Feature Based Representation for Audio-Visual Speech Recognition

Partha Niyogi, Eric Petajan, Jialin Zhong

Bell Labs – Lucent Technologies
Murray Hill, NJ 07974, USA.

ABSTRACT

In this paper, we consider the interaction of acoustic and visual stimuli at the subphonemic level of the distinctive feature. We argue that this provides a natural intermediate level for audio-visual integration and discuss the visual and acoustic feature detection problems that are associated with this task.

1. RECOGNITION USING AUDIO-VISUAL CUES

While compelling psychophysical evidence exists [9] suggesting that the auditory and visual modalities are strongly linked in word and syllable recognition, it is generally unclear how to embody such audio-visual links in a coherent and insightful computational framework. In speech recognition systems that are based on HMMs, the core recognition engine (an HMM) is constrained to receive inputs that are a sequence of vectors in a finite-dimensional space. As a result, most approaches using such systems end up simply concatenating (at each point in time) the video frame and audio frame into one audio-visual frame and performing recognition with such combined inputs. Alternatively, one uses essentially two HMM systems — one based entirely on video inputs and another entirely on audio inputs and combines likelihood scores from the two in a late integration strategy.

In this paper, we consider the possibility of an *intermediate* level at which the interaction of visual and acoustic cues might occur. We argue that the distinctive feature provides a reasonable intermediate level at which the cues can be integrated prior to lexical access and describe the visual and auditory components of such a feature based strategy for audio-visual recognition.

1.1. Distinctive Features

An alternative framework for speech recognition has been pursued in Niyogi et al [6, 5] that utilize the notion of distinctive features [3]. Such a perspective has its roots in phonological theory that suggests that phonemes are not the atomic units of which syllables, words and other linguistic objects are composed but in fact are themselves decomposable into primitives called distinctive features. Each distinctive feature is conceptually a binary valued variable and shown in table 1 are some phonemes and their distinctive features.

The distinctive features have typically been viewed as phonological oppositions that separate minimal pairs of confusible phonemes in a language. Thus the minimal pairs p,b, t,d etc. are separated by the feature voice with /p/ and /t/ being unvoiced /-voice/ and /b/ and /d/ being voiced /+voice/. The distinctive features may also be viewed as defining a natural phonological class, e.g., labial sounds p,b,m form a class with feature value /+labial/ while the other sounds have feature value /-labial/. Phonemes are thus treated as complexes of such features and representation of the lexicon in terms of such features provides a compact description of various phonological phenomena [4].

Feature	p	b	t	s	m	n	u
Consonantal	+	+	+	+	+	+	-
Labial	+	+	-	-	+	-	-
Alveolar	-	-	+	+	-	+	-
Nasal	-	-	-	-	+	+	-
Voicing	-	+	-	-	+	+	+
Continuant	-	-	-	+	-	-	+

Table 1: Phonemes and their associated distinctive feature values.

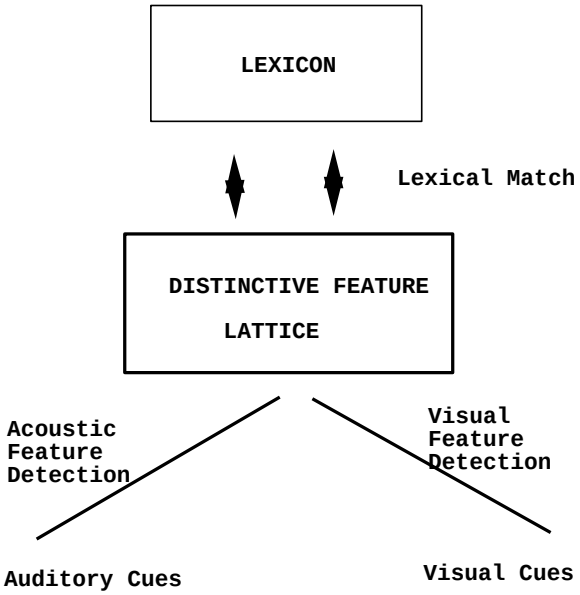

Figure 1: Schematic of Audio-Visual Word Recognition Using Distinctive Features.

It is further instructive to note that the distinctive features have a natural articulatory interpretation and because of this they have (i) acoustic consequences in the speech signal that are presumably picked up by the auditory system during lexical access (ii) direct visual cues of the articulatory gestures that are then picked up by the visual system. Therefore, the level of the distinctive feature might be appropriate for the integration of auditory and visual cues and the feature itself is then used for accessing words in the lexicon for recognition tasks. Fig. 1 characterizes this.

There are two aspects of the distinctive features and their auditory and visual correlates that are worth highlighting:

(i) Each feature has both an acoustic and a visual correlate. Each feature can therefore be extracted from audio-visual input. For example, the feature [labial] that can be + or - has a clear visual correlate — the lips close during all +labial sounds; the corresponding acoustic correlate is that the formants in the adjacent vocalic sounds are “pulled down”, i.e., take lower values than the vowel targets in the region bordering the +labial sound. In this case the visual cue is much stronger and more direct than the acoustic one. Lip rounding is another similar example. In other cases however, the primary articulatory gesture might be related to the position of tongue body (e.g. [high]) that is much less visible; however the acoustic correlate might be more distinct.

(ii) Audio-visual integration might proceed in a multi-pass stage where the acoustic signal is used to identify some features and therefore reduces the recognition problem to a sub-problem of phonemes and the visual cues are then used to disambiguate between the different phonemes. For example, the class of stop consonants (-continuant;+consonantal;-sonorant) can be determined from the acoustic signal relatively easily. However distinctions within such a class are much harder to make acoustically but relatively simple to make visually. The class of nasals are another example where the /m/-/n/ is extremely subtle acoustically but relatively trivial visually.

Given this point of view, the rest of the paper is structured as follows. In section 2, we discuss the visual processing module and how features are extracted from it. In section 3, we discuss the audio processing module and how distinctive features are extracted from the acoustics. We provide some examples of two step audio-visual cues using such an approach. Finally, an interesting interplay occurs with the related problem of audio-visual speech synthesis and we discuss the usefulness of such audio-visual speech synthesizers in providing training data for audio-visual detectors.

2. THE VISUAL SIGNAL AND ITS REPRESENTATION

Here we provide a brief account of the processing of the visual signal to obtain a representation in terms of the facial animation parameters. These parameters can then be used to extract various distinctive features. They can also be used to drive an artificial face synthesis system.

2.1. Face Analysis System

Practical Assumptions The nostrils are by far the easiest facial features to be identified and tracked. They are two symmetric holes in the middle of the face which are darker than the darkest facial hair or skin and are almost never obscured by hair. Many face feature analysis applications allow camera placement suitable for viewing the nostrils (slightly below center view). We don't assume anything about the lighting conditions, skin color, eye wear, facial hair or hair style. The system is also robust in the presence of head roll and tilt (nodding and rotation in the image plane), and scale variations (face to camera distance). The system will fail during periods of excessive head rotation about the neck axis (profile view).

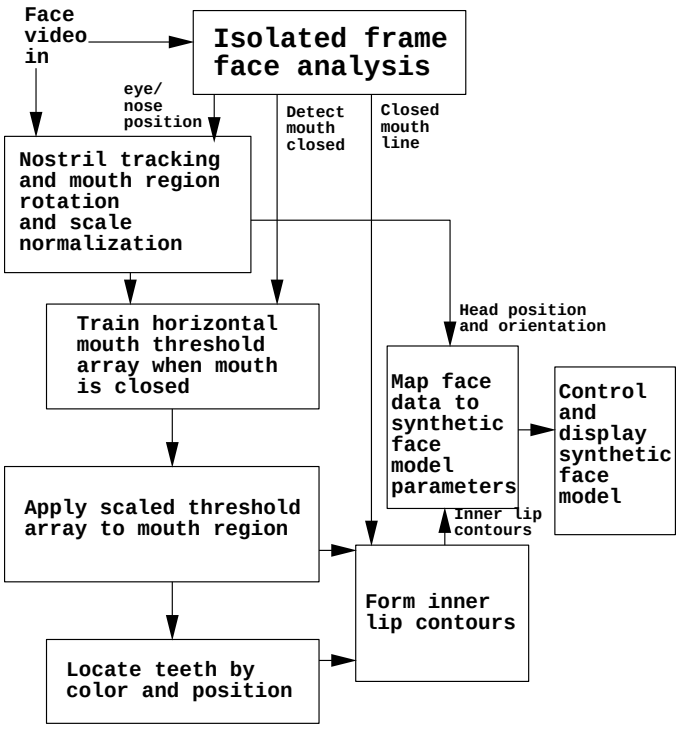

Figure 2: Facial feature analysis block diagram.

Performance Requirements The primary objective of this system is to accurately estimate the inner lip contour of an arbitrary speaker under a large range of viewing conditions. An implicit system performance requirement is extremely robust tracking of the face (eyes, nostrils, and mouth). In addition, the system must reliably detect tracking failures and never falsely indicate the positions of the nostrils and mouth. Tracking failures routinely occur due to occlusion from hands, extreme head rotation, and travel outside of the camera range. Given an accurate estimate of mouth position and orientation in the image plane, the estimated inner lip contour should always lie inside the real inner lip contour to ensure that mouth closure is accurately detected.

System Description Figure 2 shows a block diagram of the major processes used for robust face feature analysis. The process of mapping the acquired facial features to the synthetic face model parameters is also indicated. A face recognition algorithm , periodically finds faces in static frames using morphological filtering and the relative positions of eye, nose and mouth blobs. This information is used to start and then verify the nostril tracking system. A mouth window is formed using the nostril positions. A closed mouth line is combined with the resulting open mouth inner lip contours to give the final inner lip contour estimate. When the mouth is open, the mouth line lies within the inner lip contour. Figure **??** shows the block diagrams for facial feature analysis.

Mouth Detail Analysis Given the head position, scale and tilt estimates from the nostril tracking system, a window is formed around the mouth area which is a fixed distance from the nostrils. To avoid interlace artifacts, each field is analyzed separately. After compensating the mouth image for image plane rotation and scale, a horizontal array of inner lip color thresholds are trained

whenever the mouth is closed. Mouth closure is detected by the face recognition system which looks for a single horizontal valley in the mouth region. The inner lip thresholds are taken to be 90% of the minimum color intensity in each column of pixels in the mouth window resulting in a color threshold for each horizontal position. The inner lip contour threshold array is then applied to the mouth window. Each pixel which is below threshold is labeled as an inner mouth pixel and isolated inner mouth pixels are removed. The teeth are then detected by forming a bounding box around the inner mouth pixels and testing all non-inner mouth pixels for teeth color given as a fixed set of RGB ranges (or YUV ranges). Then a closed contour is formed around the inner mouth and teeth pixels starting from the mouth corners. The upper contour is constrained to have increasing or constant height as the face centerline is approached. The lower contour is constrained to have increasing or constant distance from the upper contour. These constraints were derived from anatomical considerations.

Tracking Failure Detection Applications of face analysis to automatic speech recognition (ASR) require accurate detection of tracking failure. In case of failure, the ASR system would switch to using only the acoustic speech signal. Our system detects tracking failure by applying a combination of constraints on the face parameters which must be met for tracking to proceed. The static face recognition system provides the tracking system all possible eye-nose- mouth(ENM) candidates ordered from most to least likely. For greater efficiency, nostril tracking proceeds using the first ENM candidate if the scale, orientation, and position of the nostrils are sufficiently close to the values from the previous frame. If this is not the case, all ENM candidates are evaluated and a best candidate is selected. If no ENM candidate is close enough to the previously tracked ENM, then multiple frames of ENMs are compared and tracked until a best ENM through time is determined. This startup procedure must be used during system initialization and after the face is either out of camera range or occluded by the hands.

2.2. Visual Feature Representation

There is plenty of research work on tracking facial features. Unfortunately, little work has been done toward methods of selecting a necessary and sufficient set of facial features for visual speech analysis and synthesis. For our work, we use the Facial Definition Parameters (FDP) set defined in MPEG-4/Synthetic and Natural Hybrid Coding (SNHC) [2]. There are 57 FDP points defined in MPEG-4/SNHC. However, for audio/visual speech processing, only those feature points related to speech production, i.e. the feature points on lips and their surrounding area, are relevant. Figure 3 shows those feature points. The displacements of FDP points in MPEG-4/SNHC are measured in terms of Facial Animation Parameters (FAP). FAP is a set of scalar values that represent the displacements of FDP points from their corresponding positions on a neutral face, which is defined as when mouth is close and there is not facial expressions. The FAPs are normalized in terms of mouth width in x direction and the distance between nose tip to mid mouth point in y direction to roughly compensate the size difference between different head models.

All MPEG-4 FAPs are defined as deviations from the neutral face. The face is assumed to be in the neutral position when mouth closure is detected for the first time. The position of the mouth corners, center vertical position, and the vertical positions of the midpoints between the mouth center and the mouth corners is stored for comparison with subsequent frames. The inner lip FAPs for a subsequent frame are computed by first normalizing the orientation and scale of the mouth image using the head roll estimate and the distance between the nostrils as a measure of scale. The horizontal and vertical deviation of the mouth corners are then measured relative to the neutral mouth corner positions. Then the vertical deviation of the upper and lower middle lip points are measured relative to the neutral face. Finally, the vertical deviations of the midpoints between the mouth center and the mouth corners are measured relative to the neutral face. Lip protrusion can also be specified by FAPs but this system does not estimate lip protrusion.

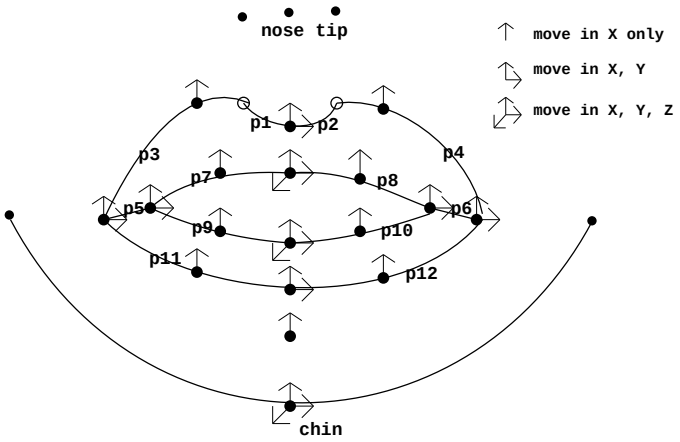

Figure 3: The FDP points around the lips and their corresponding FAPs.

3. LESSONS FROM SYNTHETIC DATA

For training purposeS, it is also possible to obtain synchronized synthetic audio/visual features. Fig. 4 shows a system to extract features from a synthetic talking head driven by a Text-to-Speech (TTS) system. The TTS is the Bell Labs English Text-to-Speech system and the facial animation system is described in [8] with a coarticulation module from [1]. Given text, the text analysis part of TTS produces corresponding phonemes and their durations, which are then mapped into a time series of action parameters through the coarticulation module. The action parameters control the facial animation directly. Synthetic visual speech is obtained by synchronizing the facial animation with the synthetic speech from TTS. This system produces visual speech with very high perceptual quality. Since we have all the 3D information, we can extract the FAPs from the this synthetic talking head directly as described in [12]. For this synthetic data, since we have the aligned ground truth of speech and FAPs segmentation, we could use them to check the validity of the algorithm.

4. DETECTION OF DISTINCTIVE FEATURES FROM ACOUSTICS

An important component of the audio-visual recognition system consists of developing distinctive feature detectors that are based

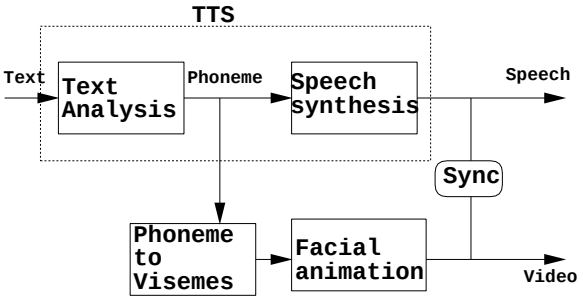

Figure 4: The scheme to generate synthetic audio/visual features.

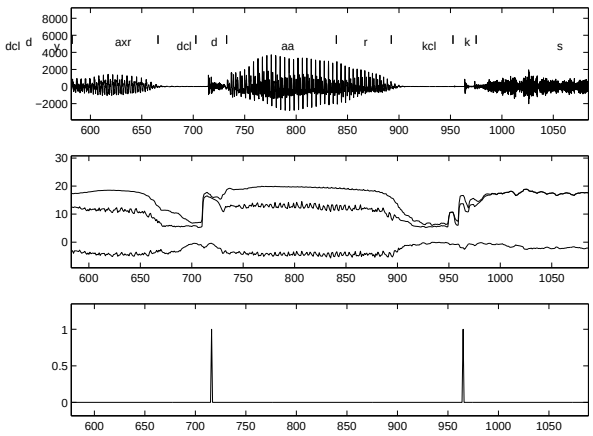

Figure 5: Portion of the speech waveform $s(n)$,(top panel), the associated three-dimensional feature vector, $\mathbf{x}(n)$ (middle panel), and the desired output $y(n)$ bottom panel marking the times of the closure-burst transition.

on the acoustic signal alone. In this section we work through two examples of such feature detectors. In the following section we will consider the interaction of auditory and visual cues to distinctive features.

4.1. Example 1. Stop Detection

Stop consonants are produced by causing a complete closure of the vocal tract followed by a sudden release. Hence they are signaled in continuous speech by a period of extremely low energy (corresponding to the period of closure) followed by a sharp, broad band signal (corresponding to the release). As a result, stops consonants are highly transient (dynamic) sounds that have a varying duration lasting anywhere from 5 to 100 ms. In American English, the class of stops consists of the sounds { p,t,k,b,d,g}.

In order to build a detector for stop consonants in running speech, the speech signal, $s(t)$, is characterized by a vector time series with three dimensions — (i) log(total Energy) (ii) log(Energy above 3kHz) (iii) spectral flatness measure based on Wiener Entropy defined as $\int log(S(f,t))df - log(\int S(f,t)df)$. All quantities are computed using 5 ms windows moved every 1 ms. Multitapered estimates [10] are computed for the spectra from which energies and Wiener entropy are then calculated. Thus, we have $\mathbf{x}(n) = [x_1(n)\ x_2(n)\ x_3(n)]'$ where n represents time (discretized in units of milliseconds) and x_1 through x_3 are the three acoustic quantities that are measured every 1 ms. Energies at 1 ms intervals potentially allow us to track rapid transitions that would otherwise be smoothed out by a coarser temporal resolution. This is particularly important since previous studies (e.g. [5]) indicate that burst durations for voiced stops could be as short as a few milliseconds. The Wiener entropy based flatness measure can be interpreted as a Kullback-Liebler divergence between $S(f,t)$ and a flat spectrum. It is also related to the predictability of the process $s(t)$.

We need to find an operator on the feature vector time series that will return a single dimensional time series that takes on large values around the times that stops occur and small values otherwise. The most natural points in time that mark the presence of stops are the transition from closure to burst release. Shown in fig. 5 is an example of a speech waveform $s(n)$, the associated feature vector time series $\mathbf{x}(n)$ and a desired output $y(n)$. The technical goal is to find an operator h on the time series $\mathbf{x}(n)$ that produces an output $y_h(n) = h \circ \mathbf{x}(n)$ such that $\parallel y - y_h \parallel$ is small in some sense (norm). Specifically, we choose the optimal operator (from some class $\mathcal{H}$ of operators) according to the criterion

$$h_{opt} = \arg\min_{h \in \mathcal{H}} R(h) = \arg\min_{h \in \mathcal{H}} E[(y - y_h)^2] \qquad (1)$$

It is easy to show that this is equivalent to approximating (by y_h) the conditional density time series $p(n) = E[\{y(n)\}|\{\mathbf{x}(n)\}] = P(y(n) = 1|\{\mathbf{x}(n)\})$ for the case when $y(n)$ takes values in $\{0,1\}$. Thus $p(n)$ is the conditional probability of a stop at time n given the time series $\{\mathbf{x}(n)\}$.

Two important questions remain: (i) what is the class $\mathcal{H}$ from which the best operator is to be chosen (ii) how do we deal with the fact that we do not have access to the the true risk $R(h)$ since we don't know the underlying distributions? In [6] we discuss solutions to this problem by assuming a linear class and using the empirical risk to approximate $R(h)$. This approach reduces to a classical optimal filtering problem. In [7] we discuss more complex operators that are based on the structural risk minimization principle of Vapnik [11]. This utilizes a nonlinear class and replaces the true risk with the empirical risk and a large-deviation term. Details are provided in those papers but we show in fig. 4.1. the ROC curves that show false insertion versus deletion as the threshold of acceptance for such a feature detector is varied.

4.2. Example 2. Sonorant/Obstruent Detection

Obstruent sounds are produced with complete or partial closure of the oral tract (stops, fricatives, affricates) while sonorant sounds are produced with the oral and/or nasal tract completely open (nasals, vowels, semi-vowels). As a result of this, sonorant sounds are periodic, have well-developed formant structure, have most of their energy in low frequency regions — conversely, obstruent sounds are a-periodic and noisy, lack a strong formant structure and have considerable high-frequency energy in their spectra. Several detectors have been constructed that are based on this distinction.

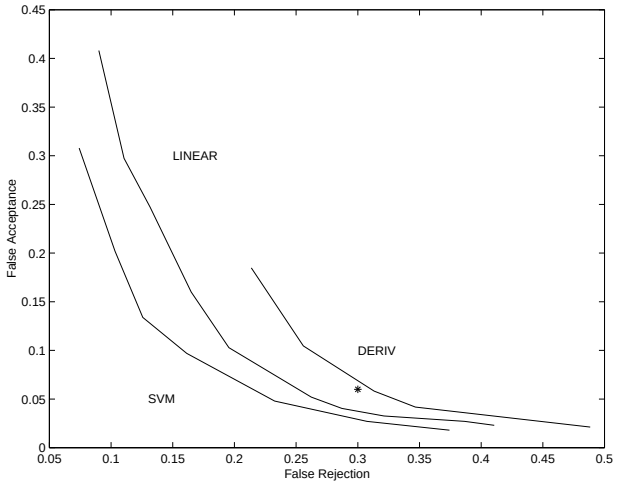

Figure 6: ROC curves for detection of stop consonants using three different algorithms. The three algorithms correspond to choices of h to be a fixed derivative operator; an optimal linear operator; and an optimal non-linear operator (Support Vector Machine; SVM). Detection performance is seen to improve with more complex detectors.

For one class of detectors, the speech signal is represented by four quantities (computed every 5 ms) (i) Ratio of Energy below 3 KHz to that above (ii) Wiener Entropy (iii) Location of Spectral Peak (iv) Dynamic Range of the spectrum. Thus we have a four-dimensional time series $\mathbf{x}(n)$ and as before, the technical problem is to find an operator $h \in \mathcal{H}$ that will map this into a new time series $y_h(n) = h \circ \mathbf{x}(n)$ such that y_h is high in all the obstruent regions and low otherwise. The sonorant/obstruent distinction is not particularly context dependent and so we can reasonably take

$$y_h(n) = h(\mathbf{x}(n))$$

The function h can be linear or non-linear and must be learned from data. Shown in fig. 4.2. is an example output of a sonorant obstruent detector operating on the speech signal. ROC curves similar to the previous case of stops have been constructed but have not been shown here for lack of space.

5. AUDIO-VISUAL DETECTORS

In the previous section, we provided a brief description of how the acoustic signal could be processed to derive distinctive features from it. The input to a particular distinctive feature detector could be derived from audio-visual input as well. In the multi-pass approach however, visual cues that directly provide information about distinctive feature values can be used to disambiguate a confusible class that has been determined by acoustic cues.

Consider fig. 5. where we show three aligned waveforms. The top panel corresponds to the raw acoustic signal that is received. The snippet shown consists of the word "cupid" (phonetic transcription: /kiupid/; only /kiupi/ is displayed here) in a continuous speech context. The speech signal is both noisy and reverberant. A stop detection algorithm based on the acoustics using the principles outlined in the earlier section outputs the time series shown in the second panel. The points where the output is high mark the

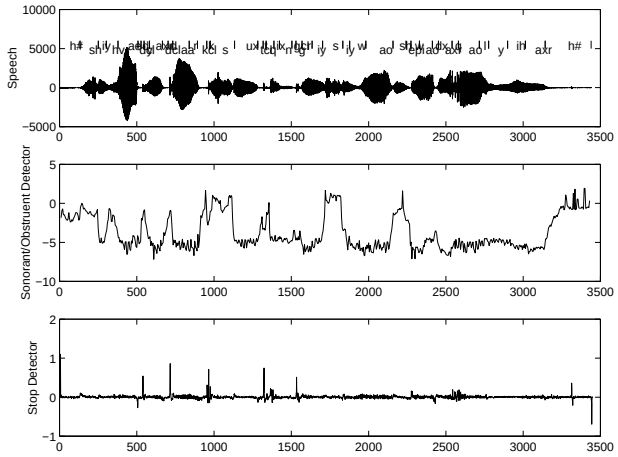

Figure 7: Detector Outputs.

onset of the burst for the two stop consonants /k/ and /p/ respectively. At the same time, a visual routine that measures the distance between the upper and lower inner lip contours (measured at the center of the lips) is plotted in the third panel. This "interlip distance" is a direct cue of lip closure and as is obvious from the figure, this helps to distinguish between the /+labial/ sound (/p/) and the (/-labial/) sound (/k/) very easily.

6. CONCLUSIONS

In this paper, we have explored the possibility of an intermediate level at which integration of visual and auditory features might occur. We have introduced the phonological notion of the distinctive feature and have discussed its articulatory and acoustic status. We have argued that such distinctive features might be recovered from (i) visual stimuli alone (ii) acoustic stimuli alone or (iii) from combined audio-visual input. We have discussed here the visual processing module that potentially allows us to track distinctive features with time. We have also provided some details of our ongoing work on statistical learning frameworks for extracting the distinctive features from the acoustics. Finally, we have shown how the audio and visual input might interact at the feature level prior to lexical access.

7. REFERENCES

1. Cohen, M. and Massaro, "Modeling Coarticulations in Synthetic Visual Speech", In N. Thalmann and D. Thalmann, ed., *Models and Techniques in Computer Animation*, pp. 139-156, Springer-Verlag, 1993.

2. Horne, C. (ed)., *MPEG4/SNHC verification model 5.0*, ISO/IEC JTC1/SC29/WG11 Document N1920, July 1997.

3. R. Jakobson, M. Halle, G. Fant, *Preliminaries to Speech Analysis*, 1952.

4. Kenstowicz, M. "Phonology in Generative Grammar," Blackwell Publishers, 1994.

5. Niyogi, P. and Ramesh, P., "Incorporating Voice Onset Time to Improve Letter Recognition Accuracies," Proceedings of ICASSP, 1998.

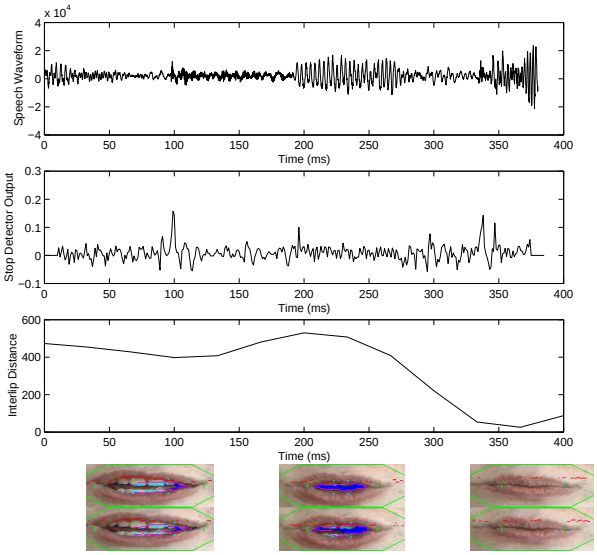

Figure 8: The speech signal, a stop detector based on acoustics, and the interlip distance showing lip closure. The underlying phoneme sequence corresponds to /Vkiupi/ where V indicates vowel. The speech fragment shown here was extracted from continously spoken speech.

6. Niyogi, P., Mitra, P., and Sondhi, M., "A Detection Framework for Locating Phonetic Events," Proceedings of ICSLP-98, Sydney, Australia.
7. Niyogi, P., Burges, C., Ramesh, P., "Distinctive Feature Detection Using Support Vector Machines," Proceedings of ICASSP-99, Phoenix, Arizona.
8. Parke, F. and Waters, K., *Computer Facial Animation*, A. K. Peters Ltd., 1996.
9. Stork, D. G., and Hennecke, M. E. (editors), "Speechreading by Humans and Machines," NATO ASI Series, Springer, 1996.
10. D. J. Thomson, "Spectrum Estimation and Harmonic Analysis," Proc. IEEE, vol. 70, 1055-1096, 1982.
11. Vapnik, V., "Statistical Learning Theory," John Wiley, 1998.
12. Zhong, J., "Flexible Facial Animation Using MPEG-4/SNHC Parameter Streams", Proceedings of ICIP, Chicago, 1998.