Articulatory Features for Robust Visual Speech Recognition

Kate Saenko, Trevor Darrell, and James Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street
Cambridge, Massachusetts, USA

{saenko,trevor,glass}@mit.edu

ABSTRACT

Visual information has been shown to improve the performance of speech recognition systems in noisy acoustic environments. However, most audio-visual speech recognizers rely on a clean visual signal. In this paper, we explore a novel approach to visual speech modeling, based on articulatory features, which has potential benefits under visually challenging conditions. The idea is to use a set of parallel SVM classifiers to extract different articulatory attributes from the input images, and then combine their decisions to obtain higher-level units, such as visemes or words. We evaluate our approach in a preliminary experiment on a small audio-visual database, using several image noise conditions, and compare it to the standard viseme-based modeling approach.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]

General Terms

Algorithms, Design, Experimentation.

Keywords

Multimodal interfaces, audio-visual speech recognition, speechreading, visual feature extraction, articulatory features, support vector machines.

1. INTRODUCTION

A major weakness of current automatic speech recognition (ASR) systems is their sensitivity to environmental and channel noise. A number of ways of dealing with this problem have been investigated, such as special audio preprocessing techniques and noise adaptation algorithms [2]. One possible approach is to take advantage of all available sources of linguistic information, including nonacoustic sensors [24], to provide greater redundancy in the presence of noise. In particular, the visual channel, while clearly not affected by audio noise, conveys complementary linguistic information. Using the images of the speaker's mouth to recognize speech is commonly known as *lipreading*. Long known to improve human speech perception [30], lipreading has been applied to ASR extensively over the past twenty years. The result is the emergence of two closely related fields of research. The first, *Visual Speech Recognition,* sometimes also referred to as *automatic lipreading* or *speechreading,* uses just the visual input to recognize speech. The second, *Audio-Visual Speech Recognition (AVSR)*, combines both modalities to improve traditional audio-only ASR. Current AVSR systems are able to achieve an effective SNR gain of around 10 DB over traditional audio-based systems [28]. Overall, automatic lipreading promises to add robustness to human–machine speech interfaces.

In practice, however, the visual modality has yet to become mainstream in spoken human–computer interfaces. This is due partially to the increased processing and storage demands, and also to the relative novelty of the field. In particular, the lack of large, commonly available audio-visual corpora has hindered the development of practical algorithms. Furthermore, the reliance of current systems on high-quality video recorded in controlled environments, where the speaker always faces the camera, is a major issue in practice. In fact, in situations where acoustic channel noise is a problem, it is possible that the visual channel will also become corrupted by noise, for example, due to inferior quality of recording equipment.

The need for improving the robustness of visual feature extraction algorithms is starting to attract attention in the research community. A recent study compared the performance of a state-of-the-art AVSR system on a typical "visually clean" studio database and a more realistic database recorded in offices and cars using an inexpensive web camera [27]. The results show that, although the visual modality remains beneficial even in such challenging conditions, the visual-only word error rate (WER) approximately doubles when moving from the studio to the office data, and triples on the automobile data.

We propose a novel approach to visual speech modeling, and show that it can lead to improved recognition rates in the presence of image noise. Our method is based on representing speech classes in terms of the underlying articulatory processes. The concept of *articulatory features (AFs)* is not new in the speech community (e.g. [21], [31],) but, to the best of our knowledge, it has never been applied in the visual domain. AF-based modeling has been used successfully in audio ASR to improve its robustness in adverse acoustic environments [16]. Our hypothesis is that the benefits of feature-based recognition would also apply in the case of visual speech.

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

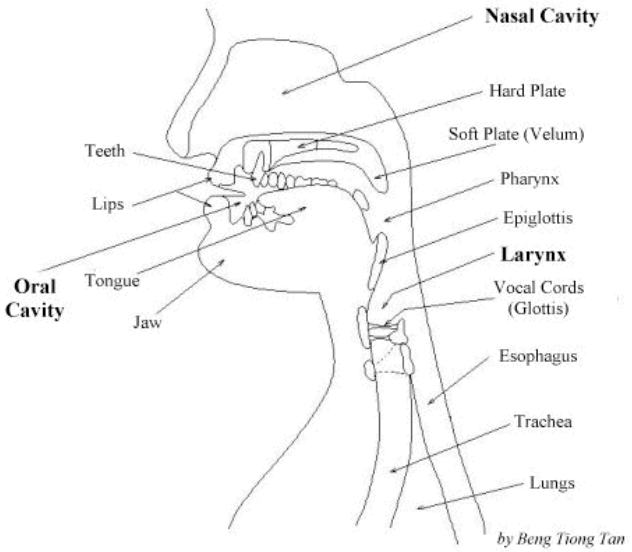

Figure 1. Human Speech Production

Section 2 summarizes previous work conducted in the field of audio-visual speech recognition and provides a background review of articulatory speech modeling in the audio domain. Section 3 describes our approach, and Section 4 presents our initial experiments and results. Section 5 discusses possible future work directions.

2. BACKGROUND

The first audio-visual speech recognizer was designed by Petajan in 1984 [26]. Since then, over one hundred research articles have been published on the subject. Applications have ranged from single-subject, isolated digit recognition [26], to speaker-independent, large-vocabulary, continuous speech recognition [23]. The majority of reported AVSR systems have achieved superior performance over conventional ASR, although the gains are usually more substantial for small vocabulary tasks and low acoustic signal-to-noise ratios [28].

The main issues involved in the development of AVSR systems are 1) *visual feature design and extraction*, 2) *the choice of speech units,* 3) *classification*, and 4) *audio-visual integration*. Although the second and third issues also apply to audio-only systems and are therefore often resolved in the same way for both modalities, the first and the last issues are unique to audio-visual systems.

2.1 Visual Feature Extraction

Visual feature design falls into three main categories: *appearance*-based, *shape*-based, and a combination of the two. Appearance-based approaches treat all intensity and color information in a *region of interest* (usually the mouth and chin area) as being relevant for recognition. The dimensionality of the raw feature vector is often reduced using a linear transform. Some examples of this "bottom-up" approach include simple gray levels [12]; principal component analysis of pixel intensities [3]; motion between successive frames [19]; transform-based compression coefficients [29]; edges [1]; and filters such as sieves [20].

In contrast, shape-based methods usually assume a "top-down" model of lip contours. The parameters of the model fitted to the image are used as visual features. Some examples of shape-based features include geometric features, such as mouth height and width [26], [1], [4]; Fourier and image moment descriptors of the lip contours [13]; snakes [14]; and *Active Shape Models (ASM)* [9]. In general, lip contours alone lack the necessary discriminative power, so they are often combined with appearance. For example, it was shown that the addition of appearance to shape significantly improves the lipreading performance of the ASM [20]. The result is an *Active Appearance Model (AAM)* [8], which combines shape and appearance parameters into a single feature vector.

2.2 Speech Unit Modeling

Traditionally, speech is assumed to consist of a sequence of contiguous basic units, or *phonemes*. This view is consistent with the early theory of *generative phonology* [6]. The English language has about 50 phonetic units (see Table 1.) In the case of visual speech, the basic units correspond to the visually distinguishable phonemes, also known as *visemes*. There are fewer visemic than phonetic units, since some phonemes, e.g. /d/ and /t/, cannot be distinguished visually. A sample viseme-to-phoneme mapping is shown in Table 1.

In recent years, a competing theory of *nonlinear phonology* has been attracting attention in the ASR community. Based on knowledge of human speech production, it views speech as the combination of multiple streams of hidden *articulatory features* [15]. A diagram describing the human speech production mechanism is shown in Figure 1. The vocal tract – the main speech organ – consists of the pharynx, the nasal cavity and the oral cavity. The glottis, the soft plate (velum), the tongue, the lips and the jaw are the articulators. The process of changing the shape of the vocal tract to produce different sounds is called articulation [11]. Thus, from the point of view of articulation, each phoneme can be defined in terms of several features, for example, *voicing, tongue body position, tongue tip position, frication*, etc.

Table 1. A sample mapping of 52 phonemes to 14 visemes

Viseme Index	Corresponding Phonemes
1	ax ih iy dx
2	ah aa
3	ae eh ay ey hh
4	aw uh uw ow ao w oy
5	el l
6	er axr r
7	y
8	b p
9	bcl pcl m em
10	s z epi tcl dcl n en
11	ch jh sh zh
12	t d th dh g k
13	f v
14	gcl kcl ng

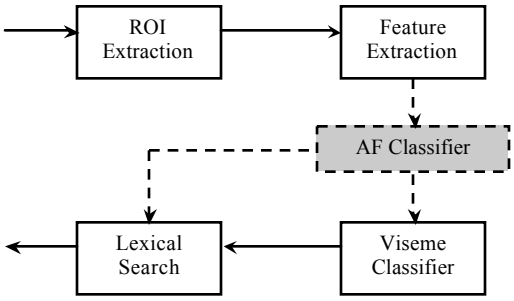

Figure 2. Articulatory-feature approach to visual speech recognition.

One of the advantages of representing speech as multiple streams of articulatory features is the ability to model each feature independently and even to allow them to de-synchronize. For example, it has been noted that spontaneous, conversational speech is difficult to transcribe in terms of conventional phoneme units, and presents a challenge for existing ASR systems. On the other hand, feature-based pronunciation models have been shown to better account for the types of pronunciation variations that occur in spontaneous speech [18].

Another advantage of AF-based modeling is its robustness in noisy environments. Experiments in acoustic speech recognition have shown that articulatory-feature systems can achieve superior performance at high noise levels [16].

2.3 Classification

Visual speech recognizers differ in their choice of classification techniques. Due to the dynamic nature of speech, the most common classifier used is a Hidden Markov Model (HMM), which allows statistical modeling of both the temporal transitions between speech classes, and the generation of class-dependent visual observations [23]. Although most HMMs use a Gaussian Mixture Model classifier for the latter task, several other classification methods have been suggested, including simple distance in feature space [26], neural networks [17] and Support Vector Machines (SVMs) [12]. In this work, we employ SVM classifiers, which are capable of learning the optimal separating hyperplane between classes in sparse high-dimensional spaces and with relatively few training examples. More details on the SVM algorithm can be found in [33].

2.4 Audio-Visual Integration

In the case of audio-visual speech recognition, a major area of ongoing research is the integration of the two modalities in such a way that the resulting recognizer outperforms both the visual-only and audio-only recognizers. Integration algorithms generally fit into one of two broad categories: *feature fusion* and *decision fusion*, sometimes also referred to as *early integration* and *late integration*. Feature fusion involves training a single classifier on the fused bimodal data vectors [32], whereas decision fusion involves training separate single-modality classifiers and then combining their outputs, for instance, as a weighted sum [10]. Decision fusion can occur at any level (e.g., HMM state, phoneme, word, or sentence,) although very early stage fusion techniques are commonly referred to as *hybrid fusion* [28], [7].

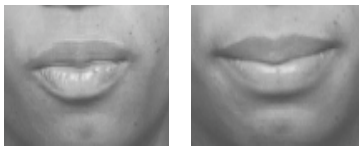

Figure 3. Full bilabial closure during the production of the words "romantic" (left) and "academic" (right)

Although we do not directly address the issue of audio-visual integration in this article, the proposed articulatory-feature model could be extended to include both acoustic and visual observations.

3. ARTICULATORY FEATURES

We propose a novel approach to modeling visual speech, derived from human speech production and inspired in part by the articulatory-feature models described in the previous section. Since we are dealing only with the visual modality, we are limited to the modeling of *visible* articulators. Given the video of the speaker's lower face region, we can obtain information about the position and relative configuration of the jaw, lips, teeth, and tongue. However, the rest of the articulators are not visible under normal circumstances. Also, in addition to static features, the video contains dynamic articulatory features, for example, lips closing and opening, tongue protruding and retracting through teeth, lower lip touching upper teeth, lips protruding, and so on.

Our approach is in many ways identical to the multistream articulatory-feature approach to audio speech modeling. We are essentially proposing to model visual speech as multiple streams of visible linguistic features, as opposed to a single stream of visemes. In fact, most of the articulatory events described above have direct equivalents in the feature set used for pronunciation modeling in [18]. For example, the visual feature of the lips closing and opening corresponds to the LIP-OPEN feature. Therefore, an integrated AF-based audio-visual speech recognizer can use the same underlying feature set. However, due to the complementary nature of the two modalities, some features may be easier to derive from the audio stream, and others from the video stream, especially in the presence of noise. For instance, it is known from perceptual studies that acoustic noise affects the detection of place of articulation (e.g. glottal, bilabial) more than voicing [22]. On the other hand, since place information is highly distinguishable visually, it might be less affected by visual noise than other features.

A typical visual speech recognition system consisting of four stages is illustrated in Figure 2. The stages are: 1) face detection and region of interest (ROI) tracking, 2) low-level image processing and feature vector extraction, 3) per-frame categorization into viseme classes, and 4) the incorporation of frame-level scores over time in order to find the most likely word sequence. Our approach introduces an extra step after the initial preprocessing of the image, but before the viseme scores are computed. In this step, the input data are classified in terms of several articulatory features by a set of parallel statistical classifiers. Afterwards, the lexical search can either proceed right away, using the obtained articulatory feature scores, or follow an additional step of classification into the higher-level visemic categories.

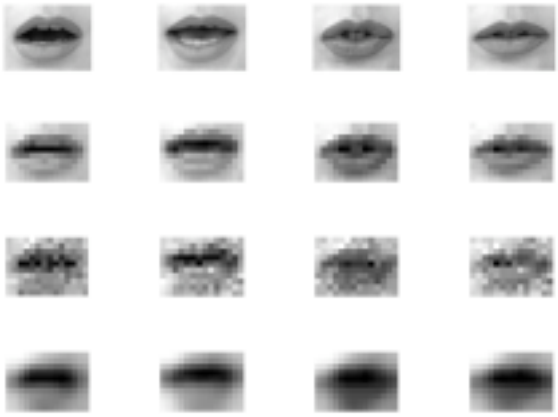

Figure 4. Sample viseme images for Speaker 1, from left to right: /ao/, /ae/, /uw/ and /dcl/. From top to bottom: the original high- resolution images, resized clean images used for training, with added 50% pixel noise, and blurred with Gaussian kernel of size 10.

The difference between our method of classifying articulatory features and the conventional method of classifying visemes is illustrated by the following example. Suppose we were to model the phoneme /m/ in two different phonetic contexts, *romantic* and *academic*. The image snapshot taken at the moment of complete closure during the production of /m/ in each context is shown in Figure 3. Both examples would be considered to belong to a single viseme class (the bilabial viseme) and to have the same open/closed feature value (fully closed.) However, their appearance is different: in the second context, the distance between the mouth corners is roughly 25% wider. This suggests the presence of contextual information. In fact, the preceding /ow/ in *romantic* causes the /m/ to be rounded, whereas the preceding /eh/ in *academic* does not. Thus, modeling lip rounding and lip opening as two separate articulatory features would allow us to recover more information than just modeling the /m/ viseme. An alternative would be to use longer units, e.g. bi-visemes or tri-visemes, however, this would lead to a decrease in the amount of training data available per class and an increase in the number of model parameters.

It is important to note that the proposed method of extracting articulatory feature information using statistical classifiers differs from extracting geometric parameters (for example, the width and height of the mouth opening) from the visual input data [25]. The latter task involves segmenting the image or fitting a lip contour model, and relies mainly on image processing algorithms. In contrast, our model can use the same preprocessing techniques as the regular viseme classifier normally would. The difference is that the feature classifier assigns abstract class labels to the data samples that correspond to various articulatory attributes, such as *rounded*, *fricated*, etc. Note, however, one of the potential benefits of our approach is the ability to use different low-level measurements for each articulatory feature. For example, the classifier for *rounded* could take optical flow measurements as input, while the *teeth* classifier could use color information.

Table 2. Viseme to Feature Mapping

Viseme	LIP-OPEN	LIP-ROUND
/ao/	Wide	Yes
/ae/	Wide	No
/uw/	Narrow	Yes
/dcl/	Narrow	No

Because of its decompositional nature, the articulatory-feature approach has several potential benefits to visual speech modeling. First of all, it combines several sources of information about the underlying speech process, derived independently via parallel classifiers. Therefore, it can take advantage of the fact that some of the features may become harder to classify than others under conditions of image noise, low resolution, or speaker differences. Confidence values on each feature can be used to assign them different weights, effectively reducing the overall number of distinguishable classes. Furthermore, as there are fewer possible values for each feature class than there are visemes, the training dataset generates more instances of each feature class value than each viseme, leading to a larger amount of training data.

4. PRELIMINARY EXPERIMENTS

The experiments described in this section investigate the performance of an AF-based classifier on visually noisy data, where the train and test noise conditions are mismatched. As this is still very much a work in progress, we have only limited initial experiments to report. Nevertheless, they indicate that our approach increases the viseme classification rate on a simple task and therefore merits further investigation.

4.1 Data Collection and Processing

We conducted our initial proof-of-concept experiments on a small two-speaker audio-visual speech corpus previously collected in our lab. The corpus consists of continuous repetitions of a nonsense utterance designed to provide a balanced coverage of English visemes. In order to facilitate the accurate extraction and tracking of the mouth region, the first speaker's lips were colored blue. A color histogram model was then used to segment the lip region of interest. The second speaker's lips were not colored, but rather segmented using correlation tracking, which resulted in imperfect ROI localization. Viseme labels were determined from an audio transcription, obtained automatically using an audio speech recognizer, via the mapping described in Table 1. Figure 4 shows some sample Speaker 1 viseme images taken from the center of the corresponding phonetic segments.

Prior to classification, the original 120x160 sample image was scaled down to 10x14 pixels in size and then vectorized to form a 140-element data vector. The decision to use very simple image features (pixels) as input to the SVM was intentional. When applied to other pattern recognition tasks, SVMs have achieved very good results using only such simple input features. Furthermore, we wanted to allow the discriminative power of the SVM determine those parts of the image that are key to a particular feature without making any prior assumptions.

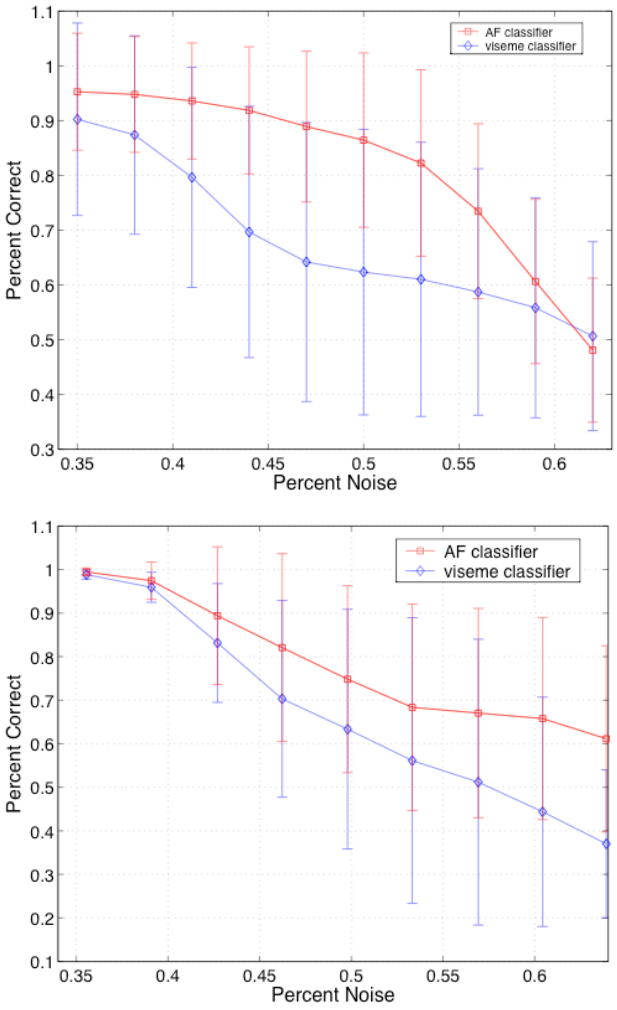

Figure 5. Comparison of viseme classification rates obtained by the AF-based and viseme-based classifiers on test data with added random pixel noise, for Speaker 1 (top) and Speaker 2 (bottom.)

We used a training set consisting of 200 samples per viseme, and a separate "visually clean" test set of 100 samples per viseme. The "visually noisy" test sets we created by either adding random Gaussian pixel noise to the down-sampled test images, or blurring the original images with a Gaussian filter to reduce their effective resolution.

4.2 Viseme Classification Using Features

As a start, we applied our approach to the task of viseme classification. For this experiment, we used only four visemes, corresponding to the phonemes /ao/, /ae/, /uw/ and /dcl/. We chose the viseme set so that it could be completely encoded by the cross product of two binary articulatory features, in this case, LIP-OPEN and LIP-ROUND. Table 2 shows the mapping from the visemes to the articulatory feature values. In the general case, there would be on the order of a few dozen visemes, and so the number of articulatory features would necessarily increase. Note that we could have used more features, such as the visibility of teeth or the tongue position, making the feature set redundant.

Table 4. Classification Rate for Low-Resolution Data

Kernel Size	Viseme	OPEN	ROUND	Combined
None	99	100	99	99
9	97	100	99	99
10	90	99	99	98

A separate SVM classifier was trained for the four visemes, as well as for each of the two features, using LIBSVM software [5], which implements the "one-against-one" multi-class method. We chose to use the radial basis function (RBF) kernel in all experiments, as we found it to give the best performance with the fewest free parameters. The RBF kernel is defined as follows:

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0,$$

where x_i, x_j are training samples. Therefore, in addition to the penalty parameter of the error term, C, the RBF-based SVM has another free parameter γ. To find the optimal values for these two parameters, we performed a grid search over a range of parameter values, using randomized v-fold cross-validation on the training data.

During classification, feature labels were converted to viseme labels using the mapping shown in Table 2. This is the simplest possible combination rule. Another alternative would have been to train a second-level viseme classifier that takes the concatenated probabilities of the two features obtained from the two first-level classifiers as input.

4.3 Results

Figure 5 shows the classification rates obtained by each classifier across several levels of random pixel noise, averaged over 20 training and testing runs. The horizontal axis shows the percentage of Gaussian noise that was added to the test images. The vertical axis shows the correct viseme classification rate. Results for each speaker are shown on separate plots. Table 4 shows the classification results on the low-resolution test data for Speaker 1. The first column shows the size of the Gaussian kernel used to blur the original high-resolution images. The second column shows the viseme classification rate obtained by the viseme classifier, and the next two columns show the respective LIP-OPEN and LIP-ROUND feature classification rates. The last column shows the viseme classification rate obtained by combining the results of the individual feature classifiers. One interesting fact is the resilience of the SVM to significant amounts of noise and blurring. This could be attributed to the fact that the four chosen visemes can be distinguished using mostly low-frequency information. The same result may not hold for other visemes that can only be distinguished by high-frequency information, such as a small opening between the lips, etc.

5. CONCLUSION AND FUTURE WORK

Overall, the results of our preliminary experiments show the advantage of using articulatory feature modeling for viseme recognition from noisy images. While the viseme classifier's performance degrades with increasing noise levels, the combined articulatory feature-based classifier retains a significantly higher recognition rate.

As this research is still in its early stages, there are many interesting open issues to pursue in the future. We have already started conducting experiments on a larger, multi-speaker database. Since we used the SVM classifier for our experiments, we would like to explore whether other classifiers benefit from the articulatory feature modeling approach as well. In addition, we plan to extend the feature set to cover the set of all possible visemes. Another direction for future work is incorporating articulatory features into the framework of word recognition. It has been noted that using articulatory features overlapping in time leads to advantages in context modeling over traditional multi-phone models [31]. Since the feature spreading property is particularly noticeable in the lip features, it would be interesting to apply this approach to context modeling in visual speech. Finally, the merits of the feature approach in an integrated audio-visual speech recognizer should be explored.

6. ACKNOWLEDGEMENTS

This research was supported by ITRI and by DARPA under SRI sub-contract No. 03-000215.

7. REFERENCES

[1] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in HMM-based ASR," in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, pp. 461–471, 1996.

[2] S. Boll, "Speech enhancement in the 1980s: noise suppression with pattern matching," In Advances in Speech Signal Processing, pp. 309-325, Dekker, 1992.

[3] C. Bregler and Y. Konig, "Eigenlips for Robust Speech Recognition," In Proc. ICASSP, 1994.

[4] M. Chan, Y. Zhang, and T. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in Proc. Works. Multimedia Signal Processing, pp. 65–70, Redondo Beach, CA, 1998.

[5] C. Chang and C. Lin, LIBSVM: A Library For Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] N. Chomsky and M. Halle, The Sound Pattern of English, Harper and Row, New York, 1968.

[7] S. Chu and T. Huang, "Bimodal speech recognition using coupled hidden Markov models," In Proc. Int. Conf. Spoken Lang. Processing, vol. II, Beijing, China, pp. 747–750, 2000.

[8] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," In Proc. Europ. Conf. Computer Vision, Germany, pp. 484–498, 1998.

[9] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and application," Computer Vision Image Understanding, vol. 61, no. 1, pp. 38–59, 1995.

[10] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Trans. Multimedia, vol. 2, no. 3, pp. 141–151, 2000.

[11] G. Fant, Acoustic Theory of Speech Production, Netherlands: Mouton and Co., 1960.

[12] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine based dynamic network for visual speech recognition applications," EURASIP J. Appl. Signal Processing, vol. 2002, no. 11, pp. 1248–1259, 2002.

[13] S. Gurbuz, Z. Tufekci, E. Patterson, and J. Gowdy, "Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition," in Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 177–180, Salt Lake City, UT, 2001.

[14] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," Int. J. Computer Vision, vol. 1, no. 4, pp. 321–331, 1988.

[15] S. King, T. Stephenson, S. Isard, P. Taylor and A. Strachan, "Speech recognition via phonetically featured syllables," In Proc. ICSLP, Sydney, 1998.

[16] K. Kirchhoff, G. Fink and G. Sagerer, "Combining Acoustic and Articulatory-feature Information for Robust Speech Recognition," In Proc. ICSLP, pp. 891-894, Sydney, 1998.

[17] G. Krone, B. Talle, A. Wichert, and G. Palm, "Neural architectures for sensor fusion in speech recognition," In Proc. Europ. Tut. Works. Audio-Visual Speech Processing, pp. 57–60, Greece, 1997.

[18] K. Livescu and J. Glass, "Feature-based Pronunciation Modeling for Speech Recognition," In Proc. HLT/NAACL, Boston, 2004.

[19] K. Mase and A. Pentland, "Automatic Lipreading by optical flow analysis," Systems and Computers in Japan, vol. 22, no. 6, pp. 67-76, 1991.

[20] I. Matthews, T. Cootes, A. Bangham, S. Cox and R. Harvey, "Extraction of Visual Features for Lipreading," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, 2002.

[21] F. Metze, and A. Waibel, "A Flexible Stream Architecture for ASR Using Articulatory Features," In Proc. ICSLP, Denver, 2002.

[22] G. Miller and P. Nicely, "An Analysis of Perceptual Confusions among some English Consonants," J. Acoustical Society America, vol. 27, no. 2, pp. 338-352, 1955.

[23] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop," In Proc. Works. Signal Processing, pp. 619–624, Cannes, France, 2001.

[24] L. Ng, G. Burnett, J. Holzrichter, and T. Gable, "Denoising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing," In Proc. ICASSP, Istanbul, Turkey, 2000.

[25] P. Niyogi, E. Petajan, and J. Zhong, "Feature Based Representation for Audio-Visual Speech Recognition", Proceedings of the Audio Visual Speech Conference, Santa Cruz, CA, 1999.

[26] E. Petajan, "Automatic lipreading to enhance speech recognition," In Proc. Global Telecomm. Conf., pp. 265–272, Atlanta, GA, 1984.

[27] G. Potamianos and C. Neti, “Audio-visual speech recognition in challenging environments,” In Proc. Eur. Conf. Speech Comm. Tech., pp. 1293-1296, Geneva, 2003.

[28] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, In Proc. IEEE, 2003.

[29] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, “A Cascade Image Transform for Speaker-Independent Automatic Speechreading," In Proc. ICME, volume II, pp. 1097-1100, New York, 2000.

[30] W. Sumby, and I. Pollack, “Visual contribution to speech intelligibility in noise,” J. Acoustical Society America, vol. 26, no. 2, pp. 212–215, 1954.

[31] J. Sun and L. Deng, “An Overlapping-Feature Based Phonological Model Incorporating Linguistic Constraints: Applications to Speech Recognition”, J. Acoustic Society of America, vol. 111, No. 2, pp. 1086-1101, 2002.

[32] P. Teissier, J. Robert-Ribes, and J. Schwartz, “Comparing models for audiovisual fusion in a noisy-vowel recognition task,” IEEE Trans. Speech Audio Processing, vol. 7, no. 6, pp. 629–642, 1999.

[33] V. Vapnik, Statistical Learning Theory, J. Wiley, New York, 1998.