A DESCRIPTIVE APPROACH TO COMPUTER SPEECH UNDERSTANDING

P.D. GREEN and P.J. GRACE

Department of Computing, North Staffordshire Polytechnic

1 Introduction

It has become a tenet of Artificial Intelligence research that to understand something one must first be able to adequately describe it. This approach has, for instance, been used to great effect in computer vision work by Marr et al (1). It is our intention to apply description-based methods to the problem of understanding spoken sentences.

In computational terms, a description is made by the creation of a rich, symbolic data structure whose elements explicitly represent the items of interest and the relationships between these items. Understanding an utterance then amounts to finding a suitable interpretation of its description. The interpretation process will be based on comparisons of description data structures by symbolic pattern matching (2).

Unless a descriptive language is sufficiently powerful to express in a natural way all kinds of speech events, a system based on it is likely to prove unsatisfactory. We believe this is illustrated by the 'segmentation and labelling' schemes of the ARPA projects(3).

2 Preprocessing

Speech is sampled at a rate of 10 KHz. The preprocessor currently in use outputs the following parameters every 10 ms:-

Estimates of the frequencies and bandwidths of up to 6 formants, derived by an LPC algorithm (4),
The LPC error,
A 128-point approximation to the power spectrum, derived by an FFT algorithm (5)
Energy,
Zero Crossing Rate,
and an estimate of Pitch (6),

3 Describing an Utterance

An utterance description is a hierarchical data structure of descriptors. The descriptors at the bottom level of the structure are primitives, derived directly from the parametric data. We currently implement two kinds of descriptors:

3.1 Line Fragments (LFs)

An LF approximates a short time sequence of parameter values by a straight line. Two approaches to the formation of LFs are being investigated:

a) Finds all plausible LFs of a minimum duration (say 40 ms). This produces a large number of overlapping LFs.

b) Extends the lines as far as possible under control of a threshold. This produces a relatively small number of LFs, at the expense of missing some possibilities.

LFs are formed independantly for each of the Energy, Zero Crossing, Pitch and LPC error parameters. For the LPC-derived Formant Frequency data they are allowed to grow freely within the 6-parameter space (Fig.1), and the associated bandwidth values are used to provide an additional constraint on LF formation. No LFs are derived from the FFT data.

The descriptor for an LF identifies:

The parameter from which it was formed,
Its start and end times and corresponding parameter values,
Its slope,
A measure of goodness of fit of the line to the parameter points,
and, in the case of LFs derived from Formant frequencies, an average bandwidth.

3.2 Energy Band Fragments (EBFs)

To describe noiselike speech sounds like fricatives we need to focus on the spectral areas in which the noise energy lies. Using the FFT spectral data we note, for each 10ms period, the areas in which the normalised distribution of energy over the spectrum exceeds a threshold (Fig.2). Similar areas for successive time periods are then described as an EBF, whose descriptor contains similar details to that for an LF (Fig.3).

3.3 Grouping

We now proceed to find and describe relationships between primitives, then relationships between these new descriptors and so on, building the utterance description in a hierarchical manner. Each new descriptor embodies an approximation to its components and includes pointers to the descriptors from which it was formed, so that nothing is ever lost. As this grouping process iterates, the conditions for linking descriptors are gradually relaxed, so that a series of successively cruder approximations for larger groups of entities results. Two forms of grouping are considered:

Sequential Grouping seeks to associate similar descriptors which are adjacent in time, thus, for instance, linking LFs into longer (and more approximate) lines (Fig.4).

Temporal Grouping seeks to associate dissimilar descriptors which are concurrent in time, i.e. to note that several events appear to have occurred at about the same time. 'Time Groups' are found for

a) Groups of descriptors with roughly the same start time,

b) Groups with roughly the same end time,
c) Groups with both. (Fig. 5).

4 Interpreting an Utterance Description

Out approach to the understanding of an utterance by the interpretation of its description rests on Minsky's idea of 'Frames' (7). A frame is a data structure, compatible in form with the descriptors which constitute the utterance description, which represents some speech event which one hopes to identify. The identification process consists of symbolic matching between the two data structures. A matched frame is said to be instantiated.

A frame is a hierarchical structure of slots, which must be filled if the frame is to be instantiated. Slots may be filled by direct matches, or by the instantiation of other frames, thus allowing the various frames to link into a frame system. There may be other information attached to a slot, such as a default value and tolerance limits.

The unit of speech described by a frame is not fixed. It could range, theoretically, from sub-phonemic to sentential entities, but the idea of a frame description for an individual phoneme has immediate attraction. In essence we hope to describe at least some phonemes by frames which will identify with one or more time groups in the utterance description (Fig.6). The understanding of an utterance will mean finding a set of contiguous frame matches. Since the instantiation of one frame may be influenced by neighbouring frame matches through the frame system, we have a mechanism which may be capable of handling co-articulation effects.

5 References

1. D. MARR 1979 In Artificial Intelligence: An MIT Perspective, MIT Press
   Representing and Understanding Visual Information
2. P.H. WINSTON 1977 Addison-Wesley
   Artificial Intelligence
3. D.H. KLATT 1977 J. Acous. Soc. Am. 62, 6
   Review of the ARPA Speech Understanding Project
4. P.J. BRADLEY & R.C.L. O'NEIL 1977 Proc. Inst. Acous
   Linear Predictive Preprocessing for a Speech Understanding System
5. J.E. ROTHMAN 1968 Decuscope 7, 3 (1968)
   The Fast Fourier Transform and its Implementation
6. W.H. TUCKER & R.H.T. BATES 1978 IEEE Trans. ASSP 26, 6
   A Pitch Estimation Algorithm for Speech and Music
7. M. MINSKY 1975 in the Psychology of Computer Vision, P.H.Winston (ed),
   McGraw Hill
   A Framework for Representing Knowledge

A DESCRIPTIVE APPROACH TO COMPUTER SPEECH UNDERSTANDING

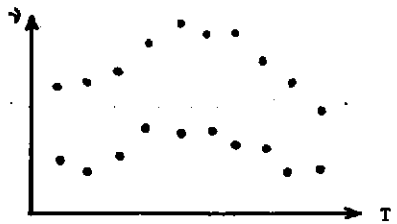

Fig. 1a): Parameters in formant space

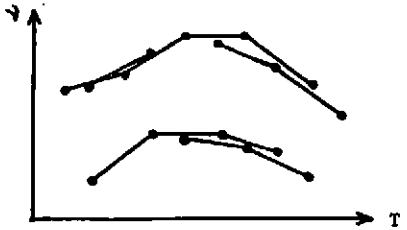

Fig. 1b): Formant LF's allowed to grow freely

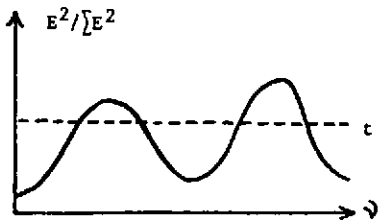

Fig. 2: EBF's where normalised energy exceeds a threshold, t

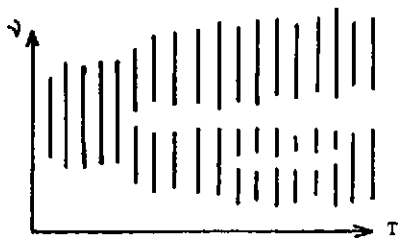

Fig. 3: Changing EBF's in time

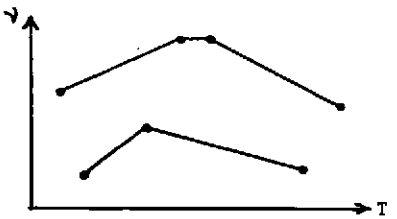

Fig. 4a): Linking LF's which lie adjacent (see Fig. 1a)

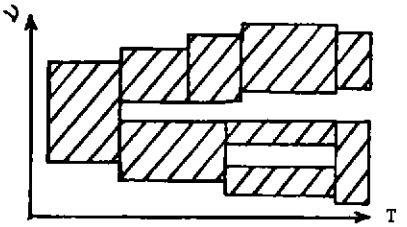

Fig. 4b): Linking EBF's which lie adjacent (see Fig. 3)

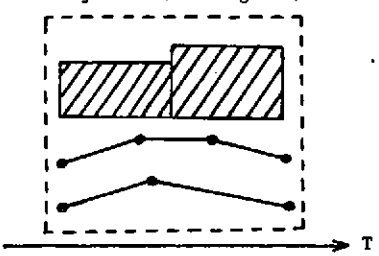

Fig. 5: A time group with similar start and end times.

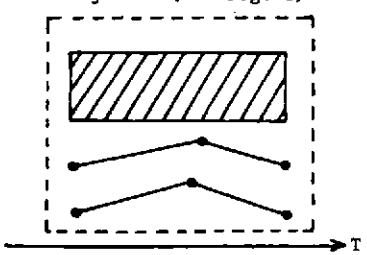

Fig. 6: A frame closely resembling the time group in Fig. 5.