants involve constrictions at certain part of the vocal tract by the speech articulators. Thus, the facial movements involved in pronunciation of consonants are more discernible. The visual

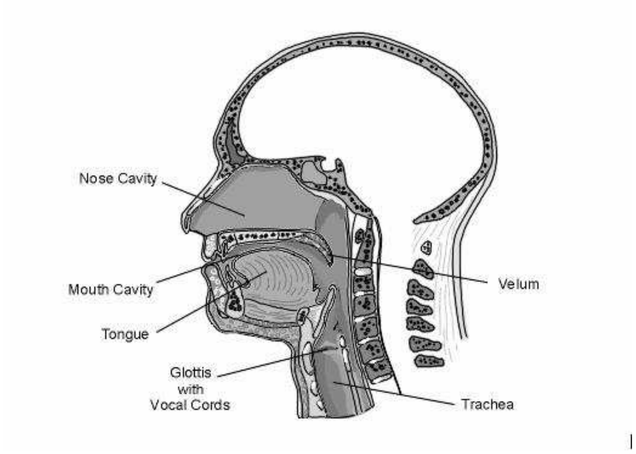

Fig. 1. Diagram of Human Speech Production System.

information is crucial in disambiguating the consonants, especially in conditions where the speech sounds are weak or in noisy environments.

The visible, facial movements associated with the articulations of different speech sounds maybe identical for certain consonants such as /p/ and /b/. Thus, the mapping of speech sounds to facial movements is many-to-one. Table 1 show a mapping of speech sounds to visual movements based on an international audiovisual object-based video representation standard known as MPEG-4.

This paper proposes a machine vision model to analyze the facial movements. The proposed model consist of three stages, which are shown in Figure 2.

2.1 Segmentation of the Lower Facial Movements from Video

This paper proposes to segment the lower facial movements from the video data using a spatio-temporal templates(STT) technique [2]. STT are grayscale images that show where and when facial movements occur in the video. The pixel locations indicate the place where movements occur and the intensity values of the pixels of the STT varies linearly with the recency of the motion. STT are generated using accumulative image difference approach.

Accumulative image difference is applied on the video of the speaker by subtracting the intensity values between successive frames to generate the difference of frames (DOF). DOF of the t^{th} frame is defined as

$$DOF_t(x,y) = |I_t(x,y) - I_{t-1}(x,y)| \quad (1)$$

where $I_t(x,y)$ represents the intensity value of pixel location with coordinate (x, y) of the t^{th} frame. a is the fixed threshold for binarisation of the DOF. $B_t(x,y)$ represents

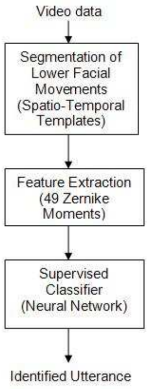
Video data
Segmentation of Lower Facial Movements (Spatio-Temporal Templates)
Feature Extraction (49 Zernike Moments)
Supervised Classifier (Neural Network)
Identified Utterance

Fig. 2. Block diagram of the proposed technique.

the binarised version of the DOF and is given by

$$B_t(x,y) = \begin{cases} 1 \text{ if } DOF_t(x,y) \geq a, \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

The intensity value of the STT at pixel location (x, y) of t^{th} frame is defined by

$$STT_t(x,y) = max \bigcup_{t=1}^{N-1} B_t(x,y) \times t \tag{3}$$

where N is the total number of frames used to capture the lower facial movements. In Eq. (3), the binarised version of the DOF is multiplied with a linear ramp of time to implicitly encode the temporal information of the motion into the STT. By computing the STT values for all the pixels coordinates (x, y) of the image sequence using Eq. (3) will produce a grayscale image (STT) that contains the spatial and temporal information of the facial movements [23]. Figure 3 illustrates the STTs of nine consonants used in the experiments.

This paper proposes the use of STT because STT is able to remove static elements from the sequence of images and preserve the short duration facial movements. STT is also invariant to the skin color of the speakers due to the image subtraction process involved in the generation of STT.

The speed of phonation of the speaker might vary for each pronunciation of a phone. The variation in the speed of utterance results in the variation of the overall duration and there maybe variation in the micro phases of the utterances. The modelling of the details of such variations is very challenging. This paper suggests a model to approximate such variations by normalizing the overall duration of the utterance. This is achieved by normalizing the intensity values of the STT to in between 0 and 1 to minimize the differences in STTs produced from different video recordings of similar phone.

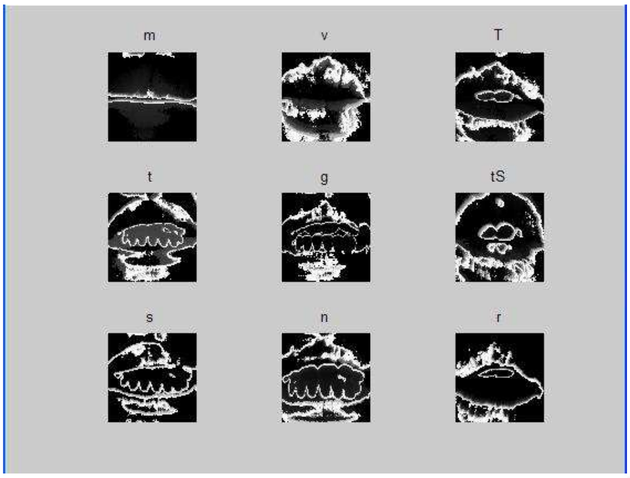

Fig. 3. Spatio-temporal templates of nine consonants that represent the different patterns of facial movements.

The proposed technique uses the discrete stationary wavelet transform (SWT) to reduce the small variations of the facial movements of the same consonant. While the classical discrete wavelet transform (DWT) is suitable for this, DWT results in translation variance [11] where a small shift of the image in the space domain will yield very different wavelet coefficients. The translation sensitivity of DWT is caused by the aliasing effect that occurs due to the downsampling of the image along rows and columns [16]. SWT restores the translation invariance of the signal by omitting the downsampling process of DWT, and results in redundancies.

2.2 Feature Extraction

This paper adopts Zernike moments as the rotation invariant features to represent the SWT approximation of the STT. Zernike moments have been demonstrated to outperformed other image moments such as geometric moments and Legendre moments in terms of sensitivity to noise, information redundancy and capability for image representation [21]. Zernike moments are computed by projecting the image function onto the orthogonal Zernike polynomial defined within a unit circle. The main advantage of Zernike moments is the simple rotational property of the features. Rotational changes of the speaker's mouth in the image results in a phase shift on the Zernike moments [22]. The absolute value of the Zernike moments are invariant to rotational changes [9, 23]. This paper proposes to use the absolute value of the Zernike moments as rotation invariant features to represent the SWT approximate image of STT.

2.3 Supervised Classifier - Artificial Neural Network

A number of possible classifiers maybe suitable for such a machine speech recognition model. The selection of the appropriate classifier would require statistical analysis of the data that would also identify the features that are irrelevant. Supervised artificial neural network (ANN) approach lends itself for identifying the separability of data even when the statistical properties and the types of separability (linear or nonlinear) is not known. While it may be suboptimum, it is an easy tool to implement as a first step.

This paper presents the use of ANN to classify the features into one of the consonants. ANN has been selected because it can solve complicated problems where the description for the data is not easy to compute. The other advantage of the use of ANN is its fault tolerance and high computation rate due to the massive parallelism of its structure [10]. A feedforward multilayer perceptron (MLP) ANN classifier with back propagation (BP) learning algorithm is used in the proposed approach. MLP ANN was selected due to its ability to work with complex data compared with a single layer network. Due to the multilayer construction, such a network can be used to approximate any continuous functional mapping [1]. The advantage of using BP learning algorithm is that the inputs are augmented with hidden context units to give feedback to the hidden layer and extract features of the data from the training events.

3 Methodology

Experiments were conducted to test the repeatability of facial movement features during articulations of consonants. The experiments were approved by the Human Experiments Ethics Committee. Nine consonants highlighted in bold font in Table 1 were used in the experiments. Each consonant represents one pattern of facial movement. The speaker pronounces each consonant in isolation.

Table 1. Visual model of English consonants based on the MPEG-4 standard.

Cluster Number	Phonemes
1	/p/,/b/,**/m/**
2	/f/,**/v/**
3	**/T/**,/D/
4	**/t/**,/d/
5	/k/,**/g/**
6	/S/, /dZ/, **/tS/**
7	**/s/**,/z/
8	**/n/**,/l/
9	**/r/**

The video data used in the experiments was recorded from a speaker using a camera that focused on the mouth region of the speaker. The camera was kept stationary throughout the experiments. The window size and view angle of the camera, background and illumination were kept constant during the recording. The video data was stored as true color (.AVI) files with a frame rate of 30 frames per second. 180 video clips were recorded and one STT was generated from each AVI file. Examples of the STT are shown in Figure 3.

SWT at level-1 using Haar wavelet was applied on the STTs and the approximate images was used for analysis. 49 Zernike moments were used as features to represent the SWT approximate image of the STTs. For further data analysis, k-means clustering algorithm was applied to the moments feature to partition the feature space into nine exclusive clusters using squared Euclidean distance. Figure 4 shows the silhouette plot of the nine clusters representing the nine consonants.

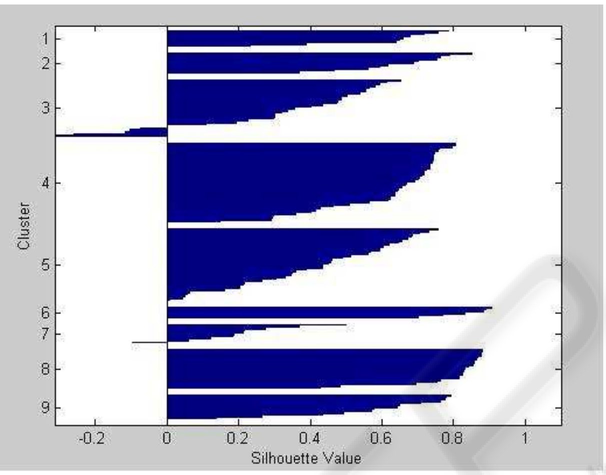

Fig. 4. Silhouette plot of the nine clusters generated using K-means algorithm.

The next step of the experiments was to classify the facial movement features using artificial neural network (ANN). The facial movement features were fed to ANN to classify the features into one of the consonants. Multilayer perceptron (MLP) ANN with backpropagation (BP) learning algorithm was used in the experiments. The architecture of the ANN consisted of two hidden layers. In the experiments, features of 90 STTs were used to train the ANN. The remaining 90 STTs that were not used in training were presented to the ANN to test the ability of the trained ANN to recognize the facial movement patterns. The experiments were repeated 10 times with different set of testing and training data through random sub sampling of the data. The mean and variance of the recognition rates for the 10 repetitions of the experiment were computed.

4 Results and Discussions

The accuracies of the neural network in recognizing the facial movement features of the nine consonants are tabulated in Table 2. The mean classification rate of the experiments is 84.7% with a standard deviation of 2.8%.

The results demonstrate that the patterns of facial movements during articulation of English consonants are highly consistent. 100% success rate is achieved using the visual system to identify the consonant /m/ due to the distinct bilabial movements while pronouncing /m/. The results suggest that facial movements can be useful as dynamic cues for machine recognition of utterances.

Table 2. Mean Classification Accuracies for nine English Consonants.

Viseme	Recognition Rate
/m/	100%
/v/	87%
/T/	65%
/t/	74%
/g/	85%
/tS/	91%
/s/	93%
/n/	74%
/r/	93%

Figure 4 shows that clusters 1, 2, 3, 6, 7 and 9 formed through k-means cluster analysis contain low or negative silhouette values. The low or negative silhouette values indicate that the facial movement features are not distinctly grouped in one cluster, or are assigned to the wrong clusters. The poor clustering results suggest that the features might not be linearly separable. Based on the preliminary data analysis using clustering algorithm, this paper proposes to use a nonlinear classifier - the multilayer perceptron (MLP) artificial neural network (ANN) to classify the facial movement features. The satisfactory classification accuracies of the ANN demonstrate the ability of the ANN to adapt and learn the patterns of the facial movements and achieve non-linear separation of features.

The classification errors can be attributed to the inability of vision-based techniques to capture the occluded movements of speech articulators such as glottis, velum and tongue. For example, the tongue movement in the mouth cavity is either partially or completely not visible (occluded by the teeth) in the video data during the pronunciation of alveolar and dental sounds such as /t/, /n/ and /T/. The STTs of /t/, /n/ and /T/ do not contain the information of the occluded tongue movements. This is a possible reason for the higher error rates of 26% and 35% for these three consonants as compare to the average error rate of 15% for all consonants. The results suggest that the facial movements of /t/, /n/ and /T/ are less distinguishable compared with other consonants.

The human perceptual analysis on visual speech using point-light displays reported in [17] clearly indicates that the dynamic component of visual speech may be the most salient informational form (versus the static face information). Our experimental results using computer-based analysis present an evidence to support the significance of time-varying information for visual perception of consonants. Our results indicate that the dynamic information in the lower face region of the speaker is useful in perceiving consonants. Nevertheless, the authors would like to point out that the proposed technique has only been tested using discrete consonants. Speech sounds are often perceived in context and not in isolation by humans. The future direction of this research is to exam-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

ine the feasibility of using dynamic visual speech information to identify speech sounds that are embedded in words.

5 Conclusion

This paper analyzes the significance of dynamic information of the lower facial movements in the visual perception of consonants using a machine-learning based technique. The proposed visual technique has been used to validate the results of perceptual analysis that shows that time-varying information is important in human perception of visual speech [17]. The outcome of the analysis using the proposed machine-learning technique indicate that dynamic speech information contained in the lower facial movements are useful in disambiguating consonants thereby supporting the findings of the study reported in [17].

The experimental results indicate that different patterns of facial movements can be used to differentiate nine consonants with accuracies of 84.7%. These results demonstrate that facial movements are reliable in representing consonants and can be useful in machine speech recognition. The proposed machine analysis provides better understanding of the cognitive process involved in human speech perception by validating the saliency of the dynamic visual speech information.

For future work, the authors intend to evaluate the reliability of facial movements in other commonly spoken languages such as German and Mandarin. Also, the authors intend to test on a larger vocabulary set covering words and phrases. Potential applications for the proposed technique include automated systems such as interfaces for users with speech impairment to control computers and control of heavy machineries in noisy factory.

References

1. Bishop, C. M.: Neural Networks for Pattern Recognition. Oxford University Press (1995)
2. Bobick, A. F., Davis, J. W.: The Recognition of Human Movement Using Temporal Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23 (2001) 257–267
3. Campbell, R.: The lateralisation of lipread sounds:A first look. Brain and Cognition. Vol. 5, (1986) 1–21
4. Campbell, R., Dodd, B., Burnham, D.:Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech.(1998) X–XIV Vol. 91 (2003)
5. Chen, T.: Audiovisual Speech Processing. IEEE Signal Processing Magazine, Vol. 18. (2001) 9–21
6. Hazen, T. J.: Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition. IEEE Transactions on Audio, Speech and Language Processing (2006) Vol. 14 No. 3 1082–1089
7. Jones, D.:An Outline of English Phonetics,W Jeffer and Sons Ltd(1969) 23
8. Kaplan, H., Bally, S. J., Garretson, C.:Speechreading: A Way to Improve Understanding.Gallaudet University Press,(1999)14–16
9. Khontazad, A., Hong , Y. H.: Invariant Image Recognition by Zernike Moments. IEEE Transactions on Pattern Analysis and Machine Intelligence (1990) Vol. 12 489–497

10. Kulkarni, A. D.: Artificial Neural Network for Image Understanding. Van Nostrand Reinhold (1994)
11. Mallat, S.:A Wavelet Tour of Signal Processing. Academic Press (1998)
12. McGurk, H., MacDonald, J.: Hearing Lips and Seeing Voices. Nature,Vol. 264 (1976)746–748
13. Petajan, E. D.: Automatic Lip-reading to Enhance Speech Recognition. In GLOBE-COM'84,IEEE Global Telecommunication Conference (2004)
14. Potamianos, G., Neti, C., Gravier, G., Senior, A.W.: Recent Advances in Automatic Recognition of Audio-Visual Speech. In Proc. of IEEE, Vol. 91 (2003)
15. Potamianos, G., Neti, C.: Improved ROI and Within Frame Discriminant Features For Lipreading. In Proc. of Internation Conference on Image Processing, (2001) 250–253
16. Simoncelli, E. P., Freeman, W. T., Adelson, E. H., Heeger, D. J.:Shiftable Multiscale Transform. IEEE Transactions on Information Theory (1992) Vol. 38 587–607
17. Rosenblum, L. D., Saldaa, H. M. : Time-varying information for visual speech perception, in Hearing by Eye: Part 2, The Psychology of Speechreading and Audiovisual Speech, R. Campbell, B. Dodd, and D. Burnham, Editors. Earlbaum: Hillsdale, NJ (1998)61–81
18. Stork, D. G., Hennecke, M. E.: Speechreading: An Overview of Image Processing, Feature Extraction, Sensory Intergration and Pattern Recognition Techiques.In the 2nd International Conference on Automatic Face and Gesture Recognition (FG '96), (1996)
19. Summerfield, A. Q.: Some preliminaries to a comprehensive account of audio-visual speech perception. Hearing by Eye : The Psychology of Lipreading (1987)
20. Sumby, W. H., Pollack, I.: Visual contributions to speech intelligibility in noise. Journal of the Acoustical Society of America, Vol. 26 (1954) 212–215
21. Teh, C. H., Chin, R. T.: On Image Analysis by the Methods of Moments. IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 10. (1988)496–513
22. Teague, M. R.: Image Analysis via the General Theory of Moments. Journal of the Optical Society of America (1980) Vol. 70 920–930
23. Yau, W. C., Kumar, D. K., Arjunan, S. P. : Visual Speech Recognition Method Using Translation, Scale and Rotation Invariant Features. IEEE International Conference on Advanced Video and Signal based Surveillance, Sydney, Australia (2006)
24. http://www.kt.tu-cottbus.de/speech-analysis/tech.html

SCITEPRESS Science and Technology Publications