ACOUSTIC, ARTICULATORY, AND PERCEPTUAL STUDIES USING THE MU+ SYSTEM FOR SPEECH DATABASE ANALYSIS

Andrew McVeigh and Jonathan Harrington

Speech Hearing and Language Research Centre, Macquarie University

ABSTRACT

mu+ is a system for retrieving and analysing speech data from large speech databases. The input to the system can be acoustic and articulatory signal files keyed to labels at different hierarchical levels. Using mu+, most combinations of labels, together with their boundary times and associated signal files can be retrieved and analysed. The system has been developed to provide a common environment for experimentation in numerous facets of speech research including: articulatory and acoustic phonetics, prosodic analysis, speech technology research, and linguistic corpus development.

INTRODUCTION

The mu+ system has been developed at the Speech Hearing and Language Research Centre over the last two years in response to the increasing availability of phonetically segmented and labelled speech databases, and the need to retrieve and analyse speech signal files in the database according to different contexts. One of the first systems for speech database analysis was the APS (Acoustic Phonetics in S) system designed at Edinburgh University (Watson, 1989; Harrington & Watson, 1990) as a tool for phonetic approaches to speech technology research. For example, in automatic speech recognition research, APS can be used to extract acoustic parameters for segments types according to different left and right contexts, while in text-to-speech development, APS can provide a knowledge of segment duration from a large speech database.

The mu+ system is an extension of the APS system in two major respects. Firstly, mu+ both derives and extracts a sets of hierarchically structured parallel labels from the acoustic phonetic transcription. The motivation for this is that speech segments often need to be extracted according to phonemic, prosodic, and lexical contexts, as well as left and right phonetic contexts (in order to model segment duration in continuous speech, for example). Secondly, mu+ includes a range of digital-signal-processing routines which can be applied to the extracted segments. This facility enables segments to be D/A converted once they have been extracted and an additional set of time and frequency routines to be applied to segments that have been extracted according to different contexts.

PREPROCESSING

The input to mu+ includes Waves+[1] derived signal files and label files. The signal files consist of the sampled speech waveform and files which are derived from the speech waveform (e.g. formant frequencies and bandwidths, fundamental frequency). There is in principle no restriction to the kinds of sampled data files that can be accepted by mu+, and in our system up to 16 channels of digitised physiological data (kinematic measures of the lip, jaw, tongue; surface EMG; airflow and air-pressure; the laryngographic waveform) can be passed from Waves+ to mu+. The Waves+ label files are produced by a manual segmentation of the speech data. Three different sets of label files are input to mu+ in our laboratory: *acoustic phonetic* labels, *acoustic intonational*, and *gestural*. Details of the acoustic phonetic segmentation are given in Croot, Fletcher & Harrington (this volume). The acoustic intonational labelling is closely based on the system described in Pierrehumbert (1980) consisting of pitch accents, phrase tones (at intermediate phrase boundaries), and boundary tones (at intonational phrase boundaries). The gestural labels are optionally included to mark events such as jaw opening and closing gestures of lip or jaw movement, if physiological data is included.

The acoustic phonetic labels are input to a mu+ preprocessing stage (known as Labed) to derive parallel sets from the acoustic phonetic labels. Labed converts the acoustic phonetic segmentation to a broad phonetic segmentation (Barry & Fourcin, 1992; Croot *et al.*, this volume) using a table of phoneme-allophone relationships. The broad phonetic segmentation is matched against a citation form

phonemic representation which is derived automatically from the orthographic form of the utterance (entered manually) using the top end of a text-to-speech system (Mannell & Clark, 1987). The matching strategy finds the word boundaries in the broad phonetic string. When the word boundaries have been found, grammatical information (content/function distinction) and the lexical stress pattern are retrieved from an on-line lexicon of the text-to-speech system. The words are syllabified (based on the maximum onset principle) and a foot structure is built between the word and syllable levels. Additional processing is carried out on the words based on the acoustic intonational transcription. The words are marked for accent (strong if the word includes a pitch-accented syllable, otherwise weak) and are grouped into intermediate phrases which are either L (low phrase tone) or H (high phrase tone). The last accented word in an intermediate phrase is also marked as strong at the nuclear accent level. Intermediate phrases are grouped into intonational phrases which are either L% (low boundary tone) or H% (high boundary tone).

Following preprocessing, each utterance has the structure shown in Figure 1. Intonation is both linear, as defined by its association to the changing shape of the fundamental frequency contour, and hierarchical, as defined by its structural association to the various levels of the tree shown in Figure 1.

THE MU+ KERNEL

The mu+ kernel consists of a set of routines for extracting most combinations of label-types and the signal files with which they are associated. The label-types are extracted as a *segment list*, consisting of the boundary times of the labels and the name of the utterance from which they were taken; the corresponding signal files are extracted as *track data*.

The expressions which can be used to create segment lists are of four kinds. Firstly, in *membership expressions*, labels can be extracted at any level (all /p/ segments; all content words; all L intermediate phrases) and labels at one level can be extracted with respect to another level (/p/ segments in strong syllables in content words). Secondly, *number expressions* extract labels according to the number of labels at another level ([s] segments in trisyllabic words; content words in intonational phrases of 8 or more syllables). Thirdly, in *sequence expressions*, labels are found with respect to preceding or following contexts at the same level (/p/ segments preceded by /i:/) or at different hierarchical levels (/p/ segments preceded by nuclear accented words). Finally in *structure expression*, labels are found according to their structural position at another level (syllable-initial /p/ phonemes; phrase-medial syllables; utterance-final words).

Once a segment list has been created, it can be passed to a separate function which extracts the corresponding signal file data. The signal file can be retrieved either from segment onset to segment offset, or else values can be retrieved at a single time-point (e.g. the first three formant frequencies at the midpoint of each segment in a segment list).

Segment lists and their associated track data can be interfaced directly to C routines, a C++ class library, a number of programs running in the UNIX environment, and also the S-PLUS environment.

ANALYSIS

At least three different kinds of analysis are possible using mu+. Firstly, analysis routines can be applied to segment lists independently of signal files to obtain information on segment duration and to tabulate segments at different levels (e.g. the number of /p/ phonemes in the database; a distribution of the number of phonemes in function words). Secondly, signal files can be analysed with respect to any sets of labels for which segment lists can be created. In its simplest form, this type of analysis can be used to display different label combinations on different kinds of signal files for each separate utterance, as in Figure 2. Analyses can be made for a large number of segments in the database, as in Figure 3 in which ellipses for three different vowels are shown in the formant plane, but with the corresponding words' orthographic forms plotted at the data points. The third type of analysis involves applying digital-signal-processing routines to the sampled speech data of existing segment lists. Segments can be D/A converted to produce a variety of stimuli directly from the database (e.g. 50 ms sections of [i:] vowels centered at the vowel target; the first 100 schwa vowels in the database produced by a particular speaker; all words spelled "the" produced by a particular speaker). The digital-signal-processing routines

include a standard set of algorithms for time and frequency analysis. They can be used in conjunction with existing S-PLUS primitives to produce displays such as Figure 4 which shows average spectral sections at the burst of [k] stops in different phonetic contexts.

SOME APPLICATIONS OF MU+

mu+ is a multidisciplinary research tool which has evolved with the aim of linking several different aspects of speech research to speech database development. In our laboratory, mu+ has been used extensively for text-to-speech development, both to model segment duration, and to extract diphones directly from the database. Another application of the system has been in classifying strong and weak vowels. In this study, weak vowels are extracted according to different phonetic contexts which are used to train a time-delay neural network system (Cassidy & Harrington, this volume). mu+ can also be used as a tool in speech database development. The D/A facilities discussed in the preceding section can be used to monitor the differences in auditory quality between segments that have been transcribed with the same acoustic phonetic label. Segment boundary placement can be monitored by rapidly processing and displaying several tokens on various acoustic parameters. Outliers in distributions can be identified in terms of an utterance identifier and position in the utterance in order to check that the hand transcription has been appropriately made. These are some of the features which we have found to be valuable in assessing the accuracy of manual segmentation and annotation as the database grows in size.

NOTES

(1) mu+ is not in principle restricted to taking data from Waves+ and can be interfaced to other speech signal processing systems (e.g. Audlab, ILS).

REFERENCES

Barry W.J. & Fourcin A.J. (1992) Levels of labelling. *Computer Speech and Language*, 6, 1-14.

Harrington J. & Watson G. (1990) APS: An Acoustic Phonetic Environment for Speech Research. Centre for Speech Technology Research, Edinburgh University.

Mannell R. & Clark J.E. (1987) Text-to-speech rule and dictionary development. *Speech Communication*, 6, 317-324.

Pierrehumbert Janet B. (1980) The Phonology and Phonetics of English Intonation. PhD dissertation, MIT, Cambridge, Ma. (Distributed by the Indiana University Linguistics Club, Bloomington).

Watson G. (1989) An environment for acoustic phonetic research (abstract). *Journal of the Acoustical Society of America*, 85, S56.

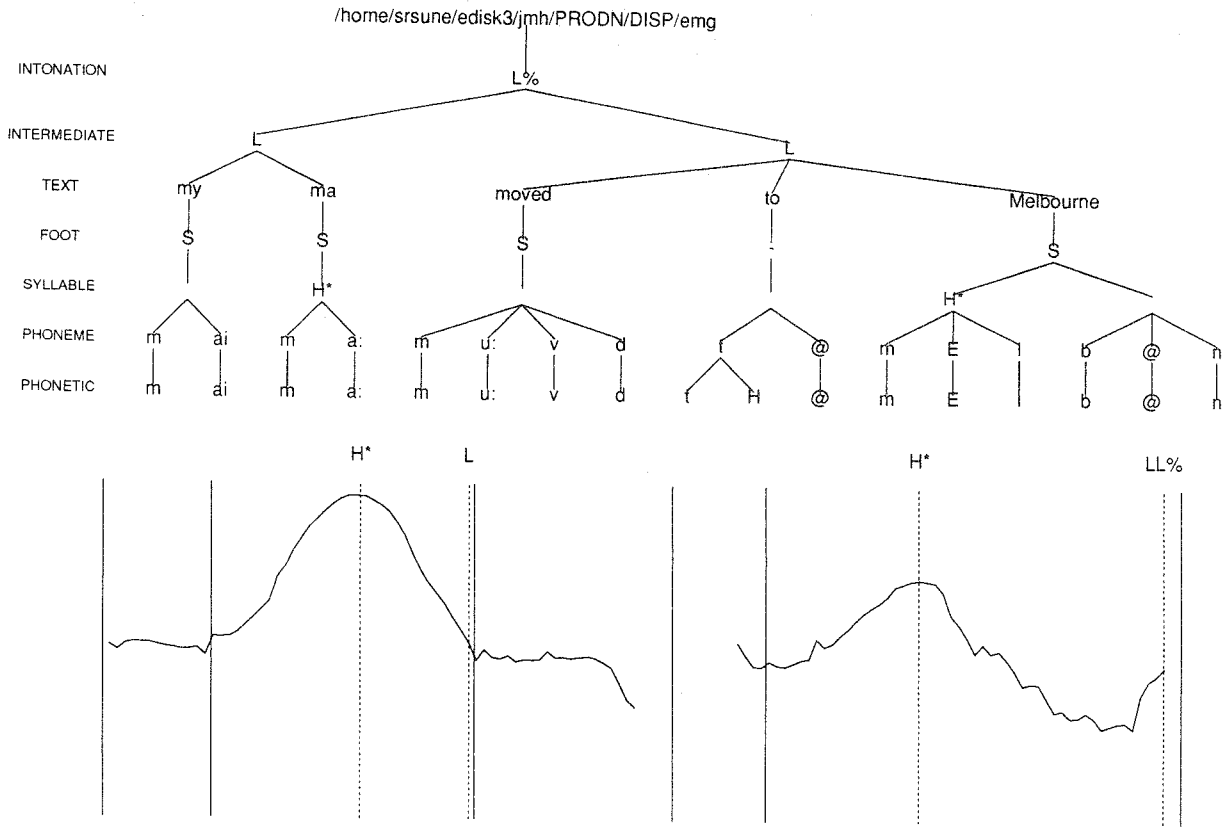

Figure 1: the structural representation of each utterance in the database and its relationship to the F0 contour (solid lines are word boundaries)

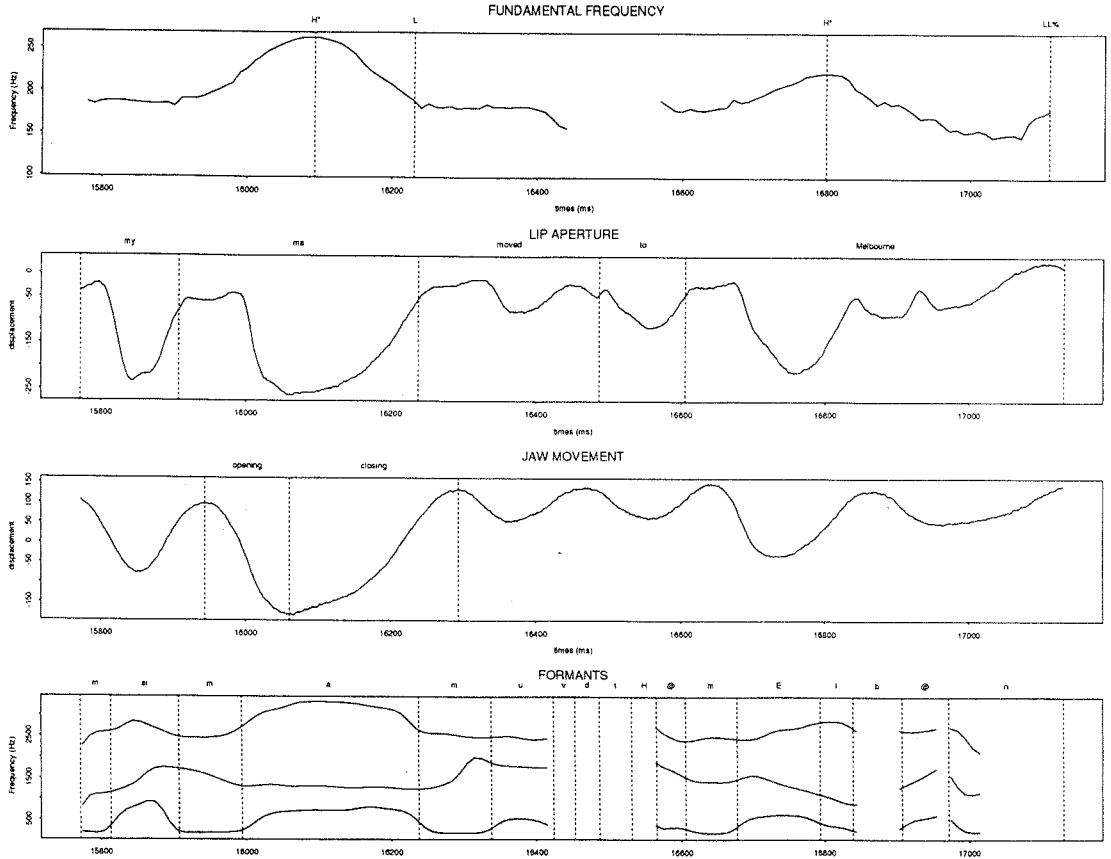

Figure 2: different levels of labelling for articulatory and acoustic data in mu+

Figure 3: ellipse plots for [i:], [u:] and [o:]

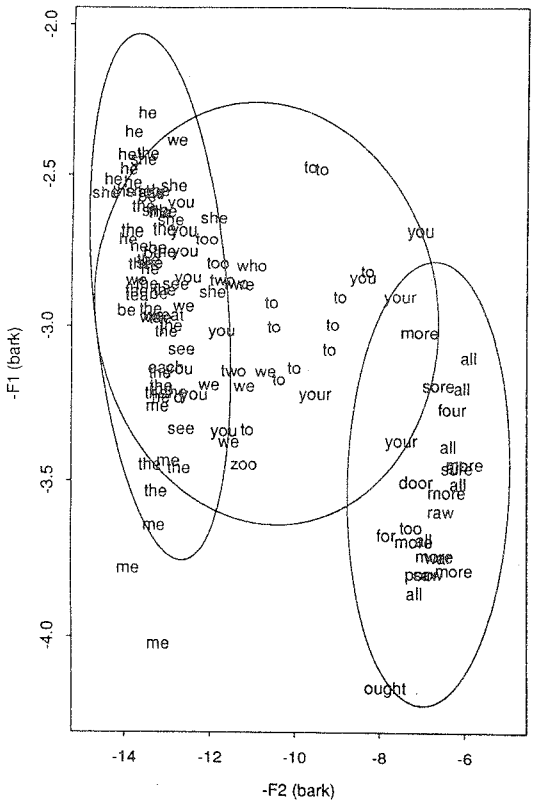

Figure 4: spectral sections each averaged according to the following vowel context