UNSUPERVISED CROSS-MODAL ANALYSIS OF PROFESSIONAL MONOLOGUE DISCOURSE

Michael A. Casey and Joshua S. Wachman
MIT Media Laboratory Perceptual Computing Group

This paper describes research in which evidence from audio and visual-kinesic data is combined to obtain an automatic, unsupervised characterization of discourse in the monologues of comedians Jay Leno and David Letterman. We describe the process of obtaining feature vectors from audio and video data and present results of classifying the feature space in terms of statistically significant clusters.

1.0 Introduction

Stand-up comedians make exceptional human expressive data; their speech contours, pauses and gestures are timed to achieve maximum impact on an audience. The goal of this research is to shed light on the problem of automatic feature extraction and classification of cross-modal gestures by the simultaneous characterization of speech and visual-kinesic data.

We break the problem into two parts; feature extraction and data classification. The goal of feature extraction is to reduce the raw signal to a representation which comprises the most perceptually salient attributes of the source. Data classification is the process of identifying and labelling statistically significant densities of these attributes.

1.1 Supervised vs. Unsupervised Analysis

Two well-understood approaches to the problem of data classification are the supervised and unsupervised families of algorithms. Supervised algorithms assume *a priori* knowledge of classes which the data is hypothesized to contain. The data is presented to the algorithm along with the classes and the algorithm finds the best fit of the data to each of the classes. Unsupervised algorithms start with an estimate of the number of classes and automatically identify clusters within the data.

We adopt the unsupervised approach: automatic identification of statistical densities in cross-modal features using the isodata algorithm. The reader is referred to [Therrien89] for details of the isodata algorithm. One benefit of the unsupervised approach is that it is relatively objective. However, the emergent clusters do not necessarily correspond to established gesture types commonly referred to in discourse research literature; see, for example [McNeill92]. In view of this, we do not attempt to relate the results to existing discourse theories; we leave this task for future research.

We now proceed to describe the methods by which a cross-modal data representation for the videotaped monologues can be constructed. The choice of features for representing the data is critical since it is by these features that the data are characterized. Aside from the selection of the features, few assumptions about speech and gesture are made.

2.0 Data Acquisition and Feature Extraction

2.1 Audio-Visual Data Acquisition

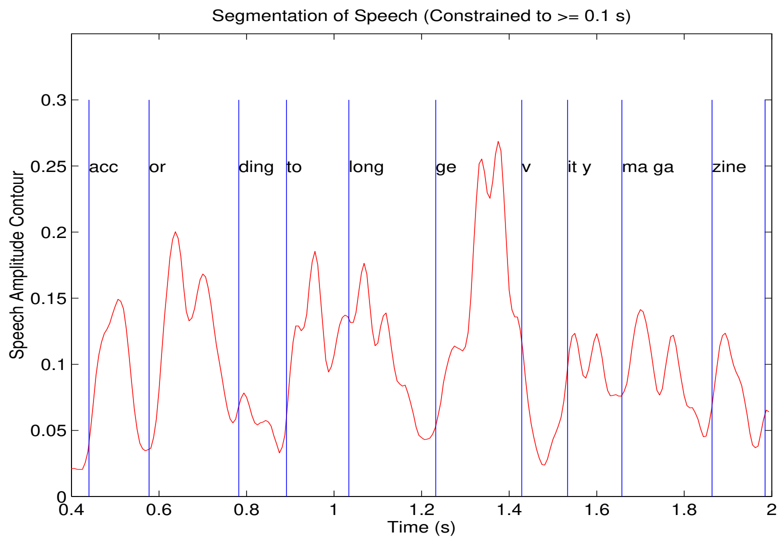
The monologs were taken from “The Tonight Show with Jay Leno” and “The Late Show with David Letterman”. The data was recorded each night for a week (Jan. 22, 1996). Uninterrupted sequences longer than 15 seconds, during which the comedian was both speaking and gesturing were selected from both shows. For Leno, four different sequences totalling 2:21 minutes were used. For Letterman, five different sequences totalling 2:24 minutes were used.

2.2 Time-Normalization and Speech Segmentation Analysis

In order to compare data sets that occupy different temporal orders, such as speech samples and image data, we needed to develop a representation that effectively factored out time. An event-oriented description of the data was developed that used syllabic speech segments as the smallest unit of temporal segmentation for all of the features. The detection of syllabic onsets was performed by extraction of an amplitude envelope for the zero-mean 8kHz speech signal by rectifying, low-pass filtering, decimating and smoothing the resulting signal with a hamming window. The low-pass filter was set to have a cutoff frequency in the order of 30Hz, the rate of syllabic onset information was below this frequency. A threshold was calculated by analysis of the variance of the noise within the first 1/10th of a second of the audio signal where there was no speech; this value was the noise variance in the signal, and the amplitude envelope was segmented at points where it crossed this threshold in the positive direction, see Figure 1.

The resulting segments were constrained to be a minimum of 0.1 seconds in length. This constraint caused some of the syllabic segments to be merged. Figure 1 shows results of the speech segmentation process; the speech, in this case, is Jay Leno saying ``According to Longevity Magazine''. We can see that the syllables ``i..ty'' and ``ma..ga'' have been assimilated to form segments that are a minimum of 0.1 seconds. The figure represents 2 seconds of audio, our data was typically arranged in 30 second sequences. In 30 seconds of speech we obtained 160-220 syllables. The inter-onset intervals were defined as the difference between the time index of successive syllables.

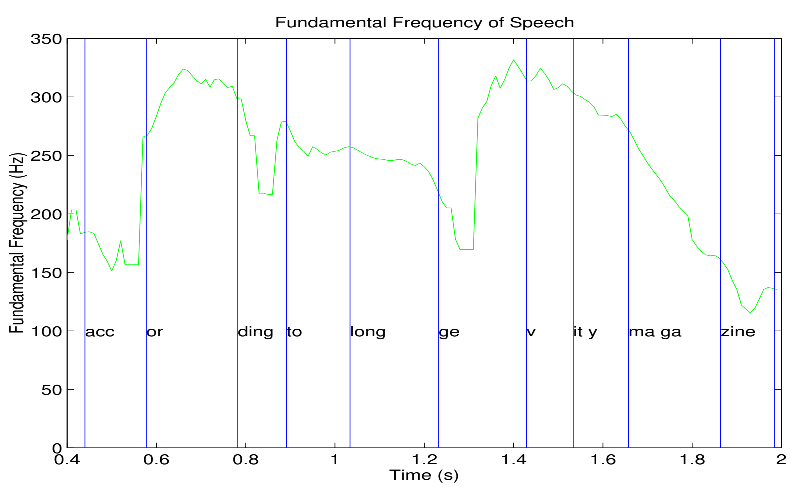
FIGURE 1. Speech Segmentation (Inter-Onset Intervals)



2.3 Fundamental Frequency Tracking

Another of the salient features in speech is the fundamental frequency of the glottal excitation of the vocal tract. The process of extracting fundamental frequency comprises estimating the period of the lowest component of the instantaneous spectrum of the speech. The analysis was done in the frequency domain, for accuracy, using a constant-Q filter bank. Figure 2 shows the continuous fundamental frequency estimates for the section of speech shown in Figure 1

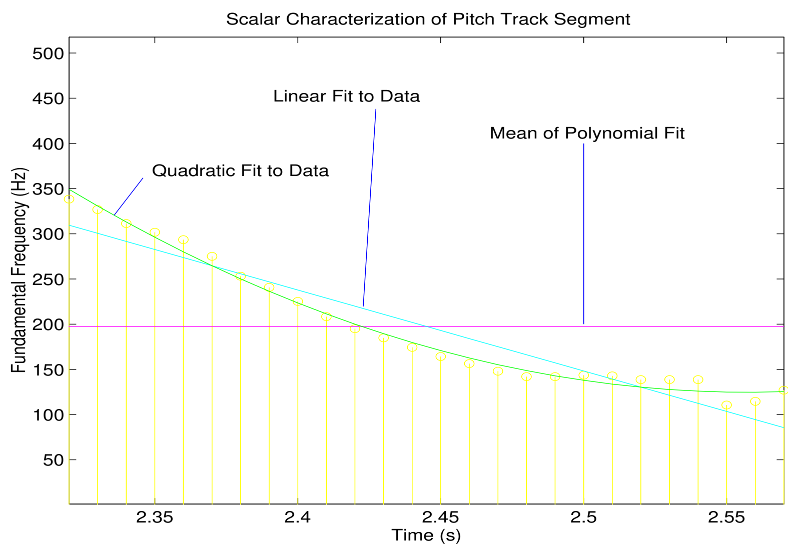
FIGURE 2. Fundamental Frequency Tracking



2.4 Scalar Characterization of Fundamental Frequency Data

In order to obtain a single value for the range of pitch-track values in each speech segment we used quadratic polynomial function approximation. The mean value of samples regenerated from this function approximation was taken as the scalar representation of the samples within the segment window. Figure 3 shows the scalar characterization of a section of the fundamental frequency signal. The stems show the original sample values, and the three lines show a quadratic fit, a linear fit and the mean value of the quadratic fit function respectively.

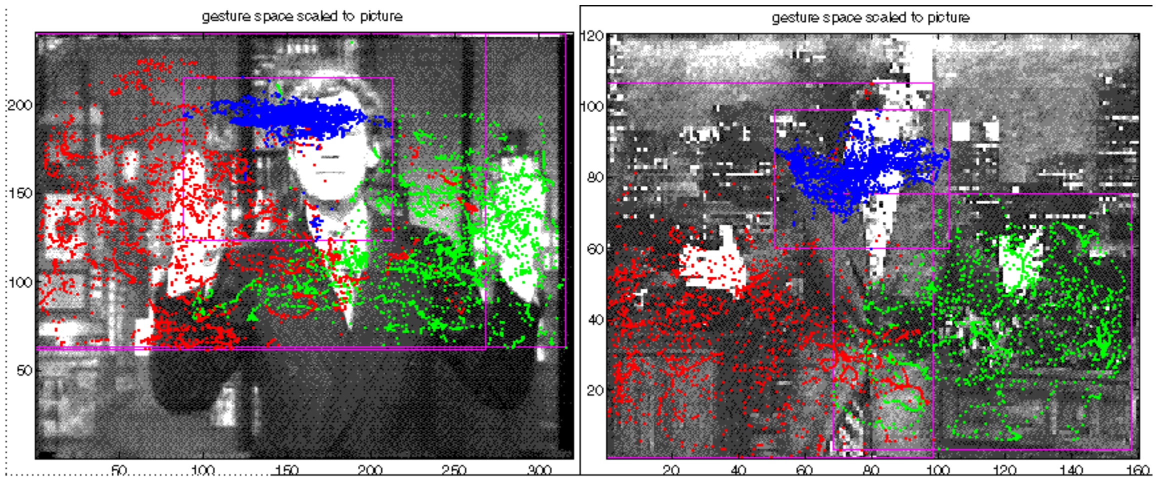
FIGURE 3. Segmentation and Scalar Characterization of Pitch Track Data



2.5 Tracking the Visual Features (Pfinder)

The primary components of the visual data for the purposes of our research were the positions of the hands and the head in each frame of the video data. These positions were tracked by a program developed at the MIT Media Laboratory called Person Finder, or "Pfinder", which tracks homogeneous clusters of luminance-invarient, color pixel data in video sequences, [Wren95]. The operator seeds the position and extent of a representative sample of each class of data. The goal of Pfinder is to find each of the classes in every frame of the sequence with minimal user intervention. Pfinder characterizes each class of data by generating a Gaussian distribution over the initial sample selection and probing the neighborhood region of subsequent frames for class membership; the decision boundaries for membership are based on the color-value variance of the Gaussian function which represents the class. Pfinder outputs the x and y means and eigenvectors for each class. The *x* and *y* means of the left hand, right hand and head constitute the raw visual data that we used. Head and hand positions for sections of monologue from both comedians is shown in Figure 4. These images illustrate the similarity in range of movement and popularity of position in the comedians respective gesture spaces.

FIGURE 4. Visualization of Gesture Space



2.6 Combined Polar Representation of Hand Positions and Velocities

The position of each of the hand classes was converted from rectangular to polar coordinates. The origin for this representation was dynamically calculated to be the *x*-position of the head class and the *y*-position of the bottom of the image frame. Since most of the video was shot as a head-on waist-up view, the head *x*-position and bottom of image frame roughly corresponds to a waist level mid-body origin. The magnitude of each hand's polar vector was then summed giving a single number for the position of both hands. This representation was pursued since it facilitated a body-centered, gesture-space normalized scalar value which could characterize the use of both hands.

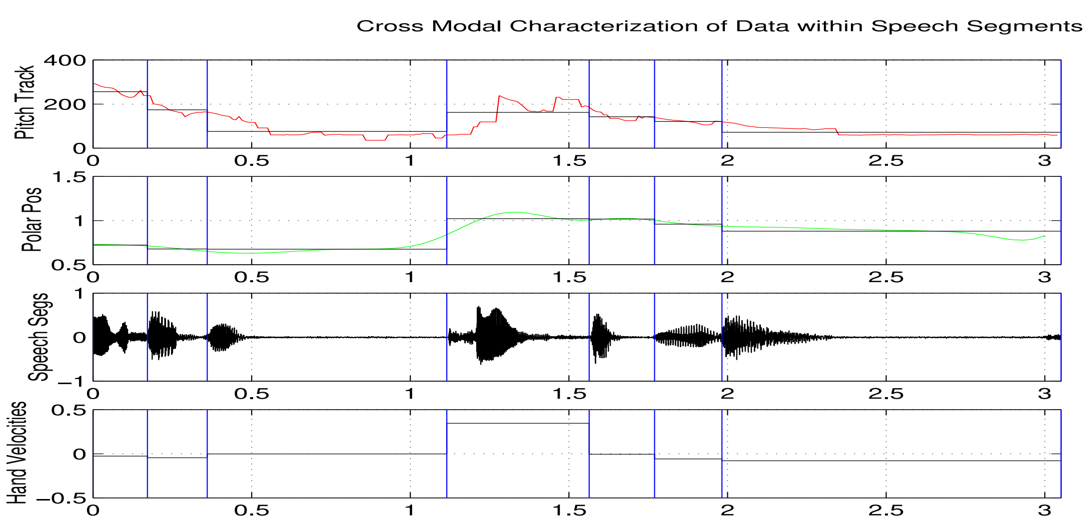
2.7 Segmentation of Visual Features

The temporal segmentation by speech inter-onset intervals in the audio domain was used to bound the gestures in the visual-kinesic domain. The summed polar-magnitude hand positions were vectorized in a manner similar to that of the pitch track. For each segment of the hand trajectory, a quadratic curve that best fit the data was computed. The mean value of the quadratic curve was used as the hand-position feature. The derivative of a linear fit was used as the hand-velocity feature.

2.8 Construction of Cross-Modal Observations

Figure 5 shows the results of forming the cross-modal observations for a section of the Jay Leno data. The staircase plots show the scalar characterization values for each segment overlaid on the continuous feature vectors. The final cross-modal observation matrix contained the syllabic inter-onset intervals, polar hand positions, polar hand velocities, log pitch track and the linear pitch track derivatives. The fourth speech segment shows a correlation in the direction of movement of the features. The ability to represent correlated activity between modalities is a desirable product of the representation. The plots show fundamental frequency, summed polar magnitude of the hands, segmented speech and hand velocity respectively.

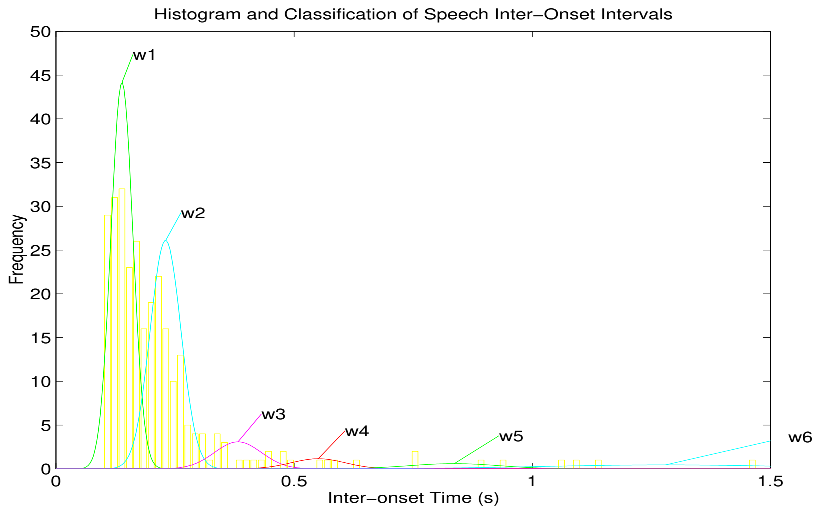
FIGURE 5. Cross Modal Feature Segmentation



3.0 Isodata Classification of Feature Data

Figure 6 shows a histogram of inter-onset intervals for Jay Leno. The overlying Gaussian functions show the classes that isodata identified. The temporal partitioning that these classes provide corresponds to findings in the literature on pauses in speech, [Brown1983].

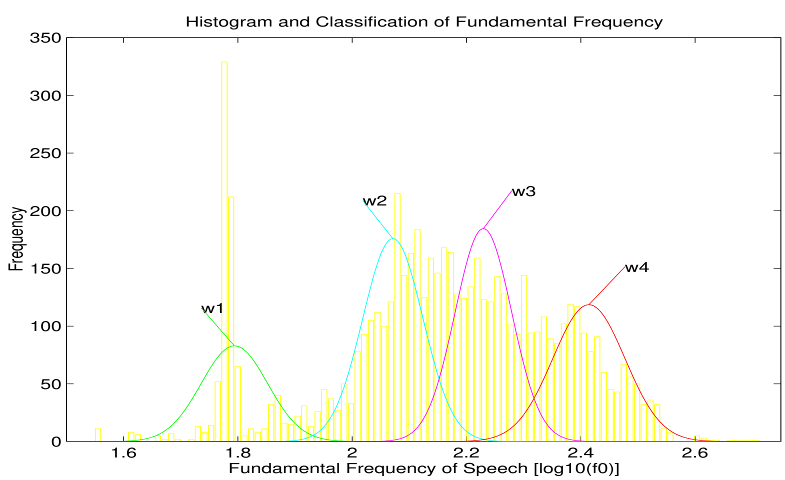
FIGURE 6. Classification of Syllabic Inter-Onset Intervals



3.1 Classification of Fundamental Frequency Usage

As with the inter-onset interval data, we found that the use of fundamental frequency could be characterized as a multi Gaussian density. Figure 7 shows the results of classifying the log of the fundamental frequency for Jay Leno's speech.

FIGURE 7. Classification of Fundamental Frequency Usage



3.2 Classification of Cross-Modal Data

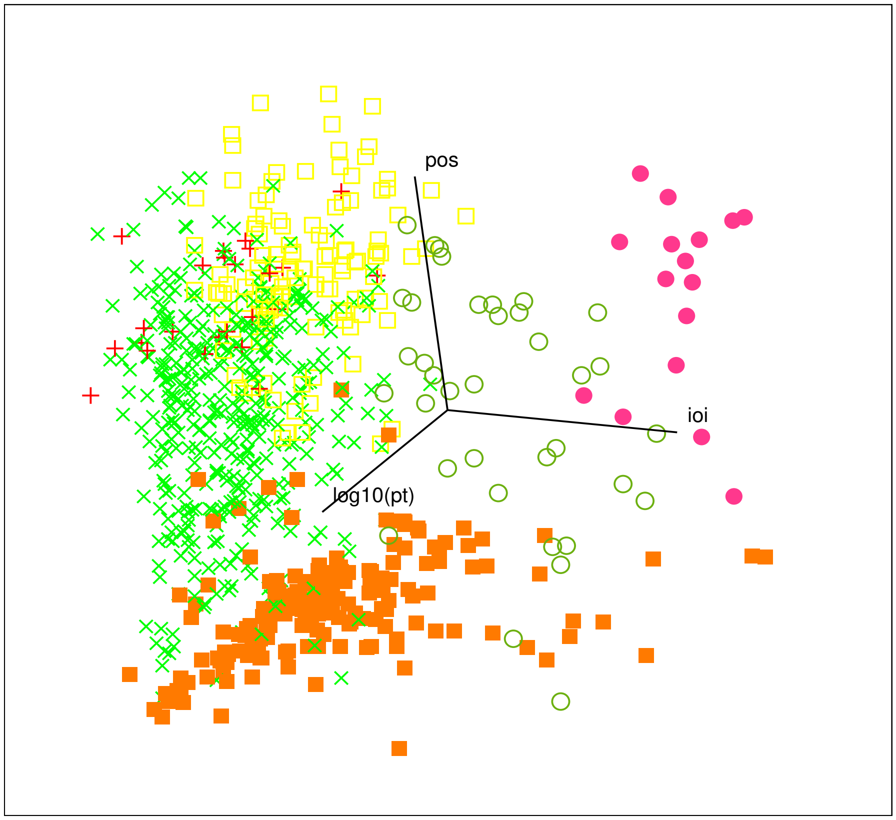
For cross-modal analysis we formed an observation matrix out of five features; syllabic inter-onset intervals, polar hand positions, velocities of the hands, log pitch track and velocities of the pitch track. All the feature vectors were normalized and any outlying data was automatically removed. This ensured that the variance of the data for each feature was conditioned for contributing to the clustering of features in the higher-order feature space.

Figure 8 shows a scatter plot of three of the five features for the combined data of Letterman and Leno: inter-onset intervals, polar hand position and log fundamental frequency. The glyphs highlight the six classes that isodata found in the combined cross-modal data. By inspecting a series of 3-dimensional projections of the five-dimensional data we were able to attribute gesture behaviours to each of the classes in terms of the cross-modal observations. Table 1 shows an analysis of the discriminating features for each class, an NC represents that the referenced feature was considered not to contribute to the given class.

TABLE 1. Characterization of the Cross-Modal Classes Analyzed in Combined Data

Class Identification Glyph	**Inter-onset interval**	**Polar hand position**	**Hand velocity**	**Log Fundament. Frequency**	**Velocity of Fundament. Frequency**
1 (Filled Cube)	NC	short (1 hand)	low-mid	NC	NC
2 (Filled Circle)	long (>1s)	NC	NC	low (<100Hz)	NC
3 (Unfilled Cube)	short (<0.2s)	NC	mid-high	low-mid (50-150Hz)	NC
4 (Unfilled Circle)	mid-long. (>0.2s)	short-med (1 or 2 hands)	mid-high	mid-high (150-400Hz)	NC
5 (Cross)	short (<0.2s)	NC	mid-high	NC	NC
6 (Plus)	short (<0.2s)	long (2 hands)	low	NC	low-mid

FIGURE 8. Clustering and Classification of Cross-Modal Features



Notable classes in this preliminary experiment are class 3- the selection of long pauses with low fundamental frequency which probably correspond to phrase terminations at the end of jokes and class 4- which selects for pauses in speech occurring with moving hands and high fundamental frequency; this class perhaps corresponds to emphasis in which pauses and small hand movements occur simultaneously.

4.0 Conclusions and Future Work

We have presented methods for extracting features for cross-modal analysis of discourse video data. A representation that combined evidence from both audio and visual data was developed that solved problems of time-alignment for the different data. We utilized an unsupervised clustering algorithm for characterizing the information in the feature space of the cross-modal data and presented preliminary results from the classification of 5 minutes of data taken from the monologues of David Letterman and Jay Leno.

Our results suggest that the techniques outlined herein can be applied usefully to the automatic analysis of speech and gesture for human discourse studies. Future work in this area will require consideration of how the emergent unsupervised categories relate to the affective roles of gesture in discourse.

5.0 References

[Brown83] G. Brown and G. Yule. *Discourse Analysis*. Cambridge University, 1983

[Cassell94] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville S. Prevost and M. Stone. Animated Conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple converstaional agents, *ACM Computer Graphics Proceedings*, 1994.

[McNeill92] D. McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago, 1992.

[Therrien89] C. W. Therrien. *Decision Estimation and Classification*, Wiley: New York, 1989.

[Wren95] C. R. Wren, A. Azarbayejani, T. Darrell and A. Pentland. Pfinder: Real-Time Tracking of the Human Body, *SPIE Photonics East*, Vol. 2615 pp. 89-98, 1995.