*Computer Speech and Language* (1993) 7, 305–331

# The mu+ system for corpus based speech research

**J. Harrington, S. Cassidy, J. Fletcher* and A. McVeigh**

*Speech Hearing and Language Research Centre, School of English and Linguistics, Macquarie University, Sydney, N.S.W. 2109, Australia*

## Abstract

mu+ is a system for corpus based speech research that can be used to retrieve and analyse segments and their associated signal files from a large speech corpus. The segments can occur at many different levels (acoustic–phonetic, phonemic, intonational, prosodic), while the signal files can include the acoustic speech waveform, analysis parameters derived from the speech waveform (e.g. formant frequencies), and various articulatory measurements (e.g. kinematic parameters from lip and jaw movement). Most combinations of segment types, together with their boundary times and the speech signal files with which they are associated, can be retrieved hierarchically (all phonemes that occur in certain words), sequentially (all phonemes that occur in a particular triphone) or hierarchically and sequentially (e.g. all phonemes that occur in content words which are preceded by an intonational phrase of a particular type). The segments and their associated signal files that are retrieved from the speech database can be analysed subsequently using a wide range of statistical primitives and digital-signal-processing routines. The system has been developed to provide a common environment for experimentation in numerous facets of corpus based speech and language research including: articulatory and acoustic phonetics, prosodic analysis, speech technology research, and linguistic corpus development.

## 1. Introduction

In recent years, there has been an increasing interest in systematically collecting and annotating large quantities of continuous speech data. In the E.C. countries, there have been several projects (e.g. SCRIBE, ACCOR) for speech database collection in the last few years (Hardcastle & Marchal, 1990; Hieronymus *et al.*, 1990; Marchal & Hardcastle, 1990; Hieronymus, 1991; Barry & Fourcin, 1992). The TIMIT database (Lamel, Kassell & Seneff, 1986) created under the DARPA program in the U.S.A. is now used worldwide and more recently DARPA has established a linguistic data consortium specifically for the further collection of spoken and written language data.

There are various reasons for the increased interest in speech database collection. One of the main motivating factors has been the growth of speech technology research. It was

* Now at: Department of Linguistics, University of Melbourne, Australia.

0885–2308/93/040305+27 $08.00/0

soon realized that most experiments in the acoustic phonetics and speech perception literature were based on isolated, citation-form speech, using mostly artificial words or words produced in a set frame such as /hVd/ by a small number of (usually male) speakers of one accent. Such data is of limited relevance to automatic speech recognizers which are ultimately designed to recognize continuous speech from a potentially unrestricted number of speakers of many accents. A second reason for speech database development is that computational resources have only recently become adequate for this task. Since segments in continuous speech are affected by so many different variables (phonetic context, speech tempo, the prosody of the utterance to name but a few), it is usually considered necessary to collect several hours of speech data in order for statistical models that are extracted from continuous speech to become sufficiently robust. In the early 1980s, most laboratories were equipped with single-user computer systems that were capable of storing only small quantities (up to a few minutes) of sampled speech data. At the beginning of the 1990s, there are many more speech laboratories that have the resources (e.g. workstations with more than 20 mips of processing power attached to several gigabytes of external disks) to digitize and analyse the large quantities of speech data that are needed for the various tasks that are relevant to speech technology research.

The increase in the size and scope of speech databases also produces a need for tools that can retrieve systematically any section of the database. More specifically, tools are needed to perform at least four tasks. Firstly, there is a need to retrieve any type of label in a speech database, together with information about the utterance from which the label has been taken, and the time at which the label is positioned relative to the beginning or end of the utterance. (A label in this context is any character which is associated with segmental boundaries marked on a speech waveform.) Secondly, it should be possible to access labels at several different levels of analysis (e.g. phonemes which occur in certain types of syllables; syllables which occur in certain types of words). Thirdly, any speech signal files should be retrievable according to any combination of labels: this would allow parameters such as the first and second formant frequencies to be extracted in relation to the boundary times of the kinds of label combinations mentioned above. Fourthly, the system for speech database analysis should include a wide range of routines for tabulating the labels, displaying the speech data, and carrying out different kinds of multivariate statistical analyses.

These four principles are incorporated in the mu+ system for speech database analysis which is described in this paper. mu+, which was inspired by the APS (Acoustic Phonetics in S) system developed at Edinburgh University (Watson, 1989; Harrington & Watson, 1990), has been developed in the context of the "Australian National Database of Spoken Language" (ANDOSL) which is a collaborative project between four laboratories in Australia (Speech Hearing and Language Research Centre, Macquarie University; National Acoustics Laboratory, Sydney; Department of Electrical Engineering, Sydney University; Department of Computer Science, Australian National University). While there are some similarities between mu+ and its predecessor, APS (e.g. both are interfaced to the S-PLUS language), there are numerous differences. A major structural difference is that APS, as described in Harrington & Watson (1990), was limited to one-dimensional (acoustic phonetic) labels, whereas mu+ both derives and accesses labels from a hierarchically structured labelling system.

The background to the ANDOSL initiative is described in Miller *et al.* (1990a,b). The database which has been created over the last two years at the Speech Hearing and

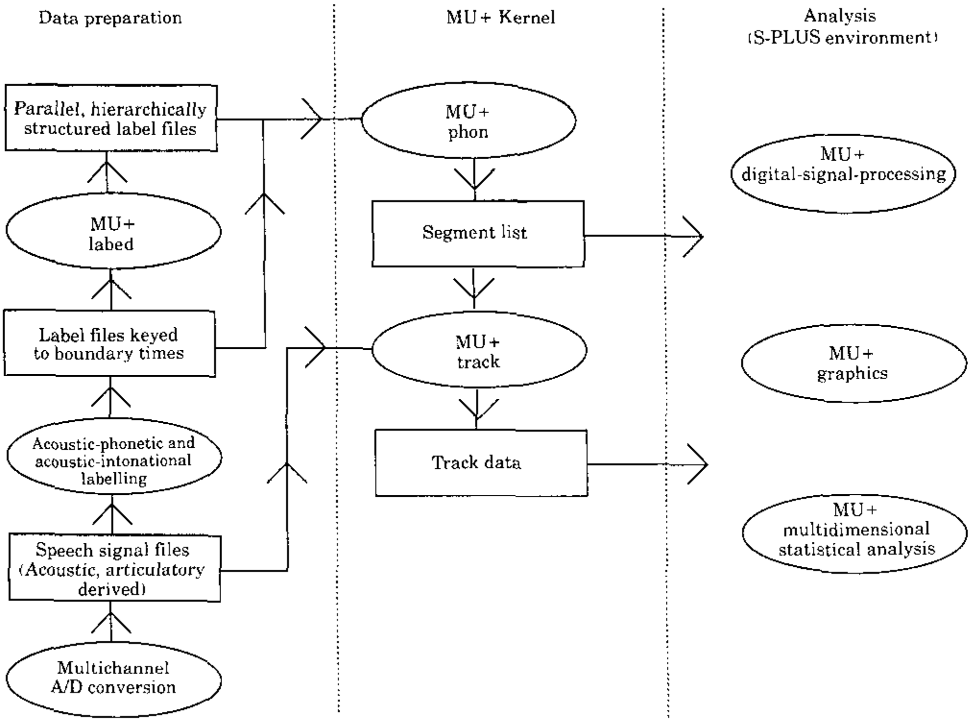

Data preparation
MU+ Kernel
Analysis
(S-PLUS environment)
Parallel, hierarchically structured label files
MU+ labed
Label files keyed to boundary times
Acoustic-phonetic and acoustic-intonational labelling
Speech signal files (Acoustic, articulatory derived)
Multichannel A/D conversion
MU+ phon
Segment list
MU+ track
Track data
MU+ digital-signal-processing
MU+ graphics
MU+ multidimensional statistical analysis

**Figure 1.** The mu+ system. Ellipses are processes, rectangles the output of processes.

Language Research Centre, Macquarie University as part of the ANDOSL initiative (the SHLRC-ANDOSL database) currently consists of 200 phonetically balanced sentences, 200 phonetically dense sentences, and a passage, all taken from the SCRIBE materials (Hieronymus *et al.*, 1990; Barry & Fourcin, 1992), as well as over 400 isolated words, each produced by five male, General Australian speakers. All this data (2000 sentences, 5 passages, 200 words) has been segmented and labelled at the various levels described in the following sections of this paper.

## 2. Overview of the system

An overview of the system for speech database analysis is presented in Fig. 1.

The first part of the system, *data preparation*, consists of three separate stages. Firstly, digitization can be carried out for both the acoustic speech signal, and up to 16 channels of articulatory data. In our laboratory, the articulatory data is taken from the Movetrack articulograph (for recording lip, jaw and tongue movement), surface electromyography (for recording activity in the facial muscles), air-flow and air-pressure equipment, and a laryngograph for recording the laryngeal waveform. Digitization is carried out at 20 kHz using the speech signal processing package Waves+, which additionally derives various analysis files from the speech pressure waveform, including formant centre frequencies and bandwidths, fundamental frequency values, and RMS values.* In the second stage, Waves+ is used for segmentation and labelling at two levels: *acoustic phonetic*, in which the boundary times of acoustic phonetic segments are marked by hand using a combination of auditory and acoustic (usually spectrographic) evidence; and *acoustic intonational* in which pitch accents and boundary tones are marked as events on the changing fundamental frequency contour. In the final stage of data preparation, parallel label files, including broad phonetic, lexical, and various kinds of prosodic labels, are derived semi-automatically from the acoustic phonetic and acoustic intonational segmentations, and the label files are hierarchically structured for each separate utterance.

In the next stage of the system, the *mu+ kernel* is used to extract data from the label files in the form of a *segment list* which consists of segment labels keyed to boundary times and the utterance name from which the segments were taken. Segment lists can be created for most kinds of labels (e.g. [w] segments in syllable-initial position in trisyllabic content words). Once a segment list has been created, *track data*, which can include any of the acoustic or articulatory signal files from the data preparation stage, can be extracted with respect to the boundary times of the segment list (e.g. the first three formant centre frequencies, or the jaw displacement trajectories, for the above [w] segments at the segment midpoint).

In the final stage of the system, most types of *analysis* are carried out on the segment lists and track data in the S-PLUS environment, although it is also possible to interface the data from the mu+ kernel directly to C routines, a C++ class library, and a number of programs running in the UNIX environment. S-PLUS has a number of features which make it particularly suitable for speech analysis, such as its ability to manipulate efficiently vectors and arrays, and the inclusion in the S-PLUS language of numerous primitives for statistical and graphical analysis. Some of the mu+ analysis routines, such as those for graphical analysis, are built directly on existing S-PLUS functions, whereas the mu+ routines for digital-signal-processing and multidimensional statistical analysis, which tend to be more computationally intensive, are built either from the C++ class library, or from C routines, and interfaced to the S-PLUS environment.

The various key stages of the system in Fig. 1 are described more fully below. Section 3 includes details on the criteria for acoustic phonetic and acoustic intonational segmentation (*data preparation*). The different kinds of parallel labels are described in section 4. The methodology for extracting the different kinds of labels and their associated signal files is described in section 5 (*mu+ kernel*). In section 6, an overview is presented of some of the statistical, graphical, and digital-signal-processing routines which can be applied to the segments that are extracted from the database (*analysis*).

## 3. Criteria for segmenting and labelling acoustic speech data

### 3.1. Acoustic phonetic segmentation

Our system for acoustic phonetic labelling of speech data is based on many of the principles that are discussed in Barry & Fourcin (1992) and also Hieronymus *et al.* (1990). As Barry & Fourcin (1992) comment:

* 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).

> "the primary criterion for determining a stretch of the speech signal for the allocation of an acoustic–phonetic label is homogeneity"

and

> "decisions about their placement often require a knowledge of what the sound is in phonetic–phonological terms"

Therefore, syllable-initial, stressed voiceless stops, which usually consist of two separate acoustic events (closure followed by frication/aspiration), are accordingly transcribed as two segments at the acoustic phonetic level. At the same time, an [o:] vowel† is labelled as a single event at the acoustic phonetic level, even though it may show as much spectral change from the onglide to the target, or from the target to the offglide, as glide-to-vowel transitions. Decisions concerning whether a section of the speech signal is to be mapped onto one, or two, acoustic phonetic events, as well as the criteria for the placement of some acoustic phonetic boundaries, are indeed sometimes arbitrary; but they are nevertheless motivated by the principal aim of the transcription at this level, which is to attempt to establish a link between the continuous speech signal and the discrete non-overlapping phonemic units of the language.

While our transcription at the acoustic phonetic level is similar in most respects to Barry & Fourcin's (1992), there are at least two differences. Firstly, our transcription is slightly broader, and makes use of fewer symbols in some cases. In the transcription of oral stops, for example, Barry & Fourcin (1992) differentiate both the closure and the release according to place of articulation (thus [bO bb] and [tO tb] for voiced bilabial and voiceless alveolar stops), whereas in our system, only the closure is distinguished (thus [b H] and [t H]). Our justification for this is economy: the place of articulation of the [H] acoustic phonetic segment is entirely predictable from the preceding segment (a bilabial release when preceded by a [b] closure and an alveolar release when preceded by [t] closure). Secondly, a *single* acoustic phonetic segment can sometimes be mapped onto *more than one* segment at the broad phonetic level in our system. This occurs when a Transcriber clearly perceives the presence of two phonetic events but with no evidence in the physical speech signal for their temporal separation. A typical instance of this would be abutting oral stops which merge to form a single stop closure (e.g. *toxic chemical*; *act*), or doubly articulated sounds as the production of [hænmeɪd] for *hand made*, in which the alveolar and bilabial nasals can be at least partially overlapping. The convention in our system is not to segment such events at the acoustic phonetic level, if there is no acoustic evidence for their separation, but to label them with a single "merged" symbol: thus [A kt H], in which the single [kt] acoustic phonetic segment spans the doubly-articulated velar-alveolar closure, or [h A nm ei d] in which [nm] represents overlapping alveolar and bilabial segments. Although no further segmentation is undertaken at an acoustic phonetic level in such cases, the corresponding broad phonetic segments are nevertheless separated, as discussed in 4.1.

### *3.2. Acoustic intonational segmentation*

Our system for acoustic intonational transcription follows that of Pierrehumbert (1980)

† See Appendix 1.

and Pierrehumbert & Beckman (1988), and more specifically that of the TOBI (tones and break indices) system which is currently used on a national project in the U.S.A. for prosodically annotating speech data (Beckman *et al.*, pers. comm.).

There are three distinct hierarchical levels at which tones can be marked. The first of these is the *pitch accent* level whose domain is the syllable. Four major accent-types are included in this system. Following Pierrehumbert (1980), two pitch accents are H* and L* respectively, the * (star) being a convention which shows that these tones are associated to strong (lexically stressed) syllables. H* corresponds to a tone target in a speaker's upper and mid pitch range. In an L* accent, the tone target occurs in the lowest portion of a speaker's pitch range. The starred tones can be modified to account for a *rising accent* which is L + H* (a peak occurs on the strong syllable, but is preceded by a sharp rise from the lower part of a speaker's range) or a *scooped accent* which is L* + H (the tone target occurs on the strong syllable in the lower part of a speaker's range and is followed by a rising pattern).

The second level at which tones can occur is the *intermediate phrase*. The intermediate phrase is a grouping of words which includes at least one pitch accented syllable: tones at this level are either L (low) or H (high) and are known as *phrase accents*. According to Pierrehumbert (1980), phrase accents are single tones that influence the stretch of speech between the last pitch accent and the edge of the phrase. One of the main criteria which determines the domain of an intermediate phrase is *downstep* which always occurs within an intermediate phrase: that is, the pitch range is reset at intermediate phrase boundaries (Beckman & Pierrehumbert, 1986). H* accents can therefore be marked for downstep (e.g. !H*), where there has been a clear lowering of the accented tone target in relation to a preceding H* pitch accent. Similarly, phrase accents can be marked for downstep in certain circumstances, e.g. in the so-called "calling contour", transcribed as H* !H L% in the TOBI system.

The third tonal level is the *intonation phrase* which is a grouping of intermediate phrases. Tones at this level are known as *boundary tones* and can be either L% (low) or H% (high) (the % sign is simply a marker to distinguish them from phrase accents). Phrase accents are combined with boundary tones to produce four major boundary contours: L–L% (falling declarative), L–H% (continuation rise), H–L% (mid-level tone) and H–H% (high rise). There are many factors which affect the placement of intonational phrase boundaries (see e.g. Halliday, 1967; Pierrehumbert, 1980) including the presence of an actual pause, or non-hesitation pause, and lengthening of the final syllable in the intonational phrase.

Following the conventions described in TOBI, an H* pitch-accent is marked at an observable peak in the F0 trace, while an L* pitch-accent is marked at the lowest observable F0 value. The starred tones of bitonal accents are marked in a similar fashion, depending on whether the major tone is L* or H*. Phrase tones and boundary tones associated with intermediate and intonational phrase boundaries are marked just inside the last segment of the phrase-final syllable. An example of a sentence which has been segmented at the acoustic intonational level is shown in Fig. 2.

## 4. Deriving parallel, hierarchically structured labels

An overview of the system for deriving parallel labels is shown in Fig. 3.

The acoustic phonetic labels which are produced from the acoustic phonetic segmentation are converted automatically by table-lookup to a broad phonetic string. This

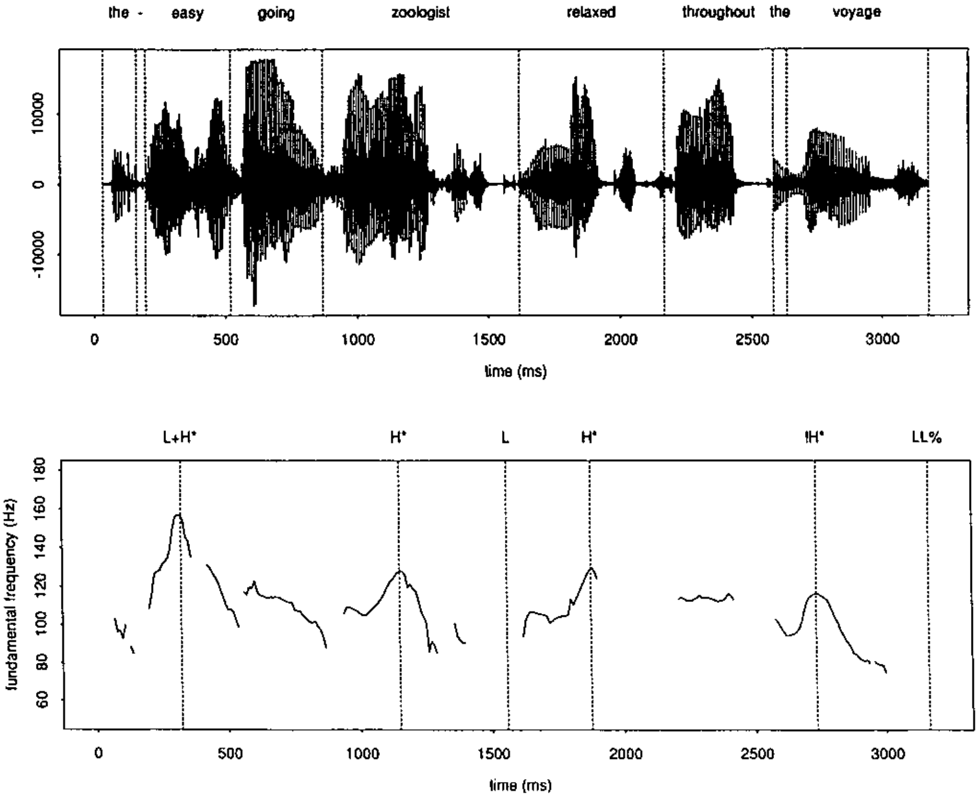


**Figure 2.** Acoustic intonational segmentation and labelling.

conversion can usually be accomplished without prior knowledge of word boundary location (e.g. mapping a voiceless bilabial nasal onto /m/), but sometimes, as in the classic case of whether [t S] is to be parsed as /t/+/S/ or /tS/in *white shoes/why choose*, word boundary information is clearly critical.

Once the broad phonetic labels have been derived, they are parsed into words using a strategy which matches a citation form phonemic representation (marked for word boundaries) to the broad phonetic string (unmarked for word boundaries). The citation form transcription is derived by converting the corresponding orthographic string (entered by the user) using the top-end of a text-to-speech system (Mannell & Clark, 1987). The next two stages involve parsing the broad phonetic segments of each word into syllables, using a table of phonotactically legal onsets, and then grouping the syllables into feet.

The output of the various stages in Fig. 3 is a "grid" showing the relationships between the labels at the different levels (Fig. 4).

The labels which are produced from the acoustic intonational transcriptions are entered on the grid (in the columns Pitch-Accent, Intermediate, and Intonation). When the acoustic intonational labels have been entered, they trigger a separate set of rules which mark any word which has a pitch accented syllable as *strong* (in the Accent column) and the final accented word of an intermediate phrase as *strong*, i.e. nuclear-accented (in the Nuclear column). The completed hierarchical structure is shown as a

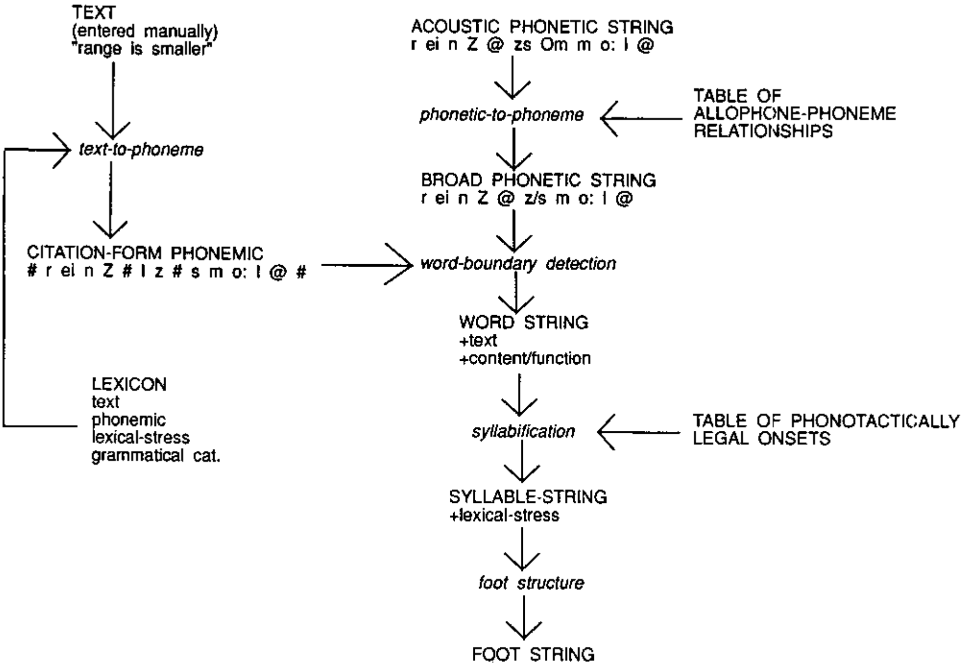


**Figure 3.** Overview of the system for deriving parallel label files.

tree-diagram in Fig. 5, which is equivalent to the "flat" parenthesis notation in the grid in Fig. 4.

Further details on the separate levels are given in sections 4.1–4.5.

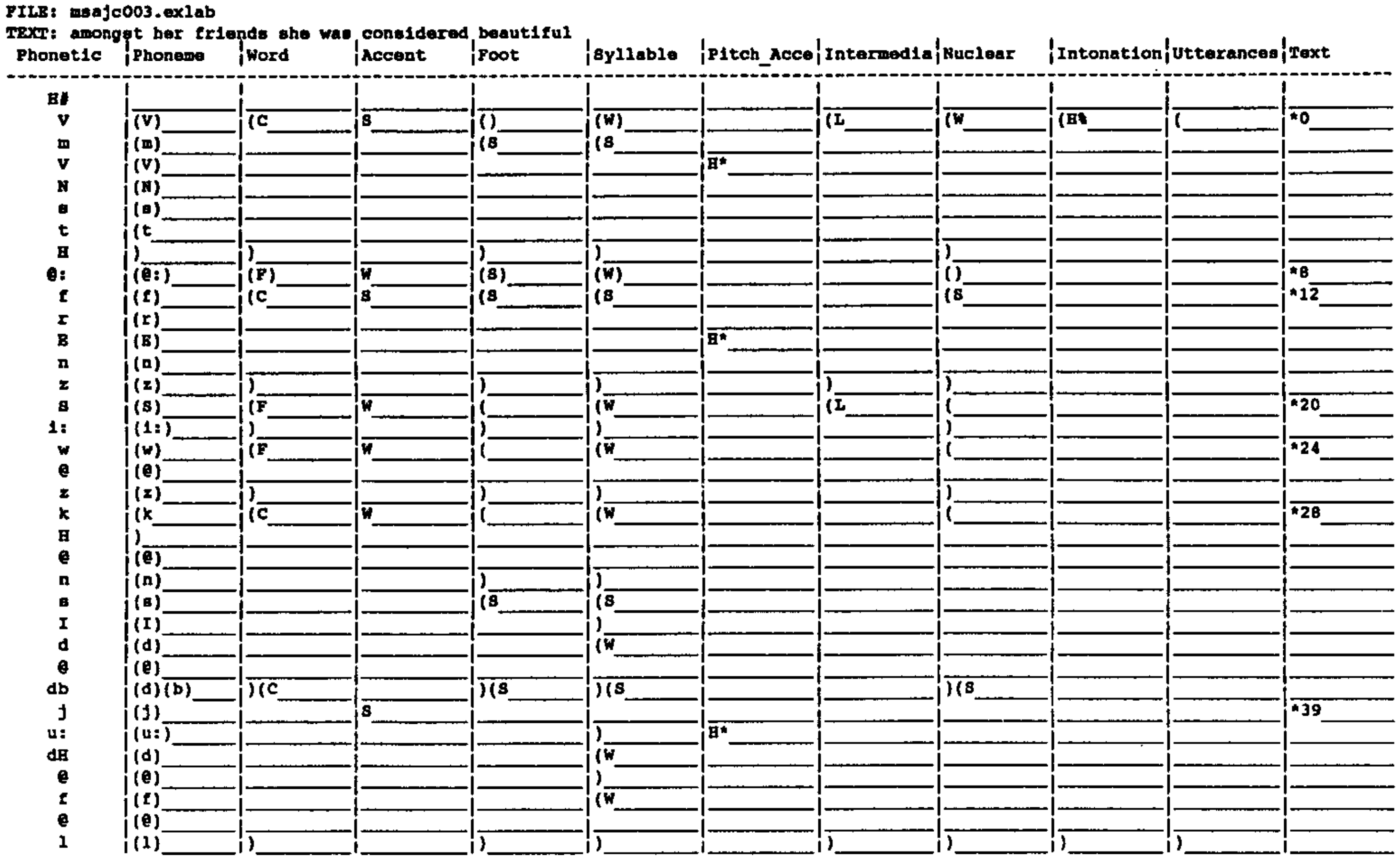
### *4.1. Broad phonetic*

A speech waveform can be annotated at many different levels of phonetic detail. One level which is commonly used in speech database annotation can be called *broad phonetic*. This type of transcription is

> "a level which only employs speech-sound symbols that have a phonemic status (i.e. they distinguish words in English: e.g. /m/ vs. /n/: map vs. nap; /p/ vs. /t/: carp vs. cart etc), but uses them to indicate non-phonemic continuous speech phenomena. ... This level is most economical in that it maximises phonetic information with minimal symbol complexity." (Barry & Fourcin, 1992).

In this type of transcription, word reductions and assimilations are shown and the inventory of labels is chosen from a finite set (e.g. /s p @u z/ for *suppose*).

The mappings from acoustic phonetic to broad phonetic segments are usually one-to-one, e.g. [s] to /s/ or [Om] (voiceless bilabial nasal) to /m/. Sometimes a sequence of acoustic phonetic segments is mapped onto a single broad phonetic segment (e.g. the closure and frication stages of an affricate corresponding to /tS/). A single acoustic phonetic segment can be mapped onto two broad phonetic segments in the examples of the merged segments described in 3.1.



FILE: msajc003.exlab

TEXT: amongst her friends she was considered beautiful

| Phonetic | Phoneme | Word | Accent | Foot | Syllable | Pitch_Acce | Intermedia | Nuclear | Intonation | Utterances | Text |
|---|---|---|---|---|---|---|---|---|---|---|---|
| H# | | | | | | | | | | | |
| V | (V) | (C | S | () | (W) | | (L | (W | (H% | ( | *0 |
| m | (m) | | | (S | (S | | | | | | |
| V | (V) | | | | | H* | | | | | |
| N | (N) | | | | | | | | | | |
| s | (s) | | | | | | | | | | |
| t | (t | | | | | | | | | | |
| H | ) | ) | | ) | ) | | | ) | | | |
| @: | (@:) | (F) | W | (S) | (W) | | | () | | | *8 |
| f | (f) | (C | S | (S | (S | | | (S | | | *12 |
| r | (r) | | | | | | | | | | |
| E | (E) | | | | | H* | | | | | |
| n | (n) | | | | | | | | | | |
| z | (z) | ) | | ) | ) | | ) | ) | | | |
| S | (S) | (F | W | ( | (W | | (L | ( | | | *20 |
| i: | (i:) | ) | | ) | ) | | | ) | | | |
| w | (w) | (F | W | ( | (W | | | ( | | | *24 |
| @ | (@) | | | | | | | | | | |
| z | (z) | ) | | ) | ) | | | ) | | | |
| k | (k | (C | W | ( | (W | | | ( | | | *28 |
| H | ) | | | | | | | | | | |
| @ | (@) | | | | | | | | | | |
| n | (n) | | | ) | ) | | | | | | |
| s | (s) | | | (S | (S | | | | | | |
| I | (I) | | | | ) | | | | | | |
| d | (d) | | | | (W | | | | | | |
| @ | (@) | | | | | | | | | | |
| db | (d)(b) | )(C | | )(S | )(S | | | )(S | | | |
| j | (j) | | S | | | | | | | | *39 |
| u: | (u:) | | | | ) | H* | | | | | |
| dH | (d) | | | | (W | | | | | | |
| @ | (@) | | | | ) | | | | | | |
| f | (f) | | | | (W | | | | | | |
| @ | (@) | | | | | | | | | | |
| l | (l) | ) | | ) | ) | | ) | ) | ) | ) | |

Figure 4. Grid showing the parallel sets of labels. The domain of any segment is denoted by bracketing. "Phonetic" and "Phoneme" correspond to acoustic phonetic and broad phonetic segmentations respectively. "Accent" and "Text" have the same domain as "Word"; "Pitch_Accent" has the same domain as "Syllable". The numerical entries in "Text" are a code for accessing the words' orthographic forms in the on-line lexicon. C/F in "Word" represent content/function labels.

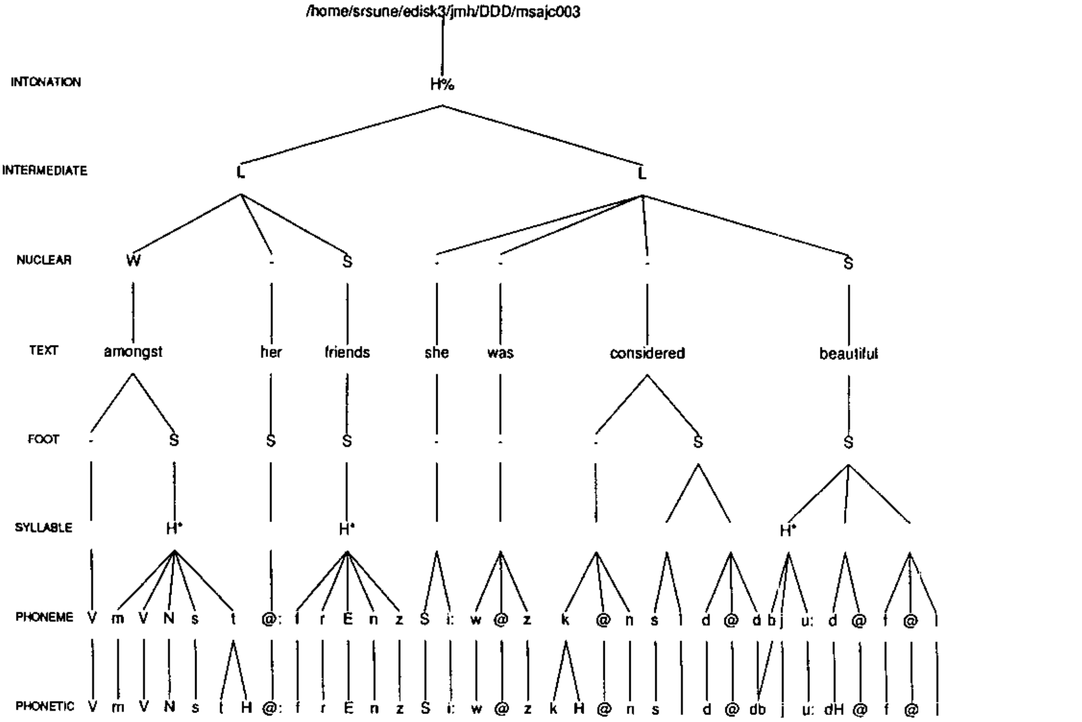

/home/srsune/edisk3/jmh/DDD/msajc003
INTONATION
H%
INTERMEDIATE
L
L
NUCLEAR
W - S - - - S
TEXT
amongst her friends she was considered beautiful
FOOT
- S S S - - - S S
SYLLABLE
H* H* H*
PHONEME
V m V N s t @: f r E n z S i: w @ z k @ n s I d @ d b j u: d @ f @ l
PHONETIC
V m V N s t H @: f r E n z S i: w @ z k H @ n s I d @ db j u: dH @ f @ l

**Figure 5.** The parenthesis notation in Fig. 4 represented as a tree diagram. ("Word", "Accent", and "Syllable", labels not shown).

The boundary times of the broad phonetic segments are inherited directly from the acoustic phonetic segments that they dominate (Fig. 6).

When there is a one-to-one correspondence between the levels, the start and end times of segments at the two levels are the same. When a broad phonetic segment corresponds to a sequence of two acoustic phonetic segments, its temporal extent spans both segments that it dominates. When a single acoustic phonetic segment maps onto two broad phonetic segments (e.g. [nm] to /n m/ in *hand made*), the boundary times of the two broad phonetic segments are the same: they are therefore defined to be *temporally overlapping*. In this sense, although there is no separation between such segments at an acoustic phonetic level, their distinction is preserved at the broad phonetic level.

### *4.2. Words, syllables and feet*

The word boundaries are derived from the broad phonetic segments using an orthographic representation of the utterance, an on-line dictionary, a set of letter-to-sound rules (Mannell & Clark, 1987), and a recursive strategy for finding word boundary candidates (see Fig. 3). In the first step of this process (Fig. 7), a text-to-phoneme conversion is carried out on the orthographic string using dictionary look-up or the letter-to-sound rules if the word is not in the dictionary.

In all cases, a citation form phonemic transcription is derived with word boundaries marked. The citation form transcription is input to the recursive parsing strategy which finds word boundary candidates in the broad phonetic transcription. The strategy aligns

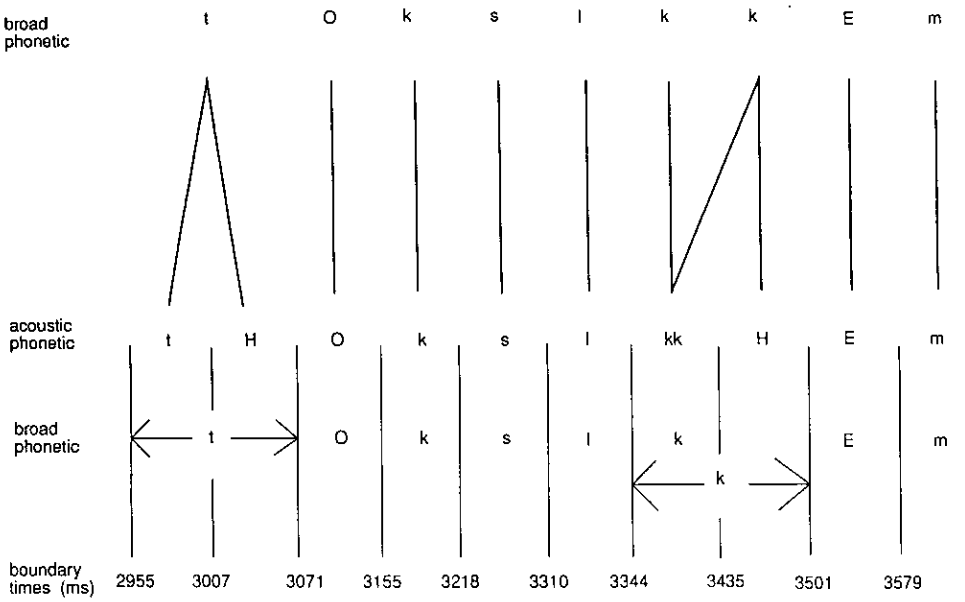


**Figure 6.** Relationship between boundary times of acoustic phonetic and broad phonetic segmentations.

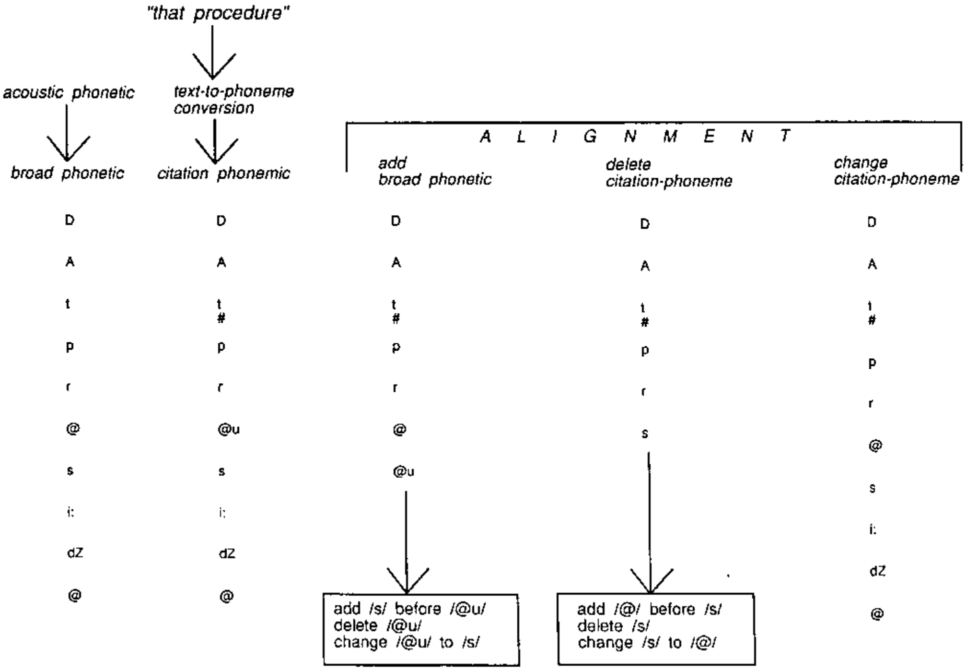


**Figure 7.** The strategy for finding word boundaries in a broad phonetic string.

the derived citation form transcription and the broad phonetic transcription and registers a correct score when the two transcriptions match. In Fig. 7, correct scores are obtained for the first five entries, but not for the sixth which has /@/ (schwa) in the broad phonetic transcription and /@u/ (a diphthong) in the citation form transcription. As a result of this mismatch, an incorrect score is registered and three separate paths are pursued corresponding to *adding* the broad phonetic segment, *deleting* the citation form phoneme, or *changing* the citation form phoneme to the broad phonetic segment. The recursive strategy is then applied to each of these three new possibilities and the scores are added: in the example in Fig. 7, changing the citation form /@u/ to /@/ results in the greatest number of matches and so this path is selected. The non-automatic part of finding word boundaries is that the orthographic forms of the utterances need to be entered: but if different speakers are producing the same sets of utterances, then the same orthographic forms can usually be used each time.

Once the word boundaries have been found, word-internal syllable boundaries are automatically derived, based on the maximum onset principle (Kahn, 1980; Selkirk, 1982), in which as many phonotactically legal consonants as possible are syllabified with a following nucleus. The syllabification rules are applied to broad phonetic segments, not acoustic phonetic segments and a syllable may not span a word boundary. When the syllable boundaries have been found, each word's lexical stress pattern is retrieved from the on-line dictionary. The lexical stress pattern is defined in terms of *s* (strong) syllables, which have a full, unreduced vowel, and *w* (weak) syllables, which usually have a schwa or [I] nucleus. Left-dominant feet are constructed from the lexical stress information by grouping strong syllables with any following number of weak syllables. A foot is marked *s* (strong) if the head syllable has primary lexical stress and as *w* (weak) if the head syllable has secondary lexical stress.

In Abercrombie's (1967) model of rhythm, feet may in fact cross word boundaries and include any word-initial weak syllables: thus *come tomorrow* would be parsed into two feet, with the first boundary occurring after /t@/. Beckman & Pierrehumbert (1988) on the other hand incorporate the notion of *extrametricality* which has been discussed extensively in connection with word stress assignment by Hayes (1982). In Beckman & Pierrehumbert (1988), feet may not cross word boundaries: any word-initial weak syllables are extrametrical (e.g. the /t@/ of *tomorrow* in the previous example) and are not grouped into feet. Abercrombie's (1967) foot is closely related to the idea that feet tend to be perceptually isochronous in English, whereas extrametricality has been shown to be an important mechanism in explaining rules in metrical phonology. There seems to be currently little phonetic evidence to choose between these alternatives and our system (somewhat arbitrarily) makes word-initial weak syllables as extrametrical, as in Beckman & Pierrehumbert (1988). From the point of view of subsequently retrieving the labels, a foot in the sense intended by Abercrombie can still be extracted by identifying all feet followed optionally by a single extrametrical syllable.

The resulting word trees are more or less equivalent to those described in Pierrehumbert & Beckman (1988) with two main notational differences. Firstly, in metrical phonological theory (see Hogg & McCully, 1987; Goldsmith, 1990; for reviews) and specifically in Beckman & Pierrehumbert (1988, p. 148), association lines bypass extrametrical syllables at the foot level and are marked directly as daughters of the word level (Fig. 8).

"Jumping a level" in this way poses various problems for accessing segments from the corpus as well as for aligning start and end times with nodes in the tree. Consequently,

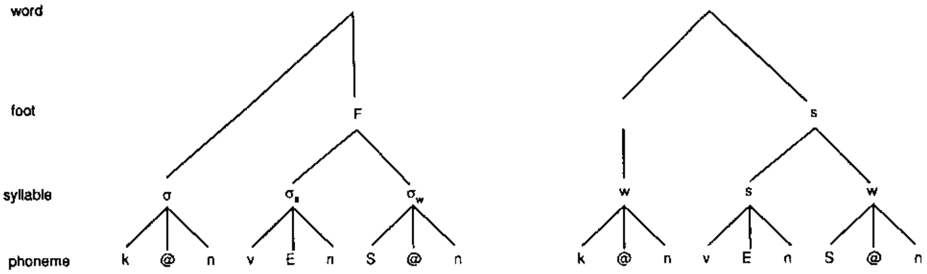


**Figure 8.** Prosodic tree structures commonly used in the metrical phonology literature (left) and in the mu+ system (right).

all syllables are parsed into feet in our system, and extrametrical syllables are marked as "-" which denotes an *empty* category (Fig. 8). Secondly, nodes are usually only explicitly labelled as *s* or *w* at the syllable and foot level in metrical phonology when they cannot be unambiguously inferred from the structure of the tree. However, this kind of economy often makes segment extraction more cumbersome. For example, syllables with primary lexical stress would have to be defined as any of the following: an unmarked foot dominating an unmarked syllable (e.g. *place*); a strong foot dominating an unmarked syllable (the second syllable of *Chinese*); an unmarked foot dominating a strong syllable (the first syllable of *only*); a strong foot dominating a strong syllable (the first syllable of *caterpillar*). By explicitly marking all nodes at the syllable and foot level as either *s* (strong) or *w* (weak), syllables with primary lexical stress can be more straightforwardly defined as *s* syllables that are dominated by *s* feet. Similarly, unstressed syllables can be defined as all *w* syllables. On the other hand, if labels were minimally specified (as in much of the metrical phonology literature), unstressed syllables would have to be defined as syllables that are unmarked and not dominated by a foot (i.e. extrametrical) or syllables which are weak ($\sigma_w$).

Since words, syllables and feet are built on the broad phonetic transcription, they also inherit the start and end times of the broad phonetic segments: therefore when two broad phonetic segments overlap in time, so do the margins of their dominating words, syllables and feet.

### 4.4. Intonation

By carrying out an acoustic intonational segmentation, as described in 2.4, pitch events are associated to the changing shape of the fundamental frequency contour. A separate aspect of intonation is the association of pitch events to hierarchically structured phonological and prosodic categories. Following the principles discussed in Beckman & Pierrehumbert (1988), pitch accents (starred tones) are associated to the phonological syllables which are output from the syllabification rules discussed above; words are grouped into L or H intermediate phrases; and intermediate phrases are grouped into L% or H% intonation phrases. Intonation in our system is therefore both linear, in being associated to sequential, temporal events of the F0 contour, and hierarchical as defined by its association to different levels of the tree (Fig. 9).

Any daughter nodes of the syllable, intermediate phrase, or intonation phrase, inherit the pitch event labels at those levels. Thus /z/ of *zoologist* (Fig. 9) is defined to be both

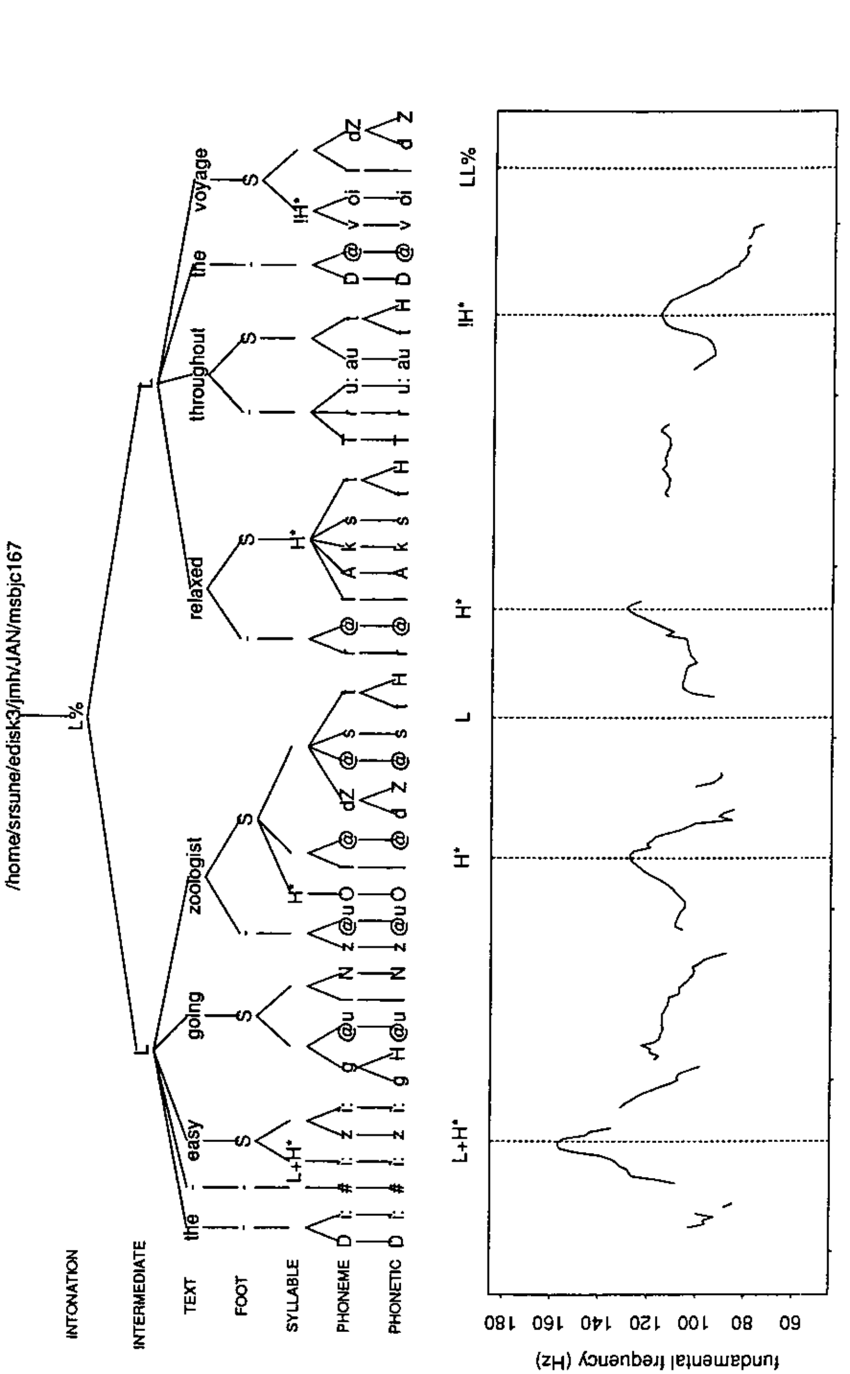


**Figure 9.** Hierarchical and linear representations of intonation.

lexically-stressed and H* because it is dominated by a lexically-stressed syllable with a high tone target; the word *zoologist* (as well as all of its phonemes) is L because it is dominated by a low phrase tone.

The association of intonation to both changes in fundamental frequency and to hierarchical structure produces two sets of times for the pitch events at each of the three different levels. For example, the H* target is defined both by a single time point on the F0 contour (at around 1800 ms), but also has left and right boundary times which span the acoustic phonetic segments [l A k s t H] that it dominates.

### *4.5. Accent and nuclear accent*

There are two types of prominence which depend on the utterance's intonational pattern. The first of these is *accent.* All words which have a pitch accented syllable are accented and *strong*; all other words are unaccented and *weak.* The second level of prominence that depends on intonation is *nuclear accent.* In English, the word which has nuclear accent is entirely predictable from the position of accented words in the intermediate phrase: the last accented word in an intermediate phrase is nuclear accented. While accent can be associated to words, it is not entirely clear what kind of prosodic category defines the domain of nuclear accents. In our system, an additional level (called *nuclear accent*) is created below the level of the intermediate phrase in order to mark the final accented word of an intermediate phrase as strong and all other accented words as weak (see Figs. 4 and 5). Unaccented words do not enter into this relationship and are (somewhat arbitrarily) defined to be extrametrical.

In the absence of further phonetic evidence (see Beckman & Edwards, 1988 and the paper by Selkirk, 1988 for a further discussion of prosodic levels below the intermediate phrase), the level of nuclear accent in Fig. 5 is little more than a notational device for distinguishing between accented words in non-nuclear and nuclear position.

## 5. Segment lists and track data

Most combinations of segments, together with their boundary times and the name of the utterance from which they have been taken, can be retrieved in a hierarchical or sequential manner from the kinds of grid structures shown in Fig. 4. The principle for retrieving segments is most economically represented in Venn diagram terms in which there is a *universe* and any number of *intersecting sets* (Fig. 10).

In general, the universe is made equal to the *label-type* of the segment a user wishes to find, where label-type denotes one of the levels of the grid in Fig. 4 (e.g. Phonetic for all acoustic phonetic segments). Making the universe equal to a label-type means that the universe consists of all the segments in the database for that label-type. The purpose of the intersecting sets is to select a subset of segments from whichever label-type is defined to be the universe. For example, in order to extract all [p] segments that occur in trisyllabic accented words, the universe would be Phonetic (i.e. all acoustic phonetic segments in the database) and there would be three intersecting sets: those acoustic phonetic segments which are [p]; those acoustic phonetic segments in three-syllable words; and those acoustic phonetic segments in accented words (Fig. 10). The desired [p] segments would then be at the intersection of the three sets.

The Venn diagram translates into a search pattern in mu+ in the following way:

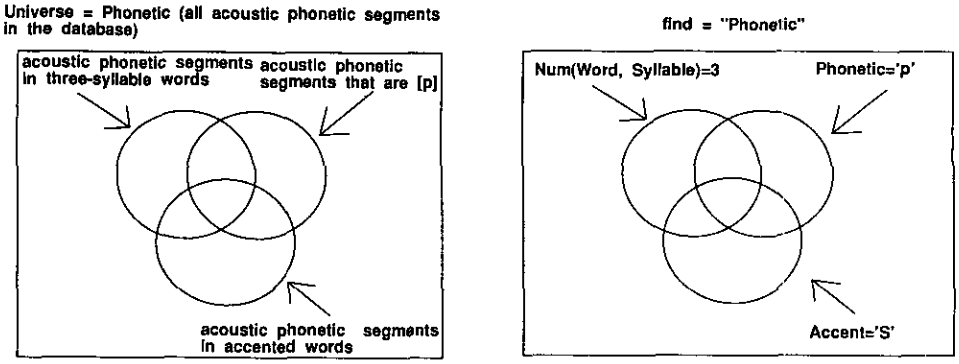

Universe = Phonetic (all acoustic phonetic segments in the database)
acoustic phonetic segments in three-syllable words
acoustic phonetic segments that are [p]
acoustic phonetic segments in accented words
find = "Phonetic"
Num(Word, Syllable)=3
Phonetic='p'
Accent='S'

**Figure 10.** The basis for creating segment lists in mu+.

(1) phon("expression 1 and/or expression 2 . . . and/or expression *n*",
find = "x")

In the above function, *x* is the label-type which is defined to be the universe and each expression represents a different set. The *and* operator denotes the intersection of sets and the *or* operator is the union of sets.

There are four different kinds of expressions that define the intersecting sets (Table I). In the first kind, a *membership* expression, segments can be selected for any label-type (e.g. all /z/ phonemes, all H* pitch accented syllables). The second kind of expression, *number*, allows the number of segments at one level to be defined with respect to another level (e.g. intermediate phrases of six or more syllables). The third type of expression, *structure*, allows the structural position of segments at one level to be defined with respect to another level (e.g. phrase-final phonemes). The last type of expression, *sequence*, allows a target segment to be found according to a preceding and/or following context, either at the same hierarchical level (/d/ phonemes preceded by /n/ phonemes) or at different hierarchical levels (/d/ phonemes preceded by L intermediate phrases). Sequence expressions can be combined with number expressions (preceding syllables in trisyllabic words), and with structure expressions (preceding word-final syllables).

The expressions in Table I allow segment lists to be created for most types of search pattern. An example of a moderately complicated pattern would be: "find the phonemes of all word-medial strong syllables in phrase-final words such that the following word in *then* or a content-word and such that the preceding phoneme is a schwa" (e.g. /mO/ in "*If Carol comes tomorrow, then*. . ." or "*If Carol comes tomorrow, arrange*. . ." etc.). The expression for this search pattern is:

(2) phon("Medial(Word, Syllable) = T and End(Intermediate, Word) = T and
(Text[1] = 'then' or Word[1] = 'C') and Phoneme[-1] = '@'",
find = "Phoneme")

Once a segment list has been created, corresponding track data can be extracted for any existing signal files:

(3) track(x, "fm[1:3]", cut = 0·5)

TABLE I. The different kinds of expressions that can be used in creating segment lists

| | |
|---|---|
| **1. Member** | |
| Phonetic = ’m’ | [m] acoustic phonetic segments |
| Foot = ’S’ | strong feet |
| Text = ’the/for’ | words spelled *the* or *for* |
| Pitch_Accent = ’H*’ | syllables with H* pitch accents |
| Nuclear = ’W’ | pitch-accented words in non-nuclear position |
| Intermediate = ’L’ | intermediate phrases with low phrase tones |
| **2. Number** | |
| Num(Utterance, Intonation) = 2 | utterances of two intonational phrases |
| Num(Intermediate, Foot) > = 3 | intermediate phrases of three or more feet |
| Num(Word, Phonetic) < 5 | words of less than five acoustic phonetic segments |
| **3. Structure** | |
| Start(Syllable, Phoneme) = T | syllable-initial phonemes |
| Medial(Intermediate, Syllable) = F | syllables that are not phrase-medial |
| End(Utterance, Phoneme) = T | utterance-final phonemes |
| **4.1 Sequence (member)** | |
| Phoneme[-1] = ’s’ | preceding /s/ phonemes |
| Text[2] = ’Carol’ | the next word plus one spelled *Carol* |
| **4.2 Sequence (number)** | |
| Num(Word[-1], Syllable) = 3 | preceding words of three syllables |
| Num(Intermediate[1], Foot) = 4 | following intermediate phrases of four feet |
| **4.3 Sequence (structure)** | |
| Start(Word, Phoneme[-1]) = T | preceding phonemes that are word-initial |
| Medial(Intonation, Syllable[1]) = T | following syllables that are phrase-medial |
| End(Syllable[-1], Phoneme) = T | syllable-final phonemes in preceding syllables |

extracts the first three formant centre frequencies for an existing segment list $x$, at the segment midpoint.

## 6. Data analysis

Once segments have been retrieved from the speech database, they can be analysed in various ways. Since segments are also retrieved together with their boundary times, they can be straightforwardly analysed for duration. In a recent study of vowel duration in Australian English, durations were analysed in 6475 vowels in 498 phonetically balanced and dense sentences (the SCRIBE corpus sets A and B).

The vowels were extracted and durations calculated according to lexical stress and

TABLE II. Table of durations (in milliseconds). Left, middle, and right columns are mean, standard deviation, and number of tokens respectively

| | Bernard and Mannell (1986) | SHLRC Pitch accented vowels | SHLRC Final vowels | SHLRC Stressed vowels |
|---|---|---|---|---|
| i: | 256 38 (170) | 124 52 (116) | 144 50 (74) | 79 27 (471) |
| I | 136 29 (169) | 61 17 (163) | 68 22 (33) | 53 18 (539) |
| E | 159 28 (170) | 83 27 (183) | 123 35 (24) | 73 22 (194) |
| A | 201 35 (170) | 120 36 (150) | 161 51 (17) | 98 30 (153) |
| a: | 315 41 (169) | 175 50 (74) | 224 56 (14) | 145 32 (71) |
| V | 165 29 (168) | 87 24 (108) | 119 32 (22) | 83 21 (126) |
| O | 182 28 (169) | 105 32 (123) | 136 41 (12) | 89 24 (131) |
| o: | 295 40 (169) | 140 51 (107) | 184 61 (20) | 114 36 (137) |
| U | 155 28 (169) | 65 20 (15) | 81 16 (2) | 59 19 (43) |
| u: | 268 40 (166) | 112 48 (93) | 151 60 (19) | 73 31 (166) |
| @: | 283 40 (167) | 141 30 (62) | 177 33 (12) | 114 30 (49) |
| ei | 303 39 (171) | 145 40 (116) | 195 55 (23) | 128 34 (184) |
| ai | 314 43 (170) | 172 51 (106) | 237 67 (21) | 135 36 (214) |
| oi | 285 39 (170) | 170 53 (21) | 178 43 (12) | 156 37 (28) |
| au | 317 41 (168) | 174 52 (43) | 217 66 (10) | 142 30 (74) |
| ou | 289 36 (171) | 146 41 (76) | 172 52 (29) | 124 32 (138) |
| i@ | 282 40 (166) | 180 77 (17) | 270 52 (12) | 118 34 (17) |
| e: | 273 44 (166) | 164 75 (27) | 271 29 (6) | 110 40 (38) |
| u@ | 265 42 (146) | — — — | — — — | — — — |
| @ | — — — | — — — | 80 42 (150) | — — — |

pitch-accent, and according to whether they occurred at intermediate or intonational phrase boundaries (Table II).

The boundary times of the segments that are retrieved from the database are also indexed into any existing signal files (via the track() function, as described above) and this makes it possible to carry out various analyses on signal files with respect to most phonetic, phonemic, lexical, and prosodic contexts. Two examples are shown. The first is taken from a study examining the relationship between vowel reduction and word-type using displays in the formant plane, but showing the words' orthographic labels, rather than the phonetic labels of the segments themselves.

The second is concerned with assessing the influence of preceding context (specifically of preceding voiced oral stops) on formant frequencies at the vowel onset and the vowel target. In this second analysis, the influence of left phonetic context was determined from time-normalized formant data and the same data was averaged according to the different phonetic contexts (Fig. 12). Additionally, locus equations were computed (Sussman *et al.*, 1991) with the aid of displays showing the computed locus frequency and the labels of the vowel tokens in the onset/vowel target plane (Fig. 13).

The digital-signal-processing routines that can be applied directly to segment lists' sampled speech data enable further kinds of analyses to be made. The simplest of these is D/A conversion which can be used to construct natural stimuli for use in speech perception experiments directly from the speech database. The stimuli can be con-

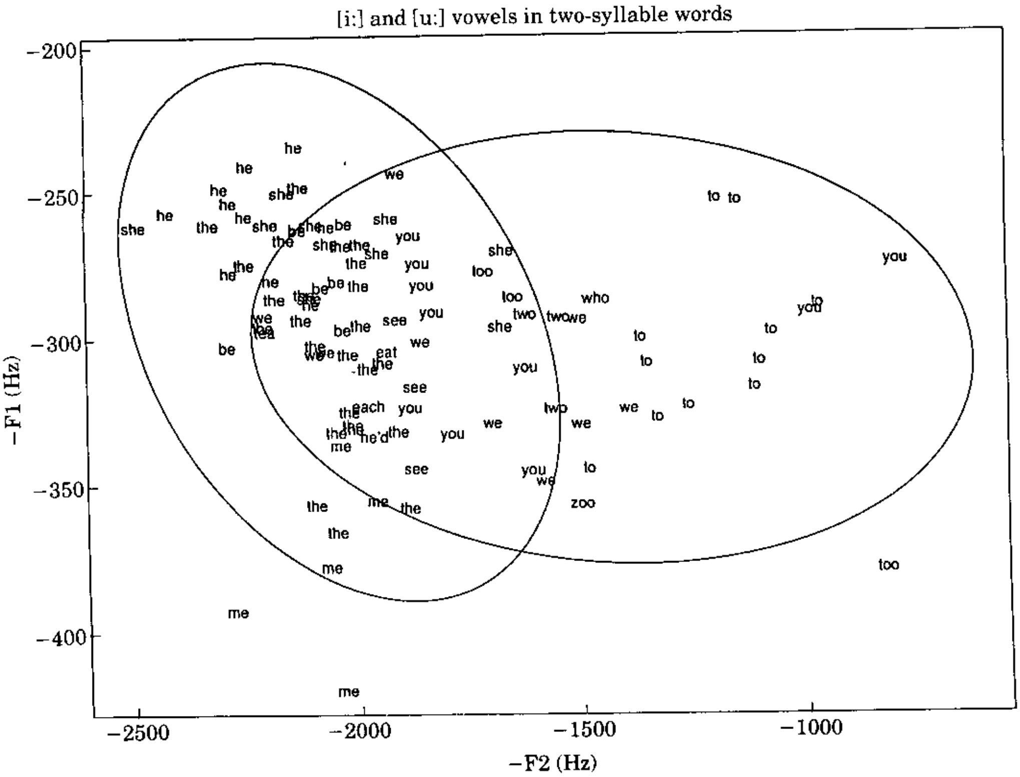


**Figure 11.** Ellipse plots for two vowels at the midpoint showing their corresponding words' orthographic forms.

structed at most levels by numerical manipulation of the boundary times of segment lists rather than by (a more time consuming) manual resegmentation and splicing of the speech data. At the phonetic level, mu+ routines and S-PLUS arithmetic primitives can be used to create cross-spliced stimuli in which, for example, any number of sections of [s] and [S] fricatives are segmented from a following vowel, spliced onto different vowels (Mann & Repp, 1980), randomized, and D/A converted with specified intervals of silence between the stimuli. Similar manipulations can be applied to create sentence-level stimuli of the kind that are used in psycholinguistic experiments, in which words with different levels of sentence stress are spliced out and replaced by other words (e.g. Cutler, 1976).

The digital-signal-processing and multidimensional statistical routines allow speech analyses of varying complexity to be carried out. An example of the kind of analysis that is possible is shown in Figs. 14 and 15. In this study, 300 voiceless fricatives (in citation form words) were divided into a set to be used for "training" and "testing". Cepstrally smoothed spectra and energy values in the first 22 critical bands were derived from 256 point Fourier transforms. The critical band space was transformed to a smaller number of dimensions using canonical discriminant analysis, and a Gaussian classification was carried out on the transformed space to quantify numerically the extent of overlap between the three fricative classes.

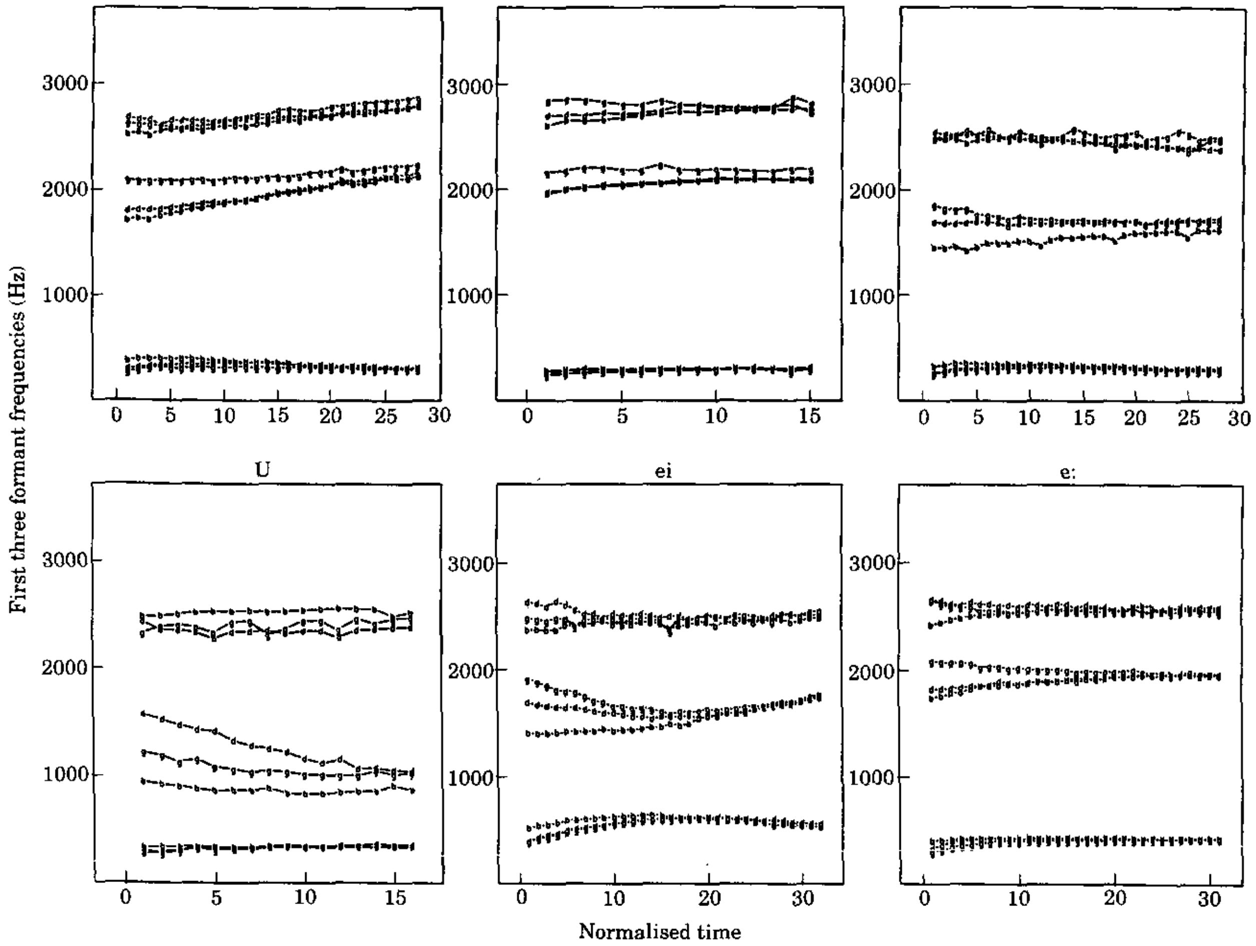


**Figure 12.** Time-normalized F1–F3 displays between the vowel onset and vowel steady-state for [b], [d], [g] in the context of various (following) vowels. Each trace represents an average of roughly 50 time-normalized tokens. Alveolars tend to have a relatively constant F2 onset frequency at 1800 Hz (see also Fig. 13).

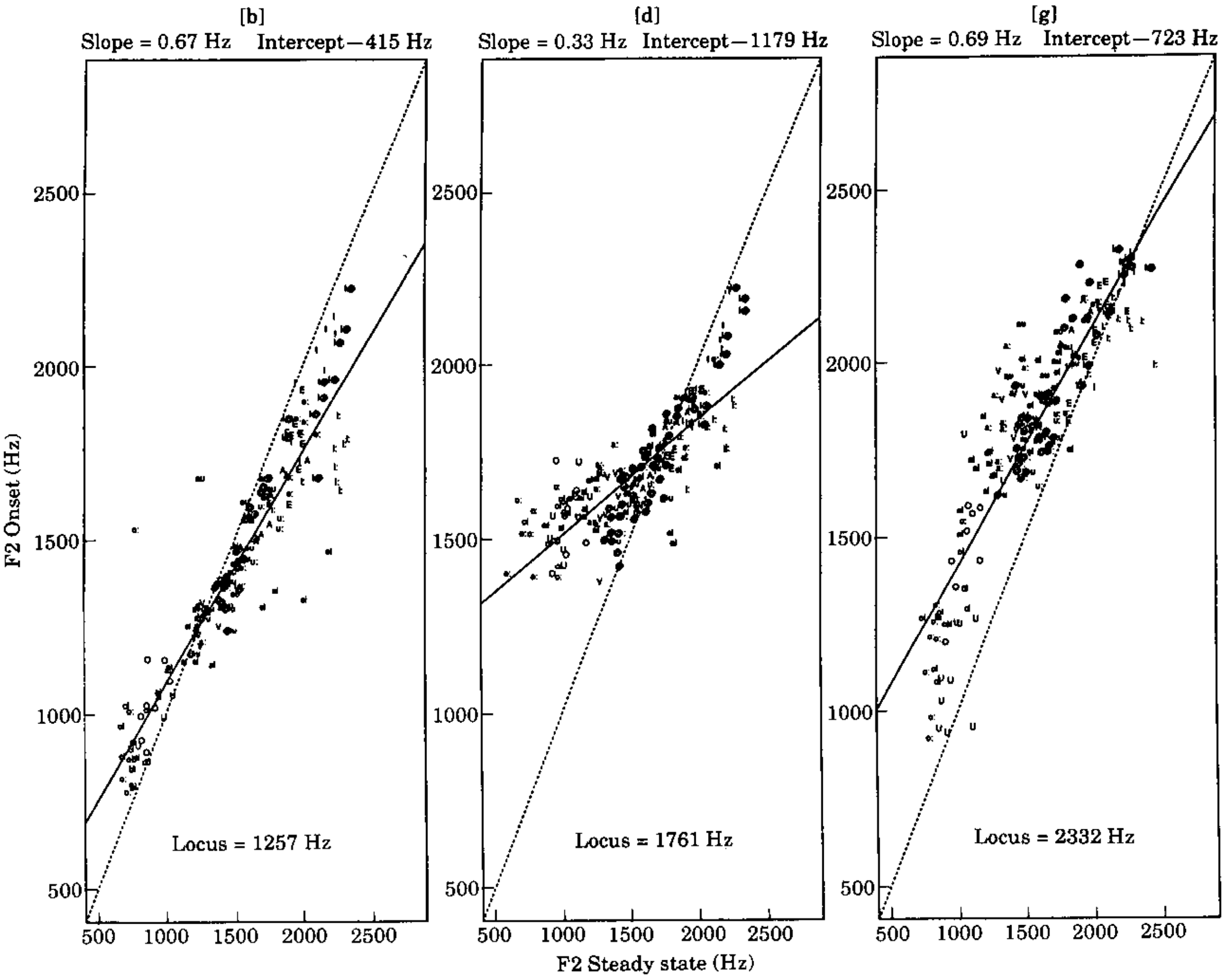


**Figure 13.** The locus is estimated by the intersection of the line $y=x$ with the regression line through the points in the F2 onset/F2 steady-state plane.

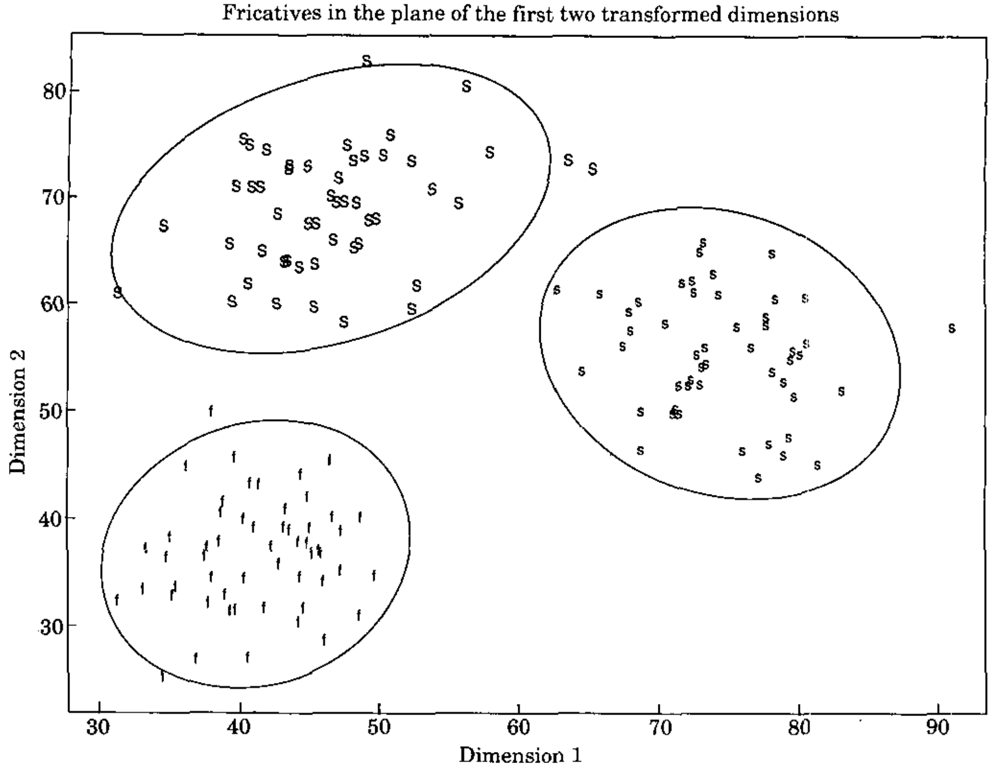


**Figure 14.** Ellipse plots for voiceless fricatives. The dimensions are derived by applying a discriminant analysis to energy values in critical bands at the segment midpoint.

## 7. Current applications and future developments

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 (Fletcher & McVeigh, 1992), and to extract diphones directly from the database. Another application of mu+ has been to monitor the development of the (SHLRC–ANDOSL) speech database itself. 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.

Currently, mu+ is being developed in two ways. Firstly, since most types of label combinations can be extracted and tabulated, a natural extension of the system is to include some of the facilities which are needed to access labels according to grammatical

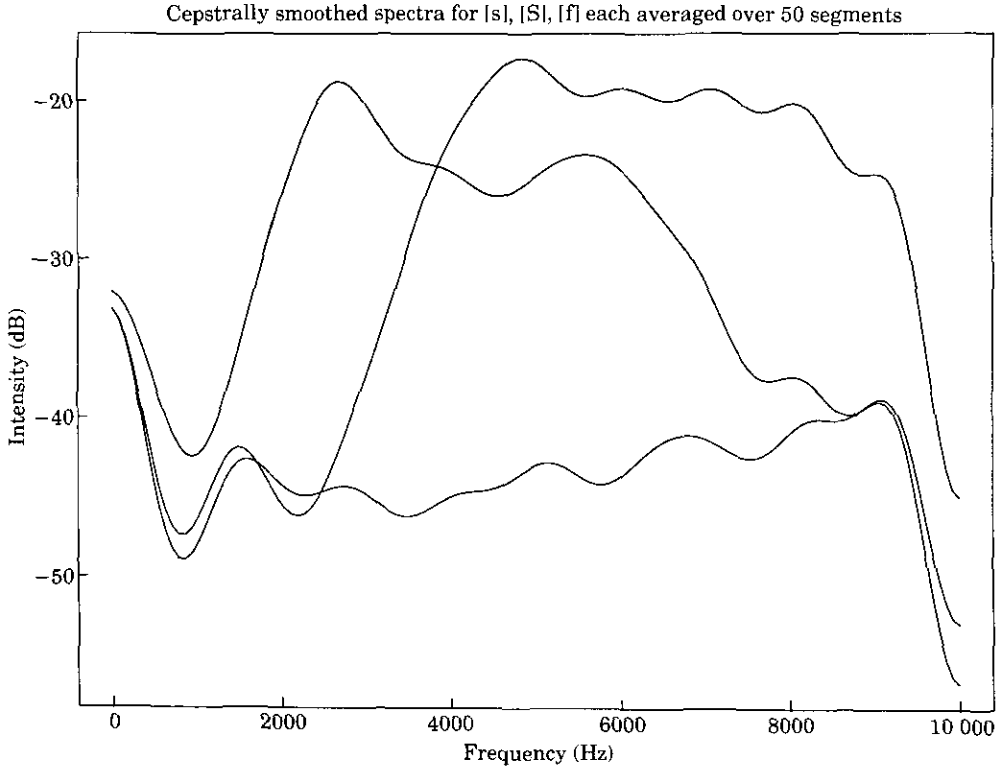


**Figure 15.** Averaged cepstrally smoothed spectra derived from 256-point FFTs at the segment midpoint.

features from a linguistic corpus. Developing a common environment for accessing linguistic corpora and speech databases is important in certain kinds of research which are at the intersection of computational linguistics and speech science, such as generating speech from semantic concepts, and mapping discourse structure onto intonation. The second, recent development concerns interfacing multichannel physiological data (also digitized using Waves+) to mu+ (Fig. 16). Since the physiological data can be extracted in the same way and for the same sets of contexts as for the acoustic speech data (e.g. jaw movement in all phrase-final accented *vs.* unaccented syllables), many investigations in the speech production literature on issues such as coarticulation, vowel undershoot, and speech timing, could also be addressed using mu+ interfaced to a systematically annotated database of speech articulation data (Hardcastle & Marchal, 1990; Marchal & Hardcastle, 1990).

## 8. Conclusions

There is a growing trend for research in speech and language to be corpus based. The main application of corpus based speech research has so far been in speech technology (in particular automatic speech recognition and understanding), but as speech databases

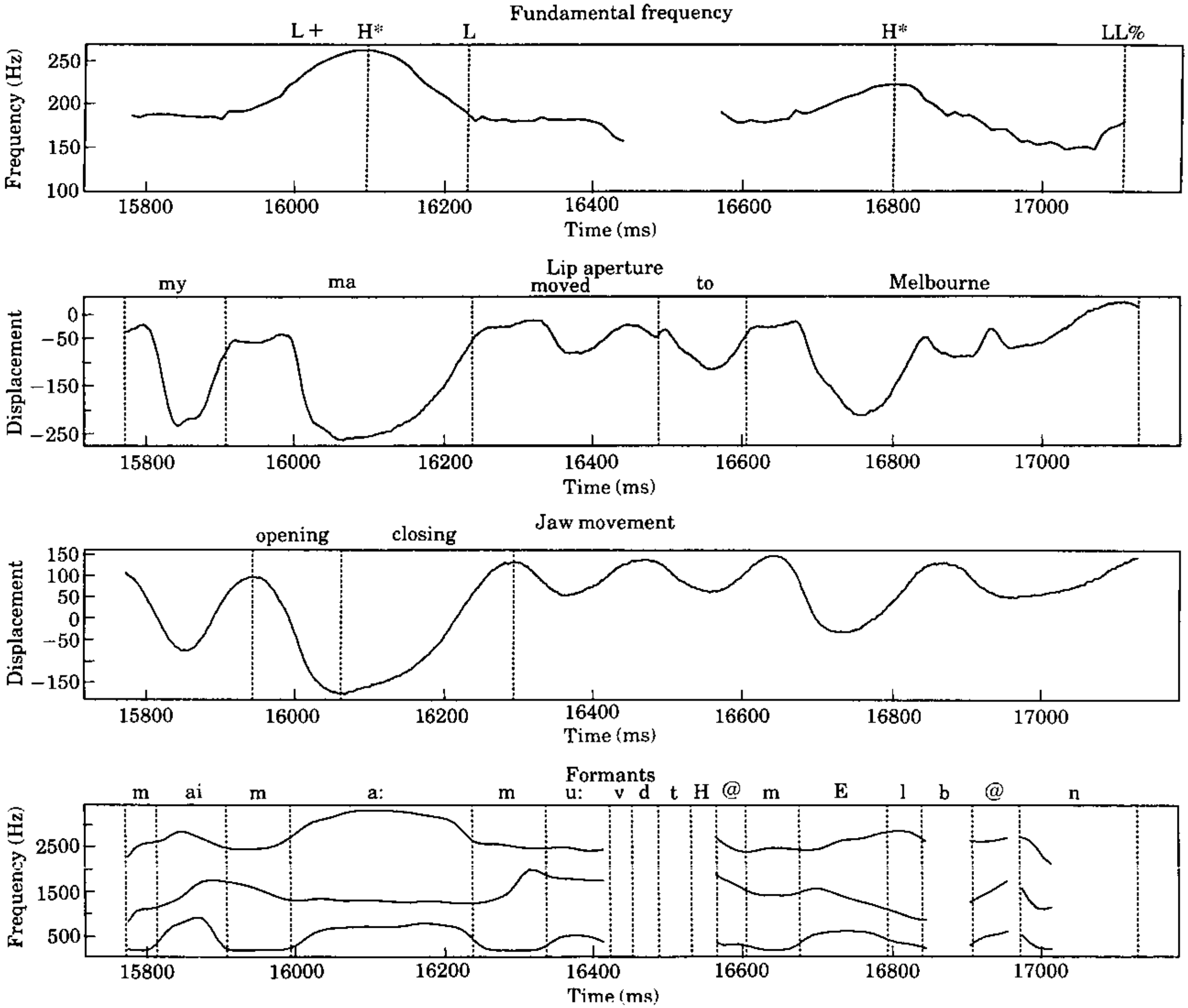


**Figure 16.** Time-aligned acoustic and physiological displays for the sentence "My ma moved to Melbourne" with superimposed intonational, orthographic, gestural, and acoustic phonetic labels. The F0 and formant traces are zeroed when the probability of voicing values are less than 50% and 90% respectively.

become more prevalent, and as the processing power and disk storage capacity of computer systems increase, it is likely that solutions to many of the "classic" problems of speech and language research will be sought by using a comprehensive database as the main resource. However, if such a database is to provide an adequate resource for the many different kinds of groups in the speech and language community, then, leaving aside questions of the database's content, at least two requirements must be fulfilled. Firstly, the database must code information at multiple levels: in addition to an acoustic phonetic segmentation, it must also include labels at various phonological, prosodic, lexical, and linguistic levels. Secondly, these multiple levels must be hierarchically coded in a way which is consistent with models of speech and language, and the hierarchical structure must also be indexed into the utterance's acoustic or articulatory speech signal files with which it is associated.

These two aims are broadly addressed by the mu+ environment which can be used both to create a hierarchical structure for each utterance, and also to extract any label, or set of labels, together with their associated speech signal files from any number of utterances in the database. Additionally, since each utterance and its association to acoustic and articulatory signals is built upon a set of statistical, graphical, and digital-signal-processing primitives, mu+ provides an integrated environment which is potentially relevant to many different kinds of research such as experimental phonetics (e.g. studies of segmental timing, vowel undershoot, coarticulation), speech perception, speech signal processing, statistical modelling of acoustic speech data, speech production modelling, and statistical modelling of the distribution of phonemic, prosodic, or linguistic categories in the database.

## Appendix 1

The machine readable phonetic alphabet for Australian English used in this paper is given below:

### Acoustic Phonetic

| *Oral stop closure* | |
|---|---|
| [p] | pie |
| [t] | tie |
| [k] | cat |
| [b] | bat |
| [d] | do |
| [g] | go |

(all can be followed by [H] denoting the release/frication/aspiration)

| *Nasal stop* | |
|---|---|
| [m] | my |
| [n] | no |
| [N] | sing |

| *Fricative* | |
|---|---|
| [f] | fan |
| [v] | van |
| [T] | think |
| [D] | the |
| [s] | so |
| [z] | zoo |
| [S] | shoe |
| [Z] | measure |
| [h] | he |

| *Affricate* | |
|---|---|
| [t S] | chew |
| [d Z] | judge |

| *Approximant* | |
|---|---|
| [l] | look |
| [r] | run |
| [w] | we |
| [j] | you |

| *Long vowels* | | *Short vowels* | | *Diphthongs* | |
|---|---|---|---|---|---|
| [i:] | heed | [I] | hid | [i@] | here |
| [u:] | who | [U] | hood | [u@] | cure |
| [e:] | there | [E] | head | [ei] | say |
| [@:] | her | [@] | the | [@u] | go |
| [o:] | hoard | [O] | hot | [oi] | toy |
| [a:] | hard | [V] | hut | [au] | loud |
| | | [A] | had | [ai] | sigh |

*Specialized notation*

| | |
|---|---|
| [=n] | syllabic (/@/ precedes as in *sudden*) |
| [l=] | syllabic (/@/ follows as in *explanation*) |
| [i:H] | breathy voice |
| [i:C] | creaky voice |
| [Om] | voicelessness |
| [dH] | tap |

**Broad Phonetic**

uses a subset of the symbols given under Acoustic Phonetic
additionally: /tS/ (choose), /dZ/ (judge)

**Prosodic**

*Syllable, Foot, Accent, Nuclear*

| | |
|---|---|
| S | strong |
| W | weak |
| - | extrametrical |

*Intonation*

| Pitch accents | | Phrase tones | | Boundary tones | |
|---|---|---|---|---|---|
| H* | High | L | low | L% | low |
| L* | Low | H | high | H% | high |
| L+H*, L*+H | Rising | | | | |
| H+L*, H*+L | Falling | | | | |
| !H* | Downstep | | | | |

## References

Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh University Press, Edinburgh.

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

Beckman, M. E. & Edwards, J. (1988). Lengthenings and shortenings and the nature of prosodic constituency. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, (Kingston, J. and Beckman, M. E., eds), pp. 152–178.

Beckman, M. E. & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, **3**, 255–310.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55–60.

Fletcher, J. & McVeigh, A. (1992). Towards a model of segment and syllable duration in Australian English. *Proceedings of the Fourth International Conference on Speech Science and Technology*, 28–33, Brisbane, Australia.

Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology*. Basil Blackwell, Oxford.

Halliday, M. A. K. (1967). *Intonation and Grammar in British English.* Mouton, The Hague.

Hardcastle, W. J. & Marchal, A. (1990). EUR–ACCOR; A multi-lingual articulatory and acoustic database. *Proceedings of the International Conference on Spoken Language Processing (ICSLP–90)*. Kobe, Japan.

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

Hayes, B. (1982). Extrametricality and English stress. *Linguistic Inquiry*, **13**, 227–276.

Hieronymus, J. (1991). Trends in speech and language databases. Paper presented at the workshop on speech databases, Language and Speech conference, Nov. 1991. Melbourne, Australia.

Hieronymus, J., Alexander, H., Bennett, C., Cohen, I., Davies, D., Dalby, J., Laver, J., Barry, W., Fourcin, A. & Wells, J. (1990). Proposed speech segmentation criteria for the SCRIBE project. SCRIBE-Project Report.

Hogg, R. & McCully, C. B. (1987). *Metrical Phonology: a Coursebook*. Cambridge University Press, Cambridge.

Kahn, D. (1980). *Syllable-based Generalisations in English Phonology*. Ph.D. dissertation, MIT. Garland Press, New York.

Lamel, L. F., Kassell, R. H. & Seneff, S. (1986). Speech database development: design and analysis of the acoustic-phonetic corpus. *Proceedings of the DARPA Speech Recognition Workshop*. (Palo Alto, Ca.).

Mann, V. A. & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception and Psychophysics*, **28**, 213–228.

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

Marchal, A. & Hardcastle, W. J. (1990). The relevance of basic research in articulatory phonetics to speech technology. *Proceedings of the Third Australian International Conference on Speech Science and Technology*. Melbourne, Australia.

Millar, J., Dermody, P., Harrington, J. & Vonwiller, J. (1990a). A national cluster of spoken language databases for Australia. *Proceedings of the Third Australian International Conference on Speech Science and Technology*. Melbourne, Australia.

Millar, J., Dermody, P., Harrington, J. & Vonwiller, J. (1990b). A national database of spoken language; concept, design, and implementation. *Proceedings of the International Conference on Spoken Language Processing (ICSLP-90)*. Kobe, Japan.

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

Pierrehumbert, J. B. & Beckman, M. E. (1988). *Japanese Tone Structure*. MIT Press, Cambridge, Ma.

Selkirk, E. O. (1982). The syllable. In *The Structure of Phonological Representations*. Part II. (van der Hulst, H. and Smith, N., eds). Foris Publications, Dordrecht.

Selkirk, E. O. (1988). On the nature of prosodic constituency: comments on Beckman's and Edward's paper. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. (Kingston, J. and Beckman, M. E., eds), pp. 179–200.

Sussman, H. M., McCaffrey, H. A., Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, **90**, 1309–1325.

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