# EXPERIMENTS ON AUTOMATIC PHONETIC SEGMENTATION AND TRANSCRIPTION OF SPEECH

MATTHEW LENNIG

Department of Electrical Engineering
McGill University
Montreal

September 1983

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering

Automatic phonetic segmentation and transcription of speech

# ABSTRACT

Understanding the structure of speech variability is necessary to advance the technology of verbal man-machine communication. However, empirical studies of variability require large, phonetically labelled speech databases. To allow automatic generation of such databases, a technique is investigated for automatic phonetic segmentation and labelling of speech, assuming its orthographic transcription is given.

A synthetic reference is first created from the given transcription. Phonetic segmentation and labelling information for the synthetic utterance is known. By using dynamic time warping to align the synthetic and natural utterances, segmentation and labelling information is mapped onto the natural utterance.

Automatically generated segmentations are compared with manual segmentations for twenty test sentences. Boundary location error rate is found to be 45 percent (N=659). Although the method gives good global alignment, it does not reliably align short-time acoustic events. The results reflect limitations in both the synthetic speech quality and the dynamic time warping alignment procedure.

# RESUME

Afin de faire avancer les techniques de communication verbale entre l´homme et l´ordinateur, il est important de comprendre la nature de la variabilité de la parole. Les études empiriques sur la variabilité exigent de grandes bases de données phonétiquement étiquetées. L´objet de la présente étude est d´évaluer une technique automatique de segmentation et d´étiquetage phonétique de la parole pour des énoncés dont la transcription orthographique est connue.

On crée d´abord un modèle de référence en utilisant un système de synthèse qui convertit le texte orthographique en discours sonore. Ce modèle de référence synthétique étant déjà doté d´une segmentation et d´un étiquetage phonétique, il s´agit alors de transférer ces informations à l´énoncé naturel en utilisant un algorithm d´anamorphose temporelle pour aligner les deux énoncés.

On a comparé, pour vingt phrases, les segmentations automatiques ainsi produites à des segmentations manuelles. Le taux d´erreur de positionnement des frontières est de 45 pourcent (N = 659), et varie selon le type de frontière. Bien que l´alignement global soit bon, la méthode présente une faiblesse au niveau des évenements acoustiques de courte durée.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1: SCOPE OF THE PROBLEM

Automatic segmentation of natural speech is a complex problem similar in many respects to other segmentation problems encountered in artificial intelligence. The signal to be segmented is the surface representation of an underlying complex process whose detailed structure is not directly accessible. It is the difficulty of the general segmentation problem that has forced most currently employed continuous speech recognition technques to replace simplistic segment-recognize recognition paradigms by hierarchic sequences of hypothesize and test operations. The segmentation problem as formulated here probes the adequacy of using our limited knowledge of speech generation to assist the analysis that may be required for recognition.

Solving the problem of automatic phonemic transcription of speech would solve the problem of automatic speech recognition. Phonemes are defined as the abstract segments representing minimal lexical distinctions. Unfortunately, the mapping from phoneme string to acoustic signal is highly variable, making the phoneme decoding problem extremely difficult. The purpose of this study is to develop a tool for studying the structure inherent in this mapping, both within a single speaker and across speaker populations. It is expected that a more complete understanding of the structure of speech variability will facilitate a solution to the speech recognition problem.

Phoneme boundaries are not well defined in the speech signal. Phonetic features associated with underlying phonemes are identifiable, but those features tend to spread out, coexisting with the features of adjacent phonemes. Nonetheless, distinct segments can be observed in sound spectrograms of speech. I will refer to these as ´phonetic segments´. Phonetic segments correspond to Fant´s (1973:21-23) sound segments, "the boundaries of which are defined by relative distinct changes in the speech wave structure...These boundaries are related to switching events in the speech production mechanism such as a shift in the primary sound source, e.g., from voice to noise, or the opening or closing off of a passage within the vocal cavities, the lateral and nasal pathways included. Less distinct sound boundaries may be defined from typical changes in the pattern of formant frequencies."

The purpose of the proposed speech analysis tool is to segment speech automatically into phonetic segments, given the underlying lexical string corresponding to the signal. The proposed tool would also label each phonetic segment using labels from a finite set of ´phones´, or phonetic segment classes. A reliable analysis tool of this sort would facilitate large scale statistical studies of speech variation and possibly lead to improved models of speech variation.

## 1.1 Approaches to Automatic Segmentation and Labelling of Speech

The automatic segmentation and labelling method investigated relies on a combination of speech synthesis from text and dynamic time warping. A synthetic utterance corresponding to the given word string is first synthesized using a text-to-speech system. The synthetic speech is then temporally aligned with the natural utterance using dynamic time warping. Assuming the segmentation and labelling of the synthetic utterance are known, they may be mapped onto the natural utterance using the temporal alignment obtained by dynamic time warping, thus inducing a segmentation and labelling of the natural utterance. This technique is termed ´synthesize-and-warp´.

Many other techniques have been used to segment and label speech automatically. Since most of these were intended for use in speech recognition systems, they do not assume the underlying string of words is known. Goldberg (1975) investigates a class of phone-level segmentation and labelling algorithms which place segmentation boundaries according to an acoustic change function. Labels are then assigned by computing distances from the segment to a set of segment templates. Gill et al. (1978) describe the ZAPDASH segmentation algorithm used in the Harpy speech recognition system. ZAPDASH uses a recursive strategy, employing time-domain acoustic parameters. It segments into the following segment types: silence, unvoiced fricative, aspiration, low amplitude voiced segment, maximum of the peak to peak differences of smoothed signal, minimum of the peak to peak differences of the smoothed signal. Mermelstein (1975a) proposes a sequential, rule-based approach to phonetic segmentation and labelling, using acoustic cue detectors. Cohen (1981) segments speech using Markov modelling, where a spectral similarity measure together with statistics on segment duration are used to find the maximum liklihood set of boundary positions.

Recently, a few experiments have been performed in which the underlying word string or phoneme string of speech to be segmented is known. Sargent and Malcolm (1979) and Sargent (1982) describe an aid for the deaf which aligns orthographic syllables with speech syllables. The method is similar to that of Mermelstein (1975b) in that it uses the energy contour to perform a syllable segmentation. However, Sargent also uses information about the expected degree of energy dip between syllable boundaries of different kinds and about expected voicing during those dips. For example, he notes that vowel-liquid-vowel syllable boundaries have a 0.3 probability of no energy dip, a 0.5 probability of a dip of less than 10 dB during which voicing is maintained, a 0.1 probability of a larger than 10 dB dip during which voicing is maintained, etc.

Another system for segmenting a known utterance is that of Wagner (1981). Wagner´s algorithm first uses the parameters of energy, voicing, and fundamental frequency to segment the utterance into voiced, unvoiced, and silent segments. Dynamic programming is then used to align substrings of these labels with the given string of phonetic labels. Néel et al. (1983) first locate "stable instants" in the speech signal and then use a dynamic programming algorithm to align these with the known phoneme string. Each phoneme is represented by one or more templates, each consisting of a single spectrum. The city block (first order Minkovski) distance determines the distance between a spectral template and a "stable instant."

Segmentation and labelling systems designed to work on a known utterance can be classified into two groups: those which use a variant of the top-down synthesize-and-warp technique employed in this study (Bridle and Chamberlain, 1983; Lennig, 1983; and Le Saint-Milon and Stella, 1983) and those which use partially bottom-up techniques (Sargent and Malcolm, 1979; Wagner, 1981; Néel et al., 1983). Applications cited for segmentation and labelling of a known utterance include aids for the handicapped (Sargent), training a recognition system (Néel et al.), creating a dictionary for diphone synthesis (Le Saint-Milon and Stella), improving quality of synthesized speech (Bridle and Chamberlain), and generation of a phonetically labelled database (Wagner, Lennig). The cited studies on segmentation of unknown utterances were directed toward speech recognition.

## 1.2 Assumptions Underlying the Synthesize-and-Warp Technique

The appeal of the synthesize-and-warp technique is that it makes use of knowledge already encoded in the text-to-speech system about the relationship between the underlying phoneme string and the acoustic signal, requiring no additional source of knowledge. Thus, to the extent that the synthesized speech is an acceptable production of the underlying phoneme string, the differences between synthetic and natural productions are comparable to the differences between natural productions by different speakers. Furthermore, alignment techniques found useful to line up different natural productions should also suffice to align synthetic productions to natural productions. However, the following crucial assumptions must hold if the synthesize-and-warp technique is to produce a correct segmentation and labelling:

1. The phonemic string underlying the natural utterance corresponds to that employed by the synthesizer.

2. The synthesizer reproduces the acoustic details of fluent speech sufficiently accurately to be used as a model in the synthesize-and-warp procedure.

3. The segmentation of the synthetic speech is correct and reflects the desired level of analysis.

4. The local distance measure used in the time warping algorithm to calculate the dissimilarity between speech frames is relatively sensitive to phonetic differences while being relatively insensitive to interspeaker differences.

5. The dynamic time warping algorithm adequately compensates for durational differences between synthetic and natural segments.

The first assumption made by the synthesize-and-warp technique is that the phonemic string underlying the natural utterance corresponds with that used by the synthesizer. This is only roughly the case. The synthesizer always employs the same lexicon, letter-to-sound rules, and phonological rules, whereas natural utterances produced by various speakers differ considerably in dialect, style, and emphasis. Bridle and Chamberlain (1983) bypass this source of variability, using hand specification of the phonemic input to the synthesizer, resulting in enhanced alignment. Because our goal is to develop an automatic system capable of working from text input, no hand specification is performed in the present study, yielding poorer, but more representative, results.

The second assumption is that the synthesized speech be realistic enough to serve as a reliable model against which to time warp the natural utterance. Although the gross acoustic features of speech are modelled by the text-to-speech synthesizer, many less salient features of fluent speech are not modelled at all. It is immediately obvious when listening to the synthetic speech that it was not produced by a human speaker. Lack of accurate reproduction of some aspects of fluent speech may adversely affect the details of the warp alignment with the natural utterance. Since the precise alignment is crucial to obtaining a correct segmentation and labelling of the natural utterance, unnaturalness of the synthetic model is a potential source of error in the synthesize-and-warp method.

The third assumption is that the implicit segmentation used to generate the synthetic speech is correct. This is not necessarily the case, since the requirements of the text-to-speech system differ from those of an analysis system. A finer level of segmentation, corresponding to phonetic segments, may be desirable for the analysis system, whereas the synthesizer may use phoneme-size segments. Boundaries between the synthesizer´s segments tend to be defined arbitrarily, while for the analysis system, it is desirable for the user to be able to specify boundary definitions. A solution to this problem is proposed in Chapt. 3.

A fourth assumption is that the local distance measure, used to calculate the dissimilarity of speech frames, is relatively sensitive to phonetic (interphone) differences while relatively insensitive to interspeaker differences. If this assumption does not hold, distances between similar phones in the natural and synthetic utterances may be larger than distances between different phones in the natural utterance. This would lead to misalignment in the time warping algorithm. The distance measure used in this study is known to be sensitive to interspeaker differences, especially for vowel-like sounds. The inability of the distance measure to distinguish between phonetically relevant and phonetically irrelevant differences is another source of alignment error.

Even if the synthetic phoneme string is correct, phonetic segments of the natural speech will certainly differ in duration from those of the synthetic model. The fifth assumption of the synthesize-and-warp method is that the dynamic time warping algorithm will adequately compensate for durational differences between synthetic and natural segments. The dynamic time warping algorithm utilizes minimum and maximum slope constraints and slope penalties to prevent excessive time dilation or compression, while still allowing needed temporal adjustments. However, since constraints and penalties are applied uniformly across the time axis of the speech, they ignore variability in the elasticity of speech (Kozhevnikov and Chistovich, 1965), allowing excessive elasticity for certain events while not enough for others.

The success of the synthesize-and-warp approach is therefore predicated on at least rough satisfaction of the above assumptions. The investigations focus on segmentation performance based on these requirements and in turn shed light on the extent that improvements to speech synthesis and comparison techniques are required before acceptable segmentation can be attained.

## 1.3 Overview of the Present Study

Chapter 2 is a discussion of preliminary experiments with the synthesize-and-warp method. The segment inventory and boundaries used were those defined by the text-to-speech system. This segment inventory and set of ad hoc boundary definitions, although suitable for the synthesizer, were determined to be unsatisfactory for the generation of a phonetically labelled database, motivating the development of a rule-based segmenter for the synthetic speech. The rule-based segmenter, described in detail in Chapt. 3, segments the parameter stream used to drive the terminal analog synthesizer. Because of the regularity of the synthetic speech and errorless estimates of the acoustic parameters, the rule-based segmenter is able to achieve precise segmentations of the model synthetic utterances. The phone inventory is modified to suit the needs of a phonetically labelled database.

Productions of a sentence from four speakers are warped against the rule-segmented synthetic model in Chapt. 4. This is compared with the use of a hand-segmented natural model. The natural model, while giving somewhat better segmentation performance than the synthetic model, still gave rise to a significant number of segmentation errors. This leads to the conclusion that the contradiction of Assumption 2 (that the synthesizer produces sufficiently natural speech) is not the major source of error in the synthesize-and-warp procedure. Chapter 4 also addresses the problem of how to evaluate the correctness of a segmentation, where the required precision in boundary placement depends on the nature of the boundary itself.

Chapter 5 describes the main experiment. The experiment consists of applying the rule-based segmenter to ten synthetic sentences warp-aligning the synthetic utterances with natural productions by two speakers. An analysis of the results indicates that brief segments are less likely to be correctly segmented than longer segments and that boundaries involving segments that are acoustically different are more likely to be correctly segmented than a boundary between acoustically similar phones.

# CHAPTER 2: PRELIMINARY EXPERIMENTS USING MITALK SEGMENTATION

The alignment method used in this study requires a segmented and labelled model utterance. The utterance for which a segmentation and labelling is desired (the test utterance) is time-warped against the model utterance. Segment boundaries and labels are then mapped from the model utterance onto the test utterance, across the warp path, inducing a segmentation on the test utterance.

In Chapt. 1, five critical assumptions were set forth. Two of these, Assumptions 2 and 3, are specifically aimed at the quality of the model: The synthetic speech must be of sufficient quality and the segmentation and labelling must be correct. Two more, Assumptions 4 and 5, deal with the fidelity of the alignment procedure, specifically, the sensitivity of the distance metric and the flexibility of the dynamic time warping algorithm. Assumption 1 addresses the issue of segmental congruence between the model and unknown. In this chapter, we examine the consequences of naively of accepting the five critical assumptions of Chapt. 1.

This chapter reports a preliminary experiment employing model utterances synthesized using the MITalk-79 text-to-speech system [1].

[1] The MITalk-79 text-to-speech system is used with permission of MIT. We have made minor modifications to the original system, in particular, by using a polynomial pulse (Rosenberg, 1971) in place of the original filtered pulse train.

Appendix A gives the set of ten phonetically balanced sentences (IEEE, 1969) which were used. MITalk-79 takes standard English orthographic input produces a synthetic speech signal. The first step that MITalk performs is to standardize the text format, spelling out numbers and abbreviations. The next few steps convert English orthography to a phonetic representation expressed in a computer-readable phonetic alphabet (Appendix B) similar to the ARPABET (Shoup, 1980). Subsequent steps assign to each phonetic segment a duration and one or two fundamental frequency targets. A phonetics module uses these segmental and suprasegmental data to generate one frame of acoustic parameters every 5 ms. The parameters control amplitudes of various types of excitation, fundamental frequency, formant frequencies, bandwidths, etc., of a series/parallel terminal analog synthesizer (Klatt, 1980).

The existence of phone durations as input to the phonetics module implies a nominal segmentation of the synthetic speech. The phonetics module actually uses these nominal boundaries to provide anchor points around which to smooth the acoustic parameters. Each phoneme has a set of target parameter values, which are affected by its phonological environment. Different smoothing rules are used depending on the nature of the segments forming the boundary. Therefore, the segment boundary locations are not absolute in any sense, but are only meaningful in the context of the specific targets and smoothing rules used in the phonetics module.

## 2.1 Nature of the Segmentation Used in MITalk

Because of its easy availability, the segmentation implicit in the MITalk text-to-speech system was used in an initial set of experiments to test the automatic alignment procedure. In this section, we examine the properties of this segmentation by looking at spectrograms of synthetic speech showing locations of MITalk's implicit phone boundaries.

Figures 2-1 through 2-4 are spectrograms of synthetic productions of sentences 1 through 4, showing the location of MITalk's implicit segment boundaries. Segments are labelled using MITalk's phonetic symbols (Appendix B). The time axis (abscissa) of each spectrogram is labelled in units of 5 milliseconds and the frequency axis (ordinate) in Hertz. In general, the segmentation seems to be a reasonable, phoneme-level segmentation. Some details of the segmentation are as follows:

(i) Stop bursts are sometimes considered part of the stop, as, for example, in the [G] of goose and in the [K] of market, of sentence 1, but sometimes partially belong to the stop and partially to the vowel, as for example in [T AH] of stubborn in Sentence 3.

(ii) Boundaries between vowels and fricatives are somewhat indeterminate: in the [UW S] of goose (Fig. 2-1a) the boundary could just as justifiably be located a few frames to the left in the period where formants and frication noise coexist. In [UW S] MITalk places the boundary at the end of formant excitation, while in the [AH Z] of was the boundary is near the onset of frication.

(iii) In the [T S] sequence in brought straight (Fig. 2-1b) the boundary presumably separates the burst of [T] from the frication of [S]. Since these are not acoustically distinguishable, the location of this boundary reflects only a convention of the synthesis system and is meaningless from a speech analysis point of view.

(iv) Liquid-vowel and glide-vowel boundaries are placed in such a way that most all of the transitional portion is identified with the vowel. Examples are [W AH] in was, [R AO] in brought, and [R EY] in straight,

(v) An indeterminacy exists at vowel-vowel boundaries, such as [IY OW] in the old (Fig. 2-1c). In this example, the boundary is placed so that the entire transition is included in the [OW] segment.

Similar phenomena are observed in the MITalk segmentations of the other sentences.

## 2.2 Segmentation of Natural Speech: Preliminary Experiment

Sentences 1 through 10 were spoken by the author (speaker ML) and time-aligned with their synthetic counterparts using a symmetric, unconstrained, decimated-grid dynamic time warping algorithm (Hunt, Lennig, and Mermelstein, 1983). Cost penalties were imposed for vertical and horizontal transitions of one half the destination local distance. The resulting warp path was used to map the MITalk segmentation of the synthetic utterance onto the natural utterance, inducing a segmentation of the latter. Induced segmentations in natural sentences 1 through 10 are shown in Figs. 2-5 through 2-14. As in the previous figures, the time axis (abscissa) of each spectrogram is labelled in units of 5 milliseconds and the frequency axis in Hertz.

One of the difficulties encountered in this preliminary experiment was the evaluation of segmentation results. Examination of Figs. 2-5 through 2-14 indicates that while most segmentation boundaries can be easily classified as correct or incorrect, there also exist a large number whose status is questionable. Some boundaries are obviously misplaced and can be readily identified as gross errors; others are almost right, but not exactly where this author would have placed them. Most boundaries appear correct. In Fig. 2-5a, for example, the [G] of goose is segmented correctly. As in the model utterance (Fig. 2-1a), the burst is considered as part of the [G]. Vowel transitions, on the other hand, are considered part of the following vowel, as in the model. The [W AH] boundary in was occurs slightly earlier than we would have preferred: The automatic segmentation indicates a completely voiceless [W]. This is the kind of minor error that is difficult to evaluate: Should the [W AH] boundary be considered correct or incorrect?

Certain types of errors are serious enough that there can be no doubt that a segmentation is wrong. This type of gross error occurs sentence 2 (Fig. 2-6b) in the [NG IH N WH IH] of thing in which. Here the [NG] is extremely short and [IH] begins where [NG] should. From about halfway through the actual [NG] to the end of [NG], the [N] label is attributed. The [WH] label is incorrectly applied to the actual [IH] of which. The [IH] label of which is incorrectly applied to the actual sequence [N WH IH]. The alignment appears to have become desychronized, causing gross errors in several consecutive segments. At [CH] of *which* the segmenter regains synchronization, although it still places the [SH W] boundary of *which we* too far to the left. (Note that [SH] is the fricative part of the final affricate of the word *which*.)

## 2.3 Segmentation Errors

Gross errors involving desynchronization occurred in the following sentences:

| | | | |
|---|---|---|---|
| Sent. 2: | [NG IH N WH IH] | thing in which | (Fig. 2-6b) |
| Sent. 3: | [AX WH IH F AX V] | a whiff of | (Fig. 2-7a) |
| Sent. 6: | [ER SIL W AX Z] | cruiser was | (Fig. 2-10b,c) |
| Sent. 8: | [DH AX L AO T K] | the lost cause | (Fig. 2-12c,d) |
| Sent. 10: | [N DH IH] | on this | (Fig. 2-14b) |

The frequent occurrence of desynchronization errors is cause for some concern. However, we will see in Chapt. 4 that many gross errors of type can be eliminated by the use of local slope constraints on the warp path.

Several other minor errors occurred, affecting only individual segment boundaries. Boundary errors other than desynchronization errors are listed below. Doubtful cases have been omitted so that only clear errors are included below:

| | | | | |
|---|---|---|---|---|
| Sent. 2: | [S IH] | sink | slightly early | (Fig. 2-6a) |
| | [SH W] | which we | much too early | (Fig. 2-6b) |
| Sent. 3: | [KP Y] | cure | late | (Fig. 2-7b) |
| | [DH AX] | the | early | (Fig. 2-7b) |
| | [M OW] | most | slightly early | (Fig. 2-7b) |
| | [T S] | most stubborn | early | (Fig. 2-7c) |
| | [ER N] | stubborn | early | (Fig. 2-7c) |
| Sent. 4: | [AX F] | the facts | late | (Fig. 2-8a) |
| | [F AE] | facts | much too late | (Fig. 2-8a) |
| | [T S] | facts | late | (Fig. 2-8a) |
| Sent. 5: | [D Z] | parades | late | (Fig. 2-9d) |
| Sent. 6: | [AX F] | the fleet | late | (Fig. 2-10d) |
| | [F L] | fleet | early | (Fig. 2-10d) |
| Sent. 7: | [F T] | left | late | (Fig. 2-11c) |
| Sent. 9: | [P R] | spring | late | (Fig. 2-13c) |
| | [IH NG] | spring | late | (Fig. 2-13c,d) |
| | [R AE] | grass | early | (Fig. 2-13d) |
| Sent.10: | [P OW] | post | early | (Fig. 2-14a) |

In the above list of errors, I have only listed boundaries which I consider to be indisputably incorrect: doubtful cases have been omitted. For example, the [T S] boundary of brought straight in sentence 1 appears to be too late: the algorithm has considered a portion of the [S] as the stop burst, but that portion appears too long. Since we have not defined any precise criteria for judging whether or not this transcription is correct, we have not listed it above as an error.

Some errors are much larger, in terms of number of frames, than others. For example, the [S IH] boundary of sink (sentence 2) is only about two frames too early. The reason we are able to list this as an error is that the [S IH] boundary is sharply defined by the onset of voicing. An example of a segmentation which we were tempted to classify as incorrect but did not is the [R AY] boundary of right in sentence 4. It seems about five frames too early. However, unless we define more precisely where liquid-vowel boundaries should be placed, we cannot justify classifying this as an error.

## 2.4 Discussion of the Preliminary Experiment

Several conclusions were drawn from the preliminary experiment described above which guided the subsequent course of this work. The discussion is divided into two sections. The first section discusses the segmentation of the synthetic model sentences themselves. The second section focusses on the segmentation induced on the natural sentences.

### 2.4.1 Implicit segmentation of the synthetic model

Examining the implicit segmentation of the synthetic sentences (Figs. 2-1 through 2-4), it was noted that the synthesizer's analysis, although plausible, does not necessarily reflect the level of segmentation or the segmentation conventions required for a particular kind of phonetic analysis. For example, in words like market (Fig. 2-1d), [AXR] is considered to be a single segment. Certain kinds of analyses may wish to separate this sequence into two segments. Similarly, for certain applications, it may be desirable to consider diphthongs as two segments rather than one. Our first conclusion is, therefore, that in order to employ the segmenter for a wide variety of tasks, the user must be able to control the level of segmentation. This is not possible if the synthesizer's segmentation is adopted directly. In other words, Assumption 3 (Chapt. 1) may not hold: the model's segmentation does not necessarily reflect the desired level of analysis.

Boundary positions in the synthesizer's segmentation are defined not by theoretical considerations of speech analysis but by the need to provide a convenient framework for the generation of acoustic parameters. This leads to a lack of consistency in the placement of certain segmentation boundaries from the point of view of phonetic analysis. For example, while the stop burst is normally considered part of the stop, as in [G] of goose (Fig. 2-1a), it is occasionally segmented as if it were part of the following segment, e.g., [B] of brought (Fig. 2-1b). The placement of certain boundaries in the synthetic model, although not incorrect by any objective criterion, seemed arbitrarily skewed. This was particulary noticable between sonorant segments (see items (iv) and (v) in Section 2.1).

Since, from the point of view of an analysis system, the model itself is sometimes incorrect in its segmentation, one cannot expect consistency in the segmentation of natural speech. In other words, Assumption 3 of Chapt. 1 does not hold. Chapt. 3 describes a rule-based system for the segmentation of the synthetic model. This system is found to produce a segmentation for which Assumption 3 holds. In addition, the proposed rule-based system allows the experimenter to specify segment boundary definitions and to specify the level of segmentation desired.

### 2.4.2 Induced segmentation of the natural utterance

An important insight gained in this preliminary experiment from the evaluation of automatic segmentation of natural utterances is an appreciation of the difficulty of such an evaluation. As noted above in Section 2.2, ambiguities, indeterminacies, and inconsistencies make reliable evaluation problematic. Quantitative scoring is necessary in order to choose among various segmentation procedures. However, as noted in Section 2.3, one segmentation error of two or three frames may be more serious from a phonetic analysis point of view than another segmentation error of five or ten frames. For example, if an automatically segmented database is used to study the spectra of stop bursts, which have a durations of the order of ten milliseconds, then a boundary placement error of the same magnitude would severly distort the results. On the other hand, an error of ten milliseconds in the position of a boundary between a liquid and a vowel is benign since the transition between these segments is continuous and relatively long. These considerations argue against a quantitative error criterion based purely on deviation from a given norm. In Chapt. 4, we propose a solution to the evaluation problem.

Although the majority of the segment boundaries are correct, the segmentation results are insufficiently reliable: Nine out of ten sentences contain at least one segmentation error. The errors are attributable to the fact that none of the five critical assumptions discussed in Chapt. 1 holds completely.

# CHAPTER 3: THE RULE-BASED SEGMENTER

One source of segmentation error in the preliminary experiment described in Chapt. 2 was the model segmentation itself: The implicit segmentation used for synthesis does not satisfy Assumption 3. In this chapter, a rule-based system is described for segmenting the synthetic model. The rule-based segmenter allows the user to define the rules used for segmenting the synthetic model. These rules are defined in terms of the stream of acoustic parameter frames which drive the terminal analog synthesizer. The resulting segmentation satisfies Assumption 3.

A disadvantage of the automatic segmentation system used for the preliminary experiment described in Chapt. 2 is that the level of segmentation produced by the system cannot be controlled by the user. As described in Section 2.4.1, the user has no choice but to consider vowel+[r] sequences, diphthongs, and stops as indivisible segments. One of the purposes of the rule-based segmenter described in this chapter is to free the user of boundary definition constraints imposed by the text-to-speech synthesis system, allowing him to define a level of segmentation appropriate to his purpose.

Of course, it would be preferable to apply the rule-based segmenter directly to the natural speech. This would eliminate the additional computation and error associated with the dynamic time warping step. However, designing a rule-based segmenter to perform successfully on natural speech is a significantly more difficult task. Three factors account for the increased difficulty. First, estimating acoustic parameters from the natural speech upon which to base segmentation rules is an inherently errorful procedure, whereas in the case of synthetic speech exact parameter values are obtained from an intermediate step in the synthesis algorithm. Secondly, many secondary cues are absent from synthetic speech which makes it easier to segment based on primary cues. Finally, although synthetic speech models phonologically conditioned variability it does so in a deterministic manner; thus, for the same input sentence, it always gives the same output. Nevertheless, the segmentation of synthetic speech provides useful insights for the direct rule-based segmentation of natural speech. Direct rule-based segmentation will be the focus of future work.

This chapter is organized in five parts. In Sect. 3.1, we define a notational device: a computer phonetic alphabet more flexible than the widely used ARPABET. In Sect. 3.2 we describe the functional structure of the rule-based segmenter and its various components. Section 3.3 describes the set of rules we have implemented, including the level of segmentation chosen and the segmentation rules themselves. Section 3.4 gives results of applying the rule-based segmenter to a set of synthetic model sentences. Finally, Sect. 3.5 discusses rule-based segmentation from a general point of view and speculates as to its applicability to natural speech.

## 3.1 Machine-readible phonetic alphabet

We have devised a computer-readable phonetic alphabet (CPA) that is designed to resemble the International Phonetic Alphabet (IPA) as closely as possible. CPA can be used with consistency for both English and French transcription. With certain extensions, CPA can be adapted for use with other languages. Table 3-1 lists the CPA symbols used in this study, together with their IPA equivalents and keywords illustrating their use in English. For a complete list, including the additional symbols necessary for transcribing French, see Lennig and Brassard (in preparation).

## 3.2 Functional Structure of the Rule-Based Segmenter

The rule-based segmenter consists of four components:

I. Determination of segment label sequence

II. Application of phonological rules

III. Determination of locally determinable boundaries

IV. Determination of other boundaries.

These four components are applied in sequence to the synthetic utterance to produce a segmentation of it.

TABLE 3-1. CPA symbols used in this study.

| CPA | keyword | IPA | CPA | KEYWORD | IPA |
|---|---|---|---|---|---|
| [i] | cream | [i] | [n] | nip | [n] |
| [I] | bit | [ɪ] | [g~] | sing | [ŋ] |
| [e] | bait | [e] | [f] | foe | [f] |
| [E] | bet | [ɛ] | [v] | very | [v] |
| [@] | bat | [æ] | [T] | thin | [θ] |
| [A] | father | [ɑ] | [D] | they | [ð] |
| [^] | but | [ʌ] | [s] | sit | [s] |
| [u] | boot | [u] | [z] | zip | [z] |
| [U] | foot | [ʊ] | [S] | chute | [ʃ] |
| [o] | boat | [o] | [Z] | vision | [ʒ] |
| [O] | caught | [ɔ] | [h] | hit | [h] |
| [*] | synthesize | [ə] | [p] | pit | [p] |
| [aj] | by | [ɑj] | [b] | bond | [b] |
| [aw] | cow | [ɑw] | [t] | tea | [t] |
| [Oj] | boy | [ɔj] | [d] | dip | [d] |
| [j] | yank | [j] | [k] | cake | [k] |
| [w] | wick | [w] | [g] | give | [g] |
| [hw] | which | [hw] | [tS] | cheek | [tʃ] |
| [l] | lap | [l] | [dZ] | jeep | [dʒ] |
| [r] | rap | [r] | [_] | (silence) | |
| [m] | map | [m] | | | |

_ (SILENCE: appended to stop to indicate closure portion)
ˈ (BURST: appended to stop symbol to indicate burst portion)

Component I is a set of rules for determining the sequence of segment labels to be associated with the synthetic utterance. In the preliminary experiment in Chapt. 2, the ARPAbet representation internal to MITalk was used directly as the segment label sequence. Flexibility is gained by allowing the user to specify a set of rules to apply either to the MITalk representation or to the original word string to determine the string of symbols that will be aligned with the synthetic speech signal. Since we are interested in a phonetic segmentation in this study, we use the ARPAbet representation as input to component I. A set of rules, described in Sect. 3.3, are applied to this MITalk representation to derive a CPA representation. The CPA representation is not a simple one-for-one translation of the MITalk transcription: CPA and MITalk representations use different levels of segmentation.

Component II involves the application of phonological rules to the output of component I. Phonological rules delete and insert segments in the segment label sequence based upon segmental context. The output of component II is a modified segment label sequence. It is the actual sequence of segments that will be located in the synthetic model utterance.

Component III takes as input both the segment label sequence produced by component II and the sequence of parameter frames produced by the text-to-speech system. The purpose of component III is to locate segment boundaries which are determinable from logical predicates defined on the acoustic parameters of two or three consecutive parameter frames. For example, the predicate large-bandwidth-change is true iff the sum of the absolute differences in formant bandwidths between the preceding and current frames is greater than a fixed threshhold. Component III is a finite state machine in which states represent classes of segments and predicates are associated with transitions.

Certain kinds of boundaries cannot be determined locally, that is, by predicates on two or three parameter frames. For example, the boundary betweeen two vowels is not characterized by a well defined acoustic event. Nonetheless, we may wish to define vowel-vowel boundaries at some agreed upon place, such as at the frequency midpoint of a formant transition. In order to locate such a boundary, it may be necessary to know first where the second vowel ends. The purpose of component IV is find boundaries which are not locally determinable but depend on the locations of exterior anchor points. When component III encounters a segment transition for which it has no segmentation rule it defers boundary placement and continues searching for the next locally segmentable boundary. When the latter boundary is found, the deferred sequence, containing two or more non-locally segmentable labels, is passed to component IV along with its endpoints. Component IV applies global rules which search for minima and maxima of parameters to determine segment boundaries within the sequence. Finally, if no rules are available in component IV to segment a particular sequence of two segments, the sequence is arbitrarily divided at its temporal midpoint and a warning message is printed.

## 3.3 Segmentation Level and Rules Used in the Present Study

The previous section described the functional structure of the rule-based segmenter. We now turn our attention to the specific segmentation rules used in the present study. We first present the level of segmentation we have chosen. This is determined by the rules in component I. The phonological rules of in component II are then presented. Next, rules governing the placement of locally determinable boundaries are discussed. Finally, we discuss rules governing the placement of non-locally determinable boundaries.

### 3.3.1 Level of segmentation

The level of segmentation we have chosen uses somewhat smaller segments than the implicit segmentation used by MITalk. Stops are segmented into two parts: the stop closure and the stop burst; vowel + [r] sequences are segmented as two units rather than as a single unit as is done in MITalk; the phoneme [hw] is described as two segments, whereas MITalk uses a single segment.

Component I consists of the rules given in Table 3-2, which are used to translate from a MITalk phonetic transcription into a phonetic label sequence. Most of the rules are a simple translation from the ARPAbet-like transcription of MITalk into CPA, however, several rules involve more than just a one-for-one substitution of symbols. All symbols on the left side of rules represent single segments. Output sequences contain either one or two segment labels. In a few cases, more than one MITalk symbol is mapped into the same CPA symbol.

TABLE 3-2. Component I: Rules used in component I to determine level of segmentation.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| WH | ---> | h w | IXR | ---> | i r | Y | ---> | j |
| SIL | ---> | _ | ER | ---> | r | YY | ---> | j |
| IY | ---> | i | EXR | ---> | E r | W | ---> | w |
| IX | ---> | I | AXR | ---> | A r | L | ---> | l |
| IH | ---> | I | OXR | ---> | O r | LX | ---> | l |
| EY | ---> | e | UXR | ---> | u r | R | ---> | r |
| EH | ---> | E | V | ---> | v | RX | ---> | r |
| AE | ---> | @ | DH | ---> | D | H | ---> | h |
| AY | ---> | aj | Z | ---> | z | P | ---> | p |
| AW | ---> | aw | GP | ---> | g | T | ---> | t |
| AA | ---> | A | G | ---> | g | CH | ---> | tS |
| AH | ---> | ^ | ZH | ---> | Z | TQ | ---> | t |
| AX | ---> | * | F | ---> | f | KP | ---> | k |
| AXP | ---> | h | TH | ---> | T | K | ---> | k |
| AO | ---> | O | S | ---> | s | Q | ---> | k |
| OY | ---> | O j | SH | ---> | S | B | ---> | b |
| OW | ---> | o | M | ---> | m | D | ---> | d |
| OH | ---> | O | N | ---> | n | DX | ---> | d |
| UW | ---> | u | NG | ---> | g~ | HX | ---> | h |
| YU | ---> | j u | EM | ---> | m | J | ---> | Z |
| UH | ---> | U | EN | ---> | n | | | |

The rules listed above are applied by component I. Even though we defined the level of segmentation to include separate segments for stop closure and stop burst, the distinction does not appear in the output of the rules given in Table 3-2. These will be inserted by the phonological component, which is described in the next section.

### 3.3.2 Phonological rules

The phonological component presently implemented contains only two rules: stop burst insertion, which translates stop consonants into a stop closure followed by a stop burst, and stop burst deletion, which deletes stop bursts before sibilants, nasals, and stops. The rules of component II are listed below in Table 3-3.

TABLE 3-3. Component II: Phonological rules of burst insertion and burst deletion.

**Burst insertion rules:**

| | | |
|---|---|---|
| p | ---> | p_ p! |
| t | ---> | t_ t' |
| k | ---> | k_ k' |
| b | ---> | b_ b' |
| d | ---> | d_ d' |
| g | ---> | g_ g' |

**Burst deletion rule:**

Delete any of the following: p! t! k! b' d! g',
before any of the following: s S z Z m n g~ p_ t_ k_ b_ d_ g_

Table 3-4 illustrates the application of three processes described above to the text input of sentences 1 through 10. First, the text is translated by the MITalk letter-to-sound conversion rules into MITalk phonetic symbols. The MITalk representation serves as input to component I, which translates it into a CPA transcription of the desired level. Finally, component II applies phonological rules of burst insertion and burst deletion to the output of component I to yield the final segment label sequence to be used for automatic alignment.

TABLE 3-4. Result of applying MITalk letter-to-sound rules to input text; CPA transcription resulting from applying component I to MITalk output; effect on CPA transcription of applying burst insertion and burst deletion rules (component II).

Sentence 1

| | |
|---|---|
| text: | The goose was brought straight from the old market. |
| MITalk: | SIL DH AX G UW S W AH Z B R AO T S T R EY T F R AX M DH IY OW LX D M AXR K AX T AXP SIL |
| comp. I | _ D * g u s w ˆ z b r O t s t r e t f r * m D i ȯ l d m A r k * t h _ |
| comp. II: | _ D * g_ g' u s w ˆ z b_ b' r O t_ s t_ t' r e t_ t' f r * m D i o l d_ m A r k_ k! * t_ t' h _ |

## Sentence 2

| | |
|---|---|
| text: | The sink ıs the thıng ın which we pile dishes. |
| MITalk: | SIL DH AX S IH NG K IH Z DH AX TH IH NG IH N WH IH CH SH W IY P AY LX D IH SH IH Z SIL |
| comp. I: | _ D * s I g˜ k I z D * T I g˜ I n h w I t S w i p aj l d I S Ī z _ |
| comp. II: | _ D * s I g˜ k_ k' I z D * T I g˜ I n h w I t_ S w ı p_ p' aȷ Ī d_ d' I S I z̄ _ |

## Sentence 3

| | |
|---|---|
| text: | A whıff of it will cure the most stubborn cold. |
| MITalk: | SIL AX WH IH F AX V IH TQ W IH LX KP YY UXR DH AX M OW S T S T AH B ER N K OW LX D AXP SIL |
| Comp. I: | _ * h w I f * v I t w I l k ȷ u r D * m o s t s t ˆ b r n k o l̄ d h _ |
| comp. II: | _ * h w I f * v I t_ t' w I l k_ k! ȷ u r D * m o s t_ s t_ t̄' ˆ b_ b' r n k_ k̄' o l d_ d' h _ |

## Sentence 4

| | |
|---|---|
| text: | The facts don´t always show who is right. |
| MITalk: | SIL DH AX F AE K T S D OW N TQ AO LX W EY Z SH OW HX UW IH Z R AY T AXP SIL |
| comp. I: | _ D * f @ k t s d o n t O l w e z S o h u I z r aj t h _ |
| comp. II: | _ D * f @ k_ t_ s d_ d' o n t_ t' O l w e z S o h u I z r aj t_ t! h _ |

Sentence 5

text: She flaps her cape as she parades the street.

MITalk: SIL SH IY F L AE P S H ER KP EY P AE Z SH IY P AX R EY D Z DH AX S T R IY T AXP

comp. I: _ S ı f l @ p s h r k e p @ z S ı p * r e d z D * s t r i t h _

comp. II: _ S i f l @ p_ s h r k_ k' e p_ p' @ z S i p_ p' * r e d_ z D * s t_ t' r i t_ t' h _

Sentence 6

text: The loss of the cruıser was a blow to the fleet.

MITalk: SIL DH AX L AO S AX V DH AX K R UW Z ER SIL W AH Z AX B L OW T AX DH AX F L IY T AXP SIL

comp. I: _ D * l O s * v D * k r u z r _ w ˆ z * b l o t * D * f l ı t h _

comp. II: _ D * l O s * v D * k_ k' r u z r _ w ˆ z * b_ b' l o t_ t' * D * f l ı t_ t! h _

Sentence 7

text: Loop the braid to the left and then over.

MITalk: SIL L UW P DH AX B R EY D T AX DH AX L EH F T AXP SIL AE N DH EH N OW V ER SIL

comp. I: _ l u p D * b r e d t * D * l E f t h _ @ n D E n o v r _

comp. II: _ l u p_ 'p! D * b_ b' r e d_ t_ t! * D * l E f t_ t' h _ @ n D E n o v r _

Sentence 8

text: Plead with the lawyer to drop the lost cause.

MITalk: SIL P L IY D W IH TH DH AX L AO YY ER T AX D R AA P DH AX L AO S T K AO Z SIL

comp. I: _ p l i d w I T D * l O j r t * d r A p D * l O s t k O z _

comp. II: _ p_ p' l ı d_ d' w I T ~~D~~ * l O j r t_ t' * d_ d' r A p_ p! D * l O s t_ k_ k' O z _

Sentence 9

text: Calves thrive on tender spring grass.

MITalk: SIL KP AE V Z TH R AY V AO N T EH N D ER S P R IH NG G R AE S SIL

comp. I: _ k @ v z T r aj v O n t E n d r s p r I g~ g r @ s _

ouput: _ k_ k' @ v z T r aj v O n t_ t' E n d_ d' r s p_ p! r I g~ g_ g' r @ s _

Sentence 10

text: Post no bills on this office wall.

MITalk: SIL P OW S T N OW B IH LX Z AO N DH IH S AO F AX S W AO LX SIL

comp. I: _ p o s t n o b I l z O n D I s O f * s w O l _

comp. II: _ p_ p! o s t_ n o b_ b' I l z O n D I s O f * s w O l _

### 3.3.3 Placement of loçally determinable boundaries

Many boundaries can be determined by observing locally definable events in the parameter stream. By locally definable events, we mean boolean-valued functions of at most three consecutive parameter frames: the preceding frame, the current frame, and the following frame. Boundaries which are not locally determinable will be discussed in Sect. 3.3.4.

Segments are classified into 16 acoustic categories. Associated with each possible transition from one category to another is an event expected in the parameter stream. For those category transitions whose boundaries are not locally determinable, the event is simply specified as <u>deferred</u>. Table 3-5 shows the 16 acoustic categories and their member segments.

TABLE 3-5. Acoustic category definitions

| CATEGORY NAME | MEMBER SEGMENTS |
|---|---|
| silence | _ |
| vowel | i I e E @ aj aw A ^ * O o u U |
| voiced-sibilant | z Z |
| voiced-nonsibilant | v D |
| voiceless-sibilant | s S |
| voiceless-nonsibilant | f T |
| nasal | m n g~ |
| glide | j w |
| liquid | l r |
| voiceless-aspirate | h |
| voiceless-stop | p t k |
| voiced-stop | b d g |
| voiced-stop-closure | b_ d_ g_ |
| voiced-stop-burst | b' d' g' |
| voiceless-stop-closure | p_ t_ k_ |
| voiceless-stop-burst | p' t' k' |

Application of the locally determinable segmentation rules can be thought of as a finite state machine in which each segmental category corresponds to a state. Symbols on arcs, which the finite state machine must consume to make a transition, are locally determinable events in the parameter stream. A special event, deferred, causes a transition to occur immediately with no segmentation boundary being generated. The boundary will be located by a subsequent process which is allowed to make use of global patterns. Deferred segmentation is treated in detail in Sect. 3.3.4.

Table 3-6 gives the 120 rules for performing locally determinable segmentation. Each rule is specified by giving its state of origin, which corresponds to the current segmental category, its destination state, which corresponds to the segmental category of the next segment in the input string, and the acoustic event required by the rule. Definitions of the acoustic events themselves are given in Table 3-7.

Rules were determined in a heuristic manner. They result in a correct segmentation of the ten phonetically balanced sentences used in this study, but are not guaranteed to work on new material. It is expected that if segmentation of a significant quantity of new material were attempted, a small number of additions or modifications to the rules would be required.

TABLE 3-6. Component III: Rules used for locally determinable segmentation of synthetic speech.

| TRANSITION | | | EVENT (see Table 3-7) |
|---|---|---|---|
| vowel | --> | vowel | deferred |
| vowel | --> | liquid | deferred |
| vowel | --> | glide | deferred |
| vowel | --> | nasal | large-bandwidth-change |
| vowel | --> | voiceless-sibilant | av-->0 |
| vowel | --> | voiceless-nonsibilant | av+avc--><45 |
| vowel | --> | voiceless-aspirate | source-->aperiodic |
| vowel | --> | voiced-sibilant | source-->aperiodic |
| vowel | --> | voiced-nonsibilant | source-->very-aperiodic |
| vowel | --> | voiced-stop-closure | av-->0 |
| vowel | --> | voiceless-stop-closure | av-->0 |
| vowel | --> | glide | source-->aperiodic |
| vowel | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| liquid | --> | vowel | deferred |
| liquid | --> | liquid | deferred |
| liquid | --> | glide | deferred |
| liquid | --> | nasal | large-bandwidth-change |
| liquid | --> | voiced-sibilant | source-->aperiodic |
| liquid | --> | voiced-nonsibilant | source-->very-aperiodic |
| liquid | --> | voiceless-sibilant | source-->aperiodic |
| liquid | --> | voiceless-nonsibilant | source-->aperiodic |
| liquid | --> | voiceless-aspirate | av--><35 |
| liquid | --> | voiced-stop-closure | av-->0 |
| liquid | --> | voiceless-stop-closure | av-->0 |
| liquid | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| glide | --> | vowel | deferred |
| glide | --> | liquid | deferred |
| glide | --> | glide | deferred |
| glide | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| nasal | --> | vowel | large-bandwidth-change-delayed |
| nasal | --> | liquid | large-bandwidth-change |
| nasal | --> | glide | large-bandwidth-change |
| nasal | --> | voiceless-stop-closure | av--><35 |
| nasal | --> | voiced-stop-closure | av--><35 |
| nasal | --> | voiced-sibilant | large-bandwidth-change |
| nasal | --> | voiced-nonsibilant | large-bandwidth-change |
| nasal | --> | voiceless-sibilant | large-bandwidth-change |
| nasal | --> | voiceless-nonsibilant | large-bandwidth-change |
| nasal | --> | voiceless-aspirate | large-bandwidth-change-delayed |
| nasal | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| voiced-sibilant | --> | vowel | source-->periodic |
| voiced-sibilant | --> | voiced-nonsibilant | avc-->>50 |
| voiced-sibilant | --> | voiceless-sibilant | av-->0 |
| voiced-sibilant | --> | voiceless-nonsibilant | av-->0 |
| voiced-sibilant | --> | voiced-stop-closure | av-->0 |
| voiced-sibilant | --> | voiceless-stop-closure | av-->0 |
| voiced-sibilant | --> | silence | av+af+ah--><40 |

| | | | |
|---|---|---|---|
| voiceless-sibilant | --> | liquid | av-->positive |
| voiceless-sibilant | --> | vowel | av-->positive |
| voiceless-sibilant | --> | glide | av+avc>af+ah |
| voiceless-sibilant | --> | nasal | av-->positive |
| voiceless-sibilant | --> | voiced-sibilant | av-->positive |
| voiceless-sibilant | --> | voiced-nonsibilant | av-->positive |
| voiceless-sibilant | --> | voiceless-stop-closure | af+ah>20-->af+ah=0 |
| voiceless-sibilant | --> | voiced-stop-closure | af+ah>20-->af+ah=0 |
| voiceless-sibilant | --> | voiceless-aspirate | af>ah-->af<ah |
| voiceless-sibilant | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| voiced-nonsibilant | --> | vowel | source-->periodic |
| voiced-nonsibilant | --> | liquid | source-->periodic |
| voiced-nonsibilant | --> | nasal | source-->periodic |
| voiced-nonsibilant | --> | voiced-sibilant | deferred |
| voiced-nonsibilant | --> | voiced-nonsibilant | deferred |
| voiced-nonsibilant | --> | voiced-stop-closure | av-->0 |
| voiced-nonsibilant | --> | voiceless-stop-closure | av-->0 |
| voiced-nonsibilant | --> | silence | av+af+ah--><40 |

| | | | |
|---|---|---|---|
| voiceless-nonsibilant | --> | liquid | av-->positive |
| voiceless-nonsibilant | --> | vowel | av-->positive |
| voiceless-nonsibilant | --> | glide | av+avc>af+ah |
| voiceless-nonsibilant | --> | nasal | av-->positive |
| voiceless-nonsibilant | --> | voiced-sibilant | av-->positive |
| voiceless-nonsibilant | --> | voiced-nonsibilant | av-->positive |
| voiceless-nonsibilant | --> | voiceless-stop-closure | af+ah>20-->af+ah=0 |
| voiceless-nonsibilant | --> | voiced-stop-closure | af+ah>20-->af+ah=0 |
| voiceless-nonsibilant | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| voiced-stop-closure | --> | nasal | av-->positive |
| voiced-stop-closure | --> | voiced-sibilant | af=0-->af>40 |
| voiced-stop-closure | --> | voiceless-sibilant | af=0-->af>40 |
| voiced-stop-closure | --> | voiceless-stop-closure | deferred |
| voiced-stop-closure | --> | voiced-stop-closure | deferred |
| voiced-stop-closure | --> | anything | af=0-->af>40 |
| voiced-stop-closure | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| voiceless-stop-closure | --> | nasal | av-->positive |
| voiceless-stop-closure | --> | voiceless-sibilant | af=0-->af>40 |
| voiceless-stop-closure | --> | voiceless-nonsibilant | af=0-->af>40 |
| voiceless-stop-closure | --> | voiced-sibilant | af=0-->af>40 |
| voiceless-stop-closure | --> | voiceless-stop-closure | deferred |
| voiceless-stop-closure | --> | voiced-stop-closure | deferred |
| voiceless-stop-closure | --> | anything | af=0-->af>40 |
| voiceless-stop-closure | --> | silence | av+af+ah+avc--><30 |

| | | | |
|---|---|---|---|
| voiced-stop-burst | --> | vowel | af>40-->af=0 |
| voiced-stop-burst | --> | glide | av-->positive |
| voiced-stop-burst | --> | liquid | af>40-->af=0 |
| voiced-stop-burst | --> | nasal | af>40-->af=0 |
| voiced-stop-burst | --> | voiced-nonsibilant | av-->positive |
| voiced-stop-burst | --> | voiced-stop-closure | af>40-->af=0 |
| voiced-stop-burst | --> | voiceless-stop-closure | af>40-->af=0 |
| voiced-stop-burst | --> | voiceless-aspirate | af>40-->af=0 |
| voiced-stop-burst | --> | silence | av+af+ah+avc--><30 |

|  |  |  |  |
|---|---|---|---|
| voiceless-stop-burst | --> | vowel | af>40-->af=0 |
| voiceless-stop-burst | --> | liquid | af>40-->af=0 |
| voiceless-stop-burst | --> | glide | af>40-->af=0 |
| voiceless-stop-burst | --> | nasal | af>40-->af=0 |
| voiceless-stop-burst | --> | voiced-nonsibilant | av-->positive |
| voiceless-stop-burst | --> | voiceless-nonsibilant | ab-->positive |
| voiceless-stop-burst | --> | voiceless-aspirate | af>40-->af=0 |
| voiceless-stop-burst | --> | voiced-stop-closure | af>40-->af=0 |
| voiceless-stop-burst | --> | voiceless-stop-closure | af>40-->af=0 |
| voiceless-stop-burst | --> | silence | av+af+ah+avc--><30 |
| voiceless-aspirate | --> | vowel | source-->periodic |
| voiceless-aspirate | --> | liquid | source-->periodic |
| voiceless-aspirate | --> | glide | source-->periodic |
| voiceless-aspirate | --> | silence | av+af+ah+avc--><30 |
| silence | --> | vowel | av-->positive |
| silence | --> | liquid | av-->positive |
| silence | --> | glide | av-->positive |
| silence | --> | nasal | av-->positive |
| silence | --> | voiced-sibilant | av-->positive |
| silence | --> | voiced-nonsibilant | av-->positive |
| silence | --> | voiceless-sibilant | af+ah<50-->af+ah>50 |
| silence | --> | voiceless-nonsibilant | af=0-->af>40 |

Table 3-7 defines the acoustic events used in the rules defined in Table 3-6. Event definitions are in terms of the acoustic parameters used to control MITalk's terminal analog synthesizer (Klatt, 1980). Acoustic events are boolean-valued functions defined on the acoustic parameters of the synthesizer. Numerical values associated with amplitude parameters (av, ab, af, ah, avc) are in decibel units.

TABLE 3-7. Definitions of acoustic events used in local segmentation rules.

ab-->positive

Amplitude of bypass in previous frame was zero and in current frame is greater than zero.

af>ah-->af<ah

In the current frame, amplitude of frication is greater than amplitude of aspiration, while in the next frame amplitude of frication is less than or equal to amplitude of aspiration.

af>40-->af=0

Amplitude of frication in previous frame was greater than 40 and amplitude of frication in current frame is zero.

af=0-->af>40

Amplitude of frication in the current frame is zero and in next frame is greater than 40.

af+ah>20-->af+ah=0

Sum of amplitudes of frication and aspiration in the current frame is greater that 20 and in next frame is zero.

af+ah<50-->af+ah>50

Sum of amplitudes of frication and aspiration in the current frame is less than 50 and in next frame is greater than or equal to 50.

av-->0

Amplitude of voicing in previous frame was positive and in current frame is zero.

av--><35

Amplitude of voicing in current frame is greater than or equal to 35 and in next frame is less than 35.

av-->positive

Amplitude of voicing in current frame is zero and in following frame is greater than zero.

av+af+ah--><40

Sum of the amplitudes of voicing, frication, and aspiration in current frame is greater than or equal to 40 and in next frame is less than 40.

av+af+ah+avc--><30

Sum of the amplitudes of voicing, frication, aspiration, and sinusoidal voicing in current frame is greater than or equal to 30 and in the next frame is less than 30.

av+avc--><45

Sum of the amplitudes of voicing and sinusoidal voicing in current frame is greater than or equal to 45 and in the next frame is less than 45.

av+avc>af+ah

In the current frame the periodic excitation (voicing plus sinusoidal voicing) is less than or equal to 45 while in the next frame periodic excitation is less than or equal to 45.

avc-->>50

Amplitude of sinuosoidal voicing in the current frame is less than or equal to 50, while in the next frame it is greater than 50.

deferred

This is not an acoustic event, but rather a signal to defer the boundary placement decision to the next stage of processing, discussed in Sect. 3.3.4.

large-bandwidth-change

The sum of the absolute differences in the bandwidths of formants one through three between the current frame and the next frame exceeds 50.

large-bandwidth-change-delayed

This event is true one frame after large-bandwidth-change is true, i.e., when the sum of the absolute bandwidths differences betweent the previous and current frames exceeds 50.

source-->aperiodic

The sum of the amplitudes of frication and aspiration is greater than or equal to the amplitude of voicing for this frame and less than the amplitude of voicing for the next frame.

source-->periodic

The sum of the amplitudes of frication and aspiration is less than the amplitude of voicing for this frame and greater than or equal to the amplitude of voicing for the next frame.

source-->very-aperiodic

The sum of the amplitudes of frication and aspiration is less than 10 plus the amplitude of voicing for this frame and is greater than or equal to 10 plus the amplitude of voicing for the next frame.

We now describe the operation of component III, which performs locally determinable segmentation using the rules in Table 3-6 and the acoustic events in Table 3-7. Sentence 1 serves as an illustration. At each stage of segmentation, the state of the segmenter is represented as three lists. The input list contains the list of segment labels to be associated with the input sentence. The current list contains the item(s) currently being segmented. The output list contains segmentation labels with their associated starting and ending frame numbers for the portion of the sentence for which segmentation has been completed.

For sentence 1, the initial state of the segmenter is as follows:

```
OUTPUT:

CURRENT:  _

INPUT:    D * g_ g' u s w ^ z b_ b' r 0 t_ s t_ t' r e t_ t' f r * m D i o
          l d_ m A r k_ k' * t_ t' h _
```

The segmenter is trying to find a boundary between silence, [_], and the segment [D] of the word "the." Since [D] is in the class voiced-nonsibilant (see Table 3-5), the rule which applies (see Table 3-6) is the one for segmenting silence followed by a voiced-nonsibilant. This rule requires that event av-->positive occur. Starting with frame 1 of the parameter stream as the current frame, the segmenter checks the truth value of av-->positive and finds it is false. The current frame is incremented to 2 and again av-->positive is false. This is continued until frame 6 is the current frame, at which point av-->positive is true. When this happens, the boundary between [_] and [D] has been found. [_]

is moved to the output list and its starting and ending frames, 1 and 6, are associated with it. [D] becomes the new current segment.

```
OUTPUT:   (1 6 _)

CURRENT:  D

INPUT:    * g_ g' u s w ^ z b_ b' r O t_ s t_ t' r e t_ t' f r * m D i o l
          d_ m A r k_ k! * t_ t' h _
```

Next, a rule is found to determine the boundary between a voiced-nonsibilant and a vowel. The event the segmenter will look for is source-->periodic. The current frame is intially frame 7, the beginning of the vowel [*], and is incremented until source-->periodic. In this example, source-->periodic at frame 17, with the following result:

```
OUTPUT:   (1 6 _) (7 17 D)

CURRENT:  *

INPUT:    g_ g' u s w ^ z b_ b' r O t_ s t_ t' r e t_ t' f r * m D i o l
          d_ m A r k_ k' * t_ t' h _
```

The next rule which applies is vowel --> voiced-stop-closure. The corresponding event is av-->0. The current frame is initially frame 18. It is incremented until at frame 29, av-->0 becomes true:

```
OUTPUT:    (1 6 _) (7 17 D) (18 29 *)

CURRENT:   g_

INPUT:     g' u s w ^ z b_ b' r 0 t_ s t_ t' r e t_ t' f r * m D i o l d_ m
           A r k_ k' * t_ t' h _
```

The next few steps are similar to those described so far:

```
RULE:      voiced-stop-closure  -->  anything
EVENT:     af=0-->af>40

OUTPUT:    (1 6 _) (7 17 D) (18 29 *) (30 40 g_)

CURRENT:   g'

INPUT:     u s w ^ z b_ b' r 0 t_ s t_ t' r e t_ t' f r * m D i o l d_ m A
           r k_ k' * t_ t' h _

RULE:      voiced-stop-burst  -->  vowel
EVENT:     af>40-->af=0

OUTPUT:    (1 6 _) (7 17 D) (18 29 *) (30 40 g_) (41 45 g')

CURRENT:   u

INPUT:     s w ^ z b_ b! r 0 t_ s t_ t' r e t_ t! f r * m D i o l d_ m A r
           k_ k! * t_ t' h _

RULE:      vowel  -->  voiceless-sibilant
EVENT:     av-->0

OUTPUT:    (1 6 _) (7 17 D) (18 29 *) (30 40 g_) (41 45 g') (46 85 u)

CURRENT:   s

INPUT:     w ^ z b_ b' r 0 t_ s t_ t' r e t_ t' f r * m D i o l d_ m A r k_
           k! * t_ t! h _

RULE:      voiceless-sibilant  -->  glide
EVENT:     av+avc>af+ah

OUTPUT:    (1 6 _) (7 17 D) (18 29 *) (30 40 g_) (41 45 g!) (46 85 u)
           (86 107 s)

CURRENT:   w
```

```
INPUT:    ^ z b_ b' r O t_ s t_ t' r e t_ t' f r * m D i o l d_ m A r k_
          k' * t_ t' h _

RULE:     glide --> vowel
EVENT:    deferred
```

At this point, since the required "event" is deferred, no segmentation is attempted. Instead, the [^] is moved from the input list to the current list and the rule vowel --> voiced-sibilant is applied:

```
OUTPUT:   (1 6 _) (7 17 D) (18 29 *) (30 40 g_) (41 45 g!) (46 85 u)
          (86 107 s)

CURRENT:  w ^

INPUT:    z b_ b' r O t_ s t_ t' r e t_ t! f r * m D i o l d_ m A r k_ k'
          * t_ t' h _

RULE:     vowel --> voiced-sibilant
EVENT:    source-->aperiodic
```

The current frame is initially frame 108. It is incremented until source-->aperiodic. This occurs at frame 127, defining the boundary between [^] and [z]. The segmenter now has available the information that the sequence [w ^] begins at frame 108 and ends at frame 127. This information, (108 127 (w ^)), is sent to component IV for further segmentation. Component IV is discussed in detail in Sect. 3.3.4. After component IV has applied, its output is added to the output list:

```
OUTPUT:    (1 6 _) (7 17 D) (18 29 *) (30 40 g_) (41 45 g') (46 85 u)
           (86 107 s) (108 117 w) (118 127 ^)

CURRENT:   z

INPUT:     b_ b' r O t_ s t_ t' r e t_ t' f r * m D i o l d_ m A r k_ k' *
           t_ t' h _)
```

RULE:      voiced-sibilant  -->  voiced-stop-closure
EVENT:     $av_T$->0

```
OUTPUT:    (1 6 _) (7 17 D) (18 29 *) (30 40 g_) (41 45 g') (46 85 u)
           (86 107 s) (108 117 w) (118 127 ^) (128 136 z)

CURRENT:   b_

INPUT:     b' r O t_ s t_ t' r e t_ t' f r * m D i o l d_ m A r k_ k' * t_
           t' h _
```

The process continues in a similar manner, performing locally determinable segmentations immediately and deferring globally determinable ones. In some cases, several consecutive segments are deferred. When the input and current lists are empty, segmentation is finally complete:

```
OUTPUT:    (1 6 _) (7 17 D) (18 29 *) (30 40 g_) (41 45 g!) (46 85 u)
           (86 107 s) (108 117 w) (118 127 ^) (128 136 z) (137 150 b_)
           (151 152 b') (153 169 r) (170 197 O) (198 205 t_) (206 223 s)
           (224 232 t_) (233 238 t') (239 257 r) (258 278 e) (279 285 t_)
           (286 290 t!) (291 309 f) (310 315 r) (316 321 *) (322 333 m)
           (334 342 D) (343 368 i) (369 394 o) (395 412 l) (413 421 d_)
           (422 437 m) (438 449 A) (450 470 r) (471 479 k_) (480 485 k!)
           (486 499 *) (500 509 t_) (510 513 t') (514 523 h) (524 692 _)

CURRENT:

INPUT:
```

The next section describes the function of component IV, which is called upon by component III to segment sequences whose boundaries are not locally determinable.

### 3.3.4 Placement of locally indeterminable boundaries

In the previous section, we described how component III finds locally determinable boundaries and mentioned that deferred boundaries, such as the one between [w] and [^], are determined by component IV. In this section, we discuss how deferred boundaries are determined.

By examining Table 3-6, we can see that the following segmentation rules are deferred:

| | | |
|---|---|---|
| vowel | --> | vowel |
| vowel | --> | liquid |
| vowel | --> | glide |
| liquid | --> | vowel |
| liquid | --> | liquid |
| liquid | --> | glide |
| glide | --> | vowel |
| glide | --> | liquid |
| glide | --> | glide |
| voiced-nonsibilant | --> | voiced-sibilant |
| voiced-nonsibilant | --> | voiced-nonsibilant |
| voiced-stop-closure | --> | voiceless-stop-closure |
| voiced-stop-closure | --> | voiced-stop-closure |
| voiceless-stop-closure | --> | voiceless-stop-closure |
| voiceless-stop-closure | --> | voiced-stop-closure |

The first nine of these rules refer to segmentation of vowels, liquids, and glides. Since no discrete event occurs at such boundaries, component IV uses global rules based on maxima and minima of formant frequencies. The remaining six rules refer to segmentation of voiced fricatives and segmentation of stop closures. The latter are clearly not segmentable by rule: the segmenter uses the heuristic of placing the boundary at the temporal midpoint of the sequence. The same heuristic is used to segment voiced fricatives, although conceivably it would be possible to segment voiced-nonsibilant/voiced-sibilant sequences on acoustic grounds. The remainder of this section describes the rules used to segment vowels, liquids, and glides.

Vowels, liquids, and glides are assigned the acoustic features shown in Table 3-8. In addition, the feature <u>anything</u> is assigned to all segments.

TABLE 3-8. Features assigned to vowels, liquids, and glides.

| SEGMENTS | FEATURES |
|---|---|
| l | lo-F1 hi-F3 |
| r | lo-F3 |
| j i I e | hi-F3 |
| w u U o | lo-F2 |
| aj aw | rising-diphthong |

Rules given in Table 3-9 are specified in terms of these features and indicate global acoustic events used to determine the segment boundaries. Global events differ from local events in that their scope is the entire range of the deferred subsequence rather than just three frames. The rules are tried in order. The first rule whose context is satisfied is used to perform the segmentation.

TABLE 3-9. Rules specifying global acoustic events used for segmenting vowels, liquids, and glides.

| FEATURE OF FIRST SEGMENT | FEATURE OF SECOND SEGMENT | EVENT |
|---|---|---|
| lo-F2 | lo-F3 | decreasing-F3 |
| lo-F2 | hi-F3 | increasing-F3 |
| lo-F2 | anything | increasing-F2 |
| hi-F2 | lo-F3 | decreasing-F3 |
| hi-F2 | anything | decreasing-F2 |
| lo-F3 | lo-F2 | increasing-F3 |
| lo-F3 | hi-F2 | increasing-F3 |
| lo-F3 | anything | increasing-F3 |
| hi-F3 | lo-F2 | decreasing-F3 |
| lo-F1 | hi-F2 | decreasing-F3 |
| lo-F1 | anything | decreasing-F3 |
| rising-diphthong | hi-F3 | increasing-F3 |
| anything | lo-F1 | decreasing-F1 |
| anything | lo-F2 | decreasing-F2 |
| anything | hi-F2 | increasing-F2 |
| anything | lo-F3 | decreasing-F3 |

We now describe the global acoustic events, of which increasing-F3 is an example. Each one consists of a direction, increasing or decreasing, and a parameter, F1, F2, or F3. If the direction is increasing, the segmenter will search for a local maximum of the parameter following a local minimum. This generates a candidate segmentation point, which is the frame between the extrema at which the parameter of interest most nearly approaches the average of the extrema. A figure of merit is assigned to the candidate segmentation point. If it exceeds a threshold, the candidate segmentation point is accepted. Otherwise, the next local minimum, local maximum sequence is evaluated. If all the extrema pairs have been evaluated and the threshold has not been exceeded, the candidate corresponding to the highest figure of merit is chosen. This algorithm results in plaçing segmentation boundaries at the frequency midpoints of formant transitions.

The figure of merit is the weighted sum of two values: (1) the extrema difference and (2) the difference between a candidate´s segmentation frame and an a priori estimate of that frame. The weight of the first value is positive, while that of the second is negative. The a priori estimate of segmentation boundary position is determined by assuming that each segment in the deferred subsequence has equal length.

## 3.4 Results of Rule-Based Segmentation of Synthetic Sentences.

Now that we have described the structure of the rule-based segmenter and the actual rules used, we examine the results of applying the segmenter to the set of ten phonetically balanced sentences. Figures 3-1 through 3-10 are spectrograms of sentences 1 through 10, respectively, which have been segmented and labelled by the rule-based segmenter. In Figs. 3-1 through 3-10, the abscissa is labelled in frames, where each frame corresponds to 5 ms. Displayed durations on each spectrogram were chosen so that the beginning and end correspond to segment boundaries, causing an integral number of segments to appear in each figure. As a result, the number of frames displayed is different for each spectrogram.

To understand the differences between the rule-based segmentation and that intrinsic to MITalk discussed in Chapt. 2, it is instructive to compare Figs. 2-1 through 2-4 with Figs. 3-1 through 3-4. Comparing Fig. 2-1a with 3-1a, the first difference observed is the treatment of the [g] stop burst. In the MITalk segmentation (Fig. 2-1a), the burst is analyzed as part of the [G] segment while the rule based segmenter further analyzes [g] into g-closure, indicated [g_], and g-burst, indicated [g']. Other examples of this difference appear at the following boundaries: [br] (Figs. 2-1b and 3-1b), [tr] (Figs. 2-1b and 3-1c), [k*] (Figs. 2-1d and 3-1e), [th] (Figs. 2-1e and 3-1e), etc. In each case, the rule-based segmenter appears to segment the burst correctly. Returning to the beginning of sentence 1, it can be seen that the MITalk and rule-based segmentations are generally in agreement until we reach the [sw] boundary, which occurs somewhat earlier in the MITalk segmentation.

Comparing Fig. 3-1b with 2-1a, the next discrepancy is in the [w^] boundary, which occurs later in the rule-based segmentation, in accordance with the, rules described in Sect. 3.3.4. These rules cause the [w^] boundary to be placed at the frequency midpoint of the second formant transition. Similarly, in Figs. 2.1b and 3-1b, the [rO] transition occurs later in the rule-based segmentation. The rule <u>increasing-F3</u> has applied, causing the [rO] boundary to be placed at the frequency midpoint of the F3 transition. Similarly, the [eo] boundary (Figs. 2-1c and 3-1d) is placed 30 ms later in the rule-based segmentation, corresponding to the frequency midpoint of F2. Other examples of the use of formant transition midpoints to place boundaries between vowels, liquids, and glides and how this compares with the MITalk segmentation are the transitions [wI] and [wi] (Figs. 2-2b and 3-2c), [Il] (Figs. 2-3a and 3-3b), [jur] (Figs. 2-3b and 3-3c), [we] (Figs. 2-4b and 3-4c), and [raj] (Figs. 2-4c and 3-4e). Note that in the segmentation of [jur], MITalk consideres [ur] a single segment, while the rule-based segmenter divides it into [u] and [r]. Similarly, [Ar] is treated as one segment in MITalk (Fig. 2-1d), while it is treated as two segments in the rule-based segmentation (3-1e).

Note the treatment of the t-burst before [s] in Fig. 2.1b as opposed to Figs. 3.1b-c. In the MITalk segmentation, 15 ms of frication is segmented as belonging to the [t], representing the burst, whereas in the rule-based segmentation, the t-burst is deleted by a phonological rule due to the presence of a following sibilant. The same phenomenon appears at the following boundaries: [tS] (Figs. 2-2b and 3-2c) and [ts] (Figs. 2-3c and 3-3d, Figs. 2-4a and 3-4a).

In Figs. 2-1d and 3-1e, we see that the [mA] boundary in the rule based-model corresponds more closely with the observed acoustic transition. The same phenomenon occurs at the [g~I] boundary in Figs. 2-2b and 3-2b and at the [mo] boundary in Figs. 2-3b and 3-3c.

MITalk tends to place boundaries between voiceless fricatives and vowels about 10 ms too early. The rule-based segmenter places such boundaries at the onset of voicing. Examples of this occur at the boundaries [sI] (Figs. 2-2a and 3-2a), [TI] (Figs. 2-2a and 3-2b), [SI] (Figs. 2-2d and 3-2e), [f*] (Figs. 2-3a and 3-3a), [f@] (Figs. 2-4a and 3-4a), and [So] (2-4b and 3-4c,d).

Occasionally, MITalk cuts the burst in two, considering the first part as belonging to the stop and the second half as part of the following vowel. The corresponding rule-based segmentations treat the burst consistently. Example of this appear in [dI] (Figs. 2-2c and 3-2c), [do] (Figs. 2-4a and 3-4b), and [tO] (Figs. 2-4b and 3-4b).

In summary, the following imprecisions and inconsistencies present in the MITalk segmentation have been overcome using the rule-based segmenter: inconsistency of burst segmentation, imprecision in segmentation of voiceless-fricative/vowel boundaries, imprecision in segmentation of nasal-consonant/vowel boundaries. Segmentation level has been modified as have the rules for determining boundaries between vowels, liquids, and glides.

### 3.5 Discussion of Rule-Based Segmentation

Results of the previous section indicate that use of a rule-based segmenter gave more precise and consistent analyses of synthetic speech than were available by simply using the nominal boundary locations of the synthesizer. In addition, the rule-based segmenter allows the user to specify segmentation level and the specific criteria used in selecting a segment boundary. It is useful at this point to speculate on the applicability of rule-based segmentation techniques to natural speech. Two facilitating factors that are present when this technique is applied to synthetic speech disappear when it is applied to natural speech. These factors are (1) the consistency (lack of free variation) inherent in synthetic speech and (2) the availability of exact values of acoustic parameters (formant frequencies, bandwidths, etc.)

Although there is no free (i.e., stochastic) variation present in the synthetic speech, the model used to generate it is sufficiently complex and rich in phonologically conditioned variation that developing a set of segmentation rules for the ten sentences discussed above was nontrivial. Getting the rules to work for the first four sentences was the most difficult. Sentences 5, 6, 7, and 9 were correctly segmented using those rules. Finally, modifications had to be made to correctly segment sentences 8 and 10.

To illustrate an insight gained from debugging the rules, the specific case of sentence 10 will be described. Using the rules developed for sentences 1 through 4, one error occurred in sentence 10: in the word <u>wall</u>, the boundary between [O] and [l] was much too early, near the beginning of [O]. The following rule had applied, causing the error:

anything + hi-F3 ----> increasing-F3

Recall that the algorithm for handling vowel-liquid segmentations is to place the boundary at a parameter value half way between the transition extrema. This rule says: find a local F3 minimum in [O]; find a local F3 maximum in [l]; place the [Ol] boundary at the frame whose F3 most closely approaches the average F3 of these two local extrema. The problem occurs because F3 increases monotonically from the beginning of [O] to the near the end of [l], so no local F3 minimum is found. The algorithm therefore uses F3 of the first frame of [O] as the local F3 minimum. This value is averaged with the local F3 maximum, found near the end of [l], to determine the F3 value at which to place the [Ol] boundary. Because the F3 minimum represents not the beginning of the [Ol] transition but a much lower F3 at the beginning of [O], the boundary is placed much too early.

This problem was solved by changing the rule to use F1 instead of F3:

anything + lo-F1 ----> decreasing-F1

This rule worked well for sentence 10 and for most other vowel-liquid boundaries, but caused an error in sentence 2. In the word pile, the boundary between [aj] and [l] was now incorrect since a local F1 maximum followed by a local F1 minimum occur within the [aj] itself, causing the segment boundary to be placed between [a] and [j] rather than between [aj] and [l]. This problem was solved by adding a new rule, which precedes the modified rule in the rule ordering:

rising-diphthong + hi-F3 ----> decreasing-F3

With this final modification, all ten sentences were correctly segmented. Although no further sentences have been tested, it is the intuitive feeling of the experimenter that most sentences would now be correctly segmented by the rule-based segmenter, but that there would still be some modifications required to handle certain phonetic sequences not yet encountered.

The representation of the rules could be greatly condensed by extending the use of feature notation to the locally determinable segmentation rules (component III). Rules in components III and IV should be expressed in the same feature-based notation. The notion of "event" could be generalized to include local and global acoustic events. Figures of merit could be computed for all events, rather than just for the global ones. This would allow the rule-based segmentation to be treated as a dynamic programming problem in which the quantity to be maximized would be the sum of the figures of merit of all the segmentation boundaries. A better heuristic is clearly needed to generate a priori duration estimates.

Some of the acoustic parameters would have to be changed in order to apply the rule-based segmentation technique to natural speech. Those parameters that are difficult to measure or difficult to distinguish from each other would be replaced by other more accessible parameters. For example, it may not be useful to try to measure quantities such as the amplitude of sinusoidal voicing (avc), the bypass amplitude (ab), or to distinguish the amplitude of frication (af) from that of aspiration (ah). On the other hand, zero-crossing rate, which is not currently in the set of parameters being used, could be very useful for locating boundaries between sonorant and nonsonorant segments.

In order to overcome phonological variability (phonemic deletion, insertion, and substitution), a more sophisticated phonological component is necessary. Our component II is categorical: A segment is either deleted or retained. In order to handle natural speech, component II must be able to mark segments as having a certain probability of being deleted. This probability could then be used in the computation of the figure of merit of a particular segmentation hypothesis.

# CHAPTER 4: PRELIMINARY EXPERIMENT USING RULE-BASED SEGMENTATION OF MODEL SPEECH

In a preliminary experiment (Lennig, 1983), a single synthetic sentence was segmented by the rule-based segmenter and used to segment four natural versions of the same sentence by applying dynamic time warping. This chapter describes the experiment, which differs from that described in Chapt. 2 where Assumption 3 did not hold. In a parallel experiment also described in this chapter, Assumption 2 is explored by using a hand segmented model in place of synthetic speech. Errors are analyzed by segment boundary type.

One conclusion drawn from this analysis is that potential weaknesses of the technique lie in the segmentation of sonorants (vowels, liquids, nasals, glides) and of short-duration events, such as stop burst. The difficulty of segmenting sonorants can be attributed to inability of the distance metric to ignore interspeaker differences while emphasizing phonetic ones (cf., Assumption 4). The second weakness is likely due to insensitivity of the time warping algorithm to short events (cf., Assumption 5).

Although segmentation errors in this experiment are determined subjectively on a case by case basis, it is evident that an objective method for evaluating the correctness of a segmentation is needed. The problem with quantitative measures based on deviation from a given norm was discussed in Chapt. 2. At the end of this chapter, a quantitative correctness measure is proposed which partially overcomes the problem.

## 4.1 Experimental Procedure

The following sentence (sentence 2) was pronounced by four male speakers and also synthesized using the MITalk text-to-speech system:

The sink is the thing in which we pile dishes.

Three of the speakers were native speakers of Montreal English; the fourth was a native speaker of New York English. The naturally produced sentences were lowpass filtered at 4.4 kHz and sampled at 10 kHz. The endpoints of each sentence were manually determined. The synthetic sentence was preprocessed to yield a mel-frequency cepstrum every 5.0 ms. The naturally produced sentences were processed similarly, except that the frame advance for each speaker was chosen between 4.7 and 6.4 ms so as to yield approximately the same number of speech frames as in the synthetic reference (454 frames). Unequal numbers of frames are undesirable in the time warping procedure because slope constraints and slope penalties are used to make the warp path to tend toward a 45 degree line.

The decimated-grid, symmetric time warping algorithm proposed by Mermelstein (1978) was used, in which grid point (i,j) is accessible from points (i-1,j-1), (i-2,j), and (i,j-2). The penalty for a vertical or horizontal step is to multiply the local distance at (i,j) by 1.5. Two different local slope constraints were tried: unconstrained and constrained slope. The constrained slope algorithm permits a maximum of one consecutive vertical or horizontal step. In a third trial, the first speaker's utterance was hand segmented and used as a reference in conjunction with the constrained slope algorithm to induce segmentations on the remaining three naturally produced sentences.

## 4.2 Results and Discussion

Segmentation and labelling induced on the natural speech was inspected by viewing spectrograms annotated with this information. Each automatically determined segmentation boundary was subjectively classified as correct or incorrect. Subjective scoring was preferred because certain segment boundaries, such as the endpoints of a stop burst, are more precisely determined by the speech signal, while others, e.g., the boundary between a vowel and a liquid or glide, may be farther away from a prescribed norm and still be considered correct.

Figures 4-1a through 4-1e show time-aligned spectrograms of the synthetic model (top) and a natural utterance (bottom) produced by one of the four speakers (DS). Slope-constrained dynamic time-warping was used. In each figure, the time scales of the two spectrograms are not identical: They have been linearly adjusted in order to display the same sequence of segments.

In Fig. 4-1a, the boundaries [D*] and [*s] appear to be a few frames late. This was not counted as an error in the subjective scoring procedure, since only serious errors were considered. Another minor error that was not counted was the early placement of the [sI] boundary. The only error in Fig. 4-1a which was considered in the results presented here was the boundary [k_ k'], which occurred late. A gross error occurred in Fig. 4-1b in the placement of the [Ig~] boundary of thing: It is at least 10 frames (64 ms) late, occurring at the end of the actual [g~] segment instead of at the beginning. The desynchronization propagates to the next boundary, causing [g~I] to be late. The only other error counted was at the [ld_] boundary (Fig. 4-1d): The boundary late.

Table 4-1 displays segmentation errors, for all speakers, according to three boundary types: boundaries between sonorant segments (vowels, liquids, nasals, glides), boundaries between nonsonorant segments, and boundaries between a sonorant and a nonsonorant segment, in either order. As expected, boundaries between similar segment types give rise to higher error rates. Segmentation performance for all boundary categories is significantly better when a slope constraint is employed and significantly better using a natural speech reference as compared with a synthetic reference. Even using a natural reference, however, the 13 percent error rate obtained is unsatisfactory. This comparison serves to quantify the importance of Assumption 2: although unnaturalness of the synthetic model is clearly introducing segmentation error, Assumptions 4 and 5 are at least as important. The difference in performance obtained with different warp constraints (cf., Assumption 5) is greater than the difference obtained using a natural versus a synthetic model.

The advantage of the constrained time warping algorithm over the unconstrained algorithm appears most saliently in the detection of sonorant/sonorant boundaries. This may be because interspeaker spectral differences in sonorants overshadow spectral differences between different sonorant segments, leaving segment duration as the only reliable alignment criterion. Use of a speaker normalization for glottal source spectrum shape may improve the accuracy of sonorant/sonorant boundary localization.

One finding obscured by Table 4-1 is that stop burst localization contributes significantly to the error rate. In the constrained, synthetic trial, for example, seven of the eight nonsonorant/nonsonorant errors occur on closure/burst boundaries. Five of the nine sonorant/nonsonorant errors occur on burst/vowel boundaries. Similar results hold for the other trials. Often, what is observed on the annotated spectrogram is an erroneous localization of the stop burst somewhere in the middle of the stop closure, temporally disjunct from the actual burst. Such errors can be explained by the relatively small distance penalty incurred by burst misplacement, due to the segment's short duration. The problem should be viewed as an inadequacy of the time warping algorithm as currently formulated.

TABLE 4-1. Error rates for segment boundary location using constrained and unconstrained time warping algorithm.

| | synthetic reference | | natural reference |
|---|---|---|---|
| | unconstrained | constrained | constrained |
| sonorant/<br>sonorant | 26/32<br>(81 %) | 10/32<br>(31 %) | 4/24<br>(24 %) |
| nonsonorant/<br>nonsonorant | 12/20<br>(60 %) | 8/20<br>(40 %) | 2/15<br>(13 %) |
| sonorant/<br>nonsonorant | 13/72<br>(18 %) | 9/72<br>(13 %) | 6/54<br>(11 %) |
| TOTAL | 51/124<br>(41 %) | 27/124<br>(22 %) | 12/93<br>(13 %) |

## 4.3 Evaluating the Correctness of a Given Segmentation

In the experiment described above, the most difficult procedure was the subjective evaluation of correctness of the resulting segmentations. Although some cases were clear-cut, others caused soul searching on the part of the experimenter before a decision could be reached. Such subjectivity is dangerous in scientific experiments because experimenter bias is easily introduced. In this section we propose an objective method of evaluating segmentations which has certain desirable properties.

One approach to removing subjectity in segmentation experiments would be to hand segment the natural utterance in advance and then measure how closely the automatic transcription corresponds to the hand segmentation. This method removes any possible experimenter bias since hand segmentation is performed prior to the experiment. Another advantage of this method is that as many segmentation experiments as desired may be run against the same hand-segmented data. Different algorithms may be compared with a common standard.

The problem with this method lies in the measurement of deviations from the hand-segmented standard. Certain phone boundaries are temporally indefinite while others are much more precisely defined. Measures based on mean square error are difficult to interpret. For example, a ten millisecond deviation from the hand segmentation at a vowel/liquid boundary is unimportant, whereas the same deviation at a stop closure/stop burst boundary is a true error. We now propose an evaluation method designed to overcome this problem.

Hand segmentation is used, but instead of specifying a specific temporal segmentation point between each pair of segments, a range of points is specified. Viewed another way, the beginning of one segment follows the end of the preceding segment with a variable number of intervening frames. This gives rise to unlabelled sections of speech between segments. If the automatic segmenter locates the segment boundary anywhere within this region, the segmentation is considered correct.

Otherwise, it is considered incorrect. Figure 4-2 illustrates the evaluation procedure for the first few phones of the phrase "She flaps..." Figure 4-2a represents the automatically segmented waveform and has a contiguous segmentation: each segment begins at the same place the previous segment ends. Figure 4-2b is the hand-segmented standard. Because the segmentation is noncontiguous, zones of indeterminacy lie between the prescribed segment locations. Since the [Si] boundary in 4-2a lies between the end of the [S] and the beginning of the [i] in Fig. 4-2b, this boundary would be considered correct. Similarly, the [if] boundary is correct. The [fl] boundary corresponds exactly with the end of [f] in the hand-segmented standard and is therefore also considered correct.

Chapter 5 describes a larger experiment in which noncontiguous hand-transcribed standard segmentations are used for evaluation.

# CHAPTER 5: EVALUATION OF AUTOMATIC SEGMENTATION TECHNIQUE

This chapter presents a quantitative evaluation of the automatic segmentation technique as applied to twenty naturally produced test sentences. Automatically generated segmentations of the sentences were scored by comparison against manual segmentations in which the experimenter specified a range of correct positions for each boundary.

## 5.1 Experimental Procedure

Sentences 1 through 10 were read once by each of two male speakers (DS and ML), low-pass filtered at 4.[illegible] kHz, and sampled at 10 kHz. DS is a native speaker of Montreal English, while ML is a native speaker of New York English. Synthetic versions of sentences 1 through 10 were produced by the MITalk synthesizer using a polynomial pulse for voiced excitation (Rosenberg, 1971).

Sentences produced by DS were hand-segmented by the author in a nonoverlapping manner: Transitional regions were not labelled but only that nuclear region of each phone which was considered to belong to that phone inalienably. Figure 5-1 is a spectrographic example of a hand-produced, nonoverlapping segmentation of the word lawyer in sentence 8, spoken by DS and segmented by the author. The short duration of the unlabelled transitional region between [l] and [O] implies that in the opinion of the transcriber, the correct segmentation point of the [10] boundary is temporally well defined. The transitional regions at the boundaries [Oj] and [jr] are much longer, implying more indeterminacy of temporal location.

Sentences produced by ML were hand-segmented by P. Boissonneault, following the same procedure. Figure 5-2, a hand-segmented spectrogram of ML´s production of the phrase the goose (sentence 1), shows how stop closures and bursts were segmented: The closure includes the whole silent portion except for relatively small transitional periods; the burst excludes any following aspiration. Figure 5-2 also illustrates how vowel/fricative boundaries were handled in the hand segmentation: The vowel was considered to end when any of the formants ceased to be excited; the fricative began when strong frication was evident.

Despite careful agreement on segmentatión criteria and comparison of partial results, one major difference in segmentation techique is evident from the data: Paul Boissonneault tended to allow somewhat shorter transitional regions as compared with those of the author, as shown in Table 5-1. Average durations of various types of hand-transcribed segments for the two speakers are given in Table 5-2.

TABLE 5-1. Average durations (ms) and standand deviations of transitional regions in hand transcribed sentences for speakers DS and ML. Transitions with silence are omitted. Speaker DS was transcribed by the author, while ML was transcribed by P. Boissonneault. N is the number of transitional regions.

| | N | AveDur | StdDev |
|---|---|---|---|
| Speaker DS | 305 | 29.4 | 43.7 |
| Speaker ML | 314 | 19.2 | 27.9 |
| Both speakers | 619 | 24.2 | 50.5 |

TABLE 5-2. Average durations (ms) of hand-transcribed segments for speakers DS and ML. N is number of segments.

| | Speaker DS | | speaker ML | |
|---|---|---|---|---|
| class | N | AveDur | N | AveDur |
| vowel | 90 | 77.0 | 91 | 87.7 |
| fricative | 67 | 63.6 | 72 | 67.6 |
| liquid | 36 | 58.8 | 36 | 71.1 |
| nasal | 15 | 64.9 | 15 | 55.3 |
| closure | 54 | 46.9 | 55 | 55.0 |
| glide | 11 | 34.8 | 11 | 53.4 |
| burst | 42 | 5.6 | 44 | 9.6 |

During hand segmentation, the transcribers used the same symbols as used by the rule-based segmenter. They were not allowed to insert or modify symbols, but they could delete symbols they judged to be absent from the natural speech. For example, the schwa in the word <u>from</u>

(sentence 1) and the initial [h] of which (sentence 3) were deleted for both speakers. In all, eleven segments were deleted in the hand segmentations of DS´s sentences, while in ML´s sentences, four segments were deleted. Segments deleted in the hand transcriptions were not included in the evaluation of segmenter performance.

Synthetic utterances were segmented and labelled using the rule-based segmenter (Chapt. 3), which also provided endpoints for each sentence. Endpoints of the naturally produced sentences were determined from the beginning and end of the hand segmentations by adding 10 ms of signal to each end. Synthetic utterances were preprocessed to derive, for each frame, the first seven mel-frequency cepstrum coefficents, A Hanning analysis window was used of 25.6 ms duration. Frame advance was 5.0 ms. Natural utterances were preprocessed in a similar manner, except that, as in the experiment described in Chapt. 4, the frame advance of each natural sentence was modified so that the total number of frames between sentence endpoints would be equal to that of the corresponding synthetic sentence.

Synthetic and natural utterances were time-aligned using the symmetric Zip algorithm (Chamberlain and Bridle, 1983). The Zip algorithm is a suboptimal version of the dynamic time-warping procedure in which the number of cumulative distances retained by the algorithm may not exceed a specified maximum of contiguous values along a diagonal of the warp space. This maximum, termed the diagonal length, was fixed at 50 for the first experiment. The grid topology is nondecimated, differing from that used in experiments described earlier. Therefore, to achieve approximately the same local constraint of a slope between 1/3 and 3, a maximum of two consecutive horizontal or vertical transitions are allowed. The cost penalty of a nondiagonal transition is 1/2 the local distance at the destination. Table 5-3 summarizes the experimental conditions.

TABLE 5-3. Experimental Conditions Employed in the Main Experiment.

| | |
|---|---|
| Horizontal transition penalty: | (1/2) (local distance at destination) |
| Vertical transition penalty: | (1/2) (local distance at destination) |
| Horizontal skip penalty: | infinite |
| Vertical skip penalty: | infinite |
| Max. consecutive horiz. trans: | 2 |
| Max. consecutive vert. trans: | 2 |
| Diagonal length (frames): | 50 |
| Anal. window length synthetic: | 25.6 ms (Hanning) |
| Anal. frame advance synthetic: | 5.0 ms |
| Anal. window length natural: | 25.6 ms (Hanning) |
| Anal. frame advance natural: | adjusted between 5.0 and 6.4 ms<br>to yield same # of frames as synthetic |
| Local distance measure: | Euclidean, using c1 through c7 |

The contiguous segmentations of the synthetic sentences produced by the rule-based segmenter were mapped across the warp paths produced by Zip to induce contiguous segmentations on the natural sentences. The following section discusses the resulting automatic segmentations of the natural utterances.

## 5.2 Evaluation of Segmentation Results

Since the automatic segmentations were contiguous, that is, the beginning of one segment corresponds to the end of the previous segment, the average durations of the automatically transcribed segments are normally longer than those of the hand-transcribed models. Table 5-4 gives these average durations for speakers DS and ML. The only exception occurs for DS´s nasals: Automatically derived duration is slightly less than that of the hand-transcribed nuclear region. This is due to alignment errors of the type seen in Chapt. 4 (Fig. 4-1b).

TABLE 5-4. Comparison of average segmental durations for hand (noncontiguous) and automatic (contiguous) segmentations.

| | | SPEAKER DS | |
|---|---|---|---|
| class | N | Ave Dur Hand | Ave Dur Automatic |
| vowel | 90 | 77.0 | 123.7 |
| glide | 11 | 34.8 | 62.0 |
| liquid | 36 | 58.8 | 93.1 |
| nasal | 15 | 64.9 | 63.1 |
| fricative | 67 | 63.6 | 89.0 |
| closure | 54 | 46.9 | 56.3 |
| burst | 42 | 5.6 | 20.8 |

| | | SPEAKER ML | |
|---|---|---|---|
| class | N | Ave Dur Hand | Ave Dur Automatic |
| vowel | 91 | 87.7 | 114.8 |
| glide | 11 | 53.4 | 54.6 |
| liquid | 36 | 71.1 | 99.9 |
| nasal | 15 | 55.3 | 64.7 |
| fricative | 72 | 67.6 | 85.0 |
| closure | 55 | 55.0 | 65.5 |
| burst | 44 | 9.6 | 19.4 |

The evaluation criterion described in Sect. 4-3 was applied to the twenty automatically segmented sentences using the hand segmentations as models. The results are shown in Table 5-5.

TABLE 5-5. Performance of the automatic segmentation algorithm. N is the total number of boundaries. %Correc is the percentage of boundaries correctly located. EarlyN is the number of boundaries positioned too early. LateN is the number of boundaries positioned too late. EarlyAve is the average error in milliseconds of the early boundaries. LateAve is the average error in milliseconds of the late boundaries. Ave Err is the average absolute error over all boundaries, correct and incorrect.

| | N | %Correc | EarlyN | LateN | EarlyAve | LateAve | Ave Err |
|---|---|---|---|---|---|---|---|
| Speaker DS | 325 | 48.0 % | 83 | 86 | 16.8 | 18.5 | 9.2 |
| Speaker ML | 334 | 41.9 % | 112 | 82 | 13.9 | 12.7 | 7.8 |
| Total | 659 | 44.9 % | 195 | 168 | 15.1 | 15.6 | 8.5 |

As can be seen in the table, the automatic segmentation algorithm achieves a correct segmentation rate of approximately 45 percent. It performs somewhat better on DS than on ML, but this may be due to the application of a more liberal hand segmentation policy to DS (see Sect. 5.1). The comparable result in Table 4-1 for the experiment described in Chapt. 4 is 68 percent correct segmentation. The difference can be attributed to the use of a more rigorous evaluation technique present experiment. Although the error rate is high, the size of the average magnitude error for incorrect boundaries is only around 15 ms, less than one fifth of the average segment duration.

In order to provide a direct comparison with the results discussed in Chapt. 4, we classify segment boundaries as follows: boundaries betweeen two nonsonorant segments (including silence), boundaries between two sonorant segments, and boundaries between sonorant and nonsonorant segments in either order. The result of this analysis is given in Table 5-6.

TABLE 5-6. Percent correct segmentation by segment category. Parenthesized values indicate the total number of boundaries in the category. Category sonorant/nonsonorant also includes nonsonorant/sonorant boundaries. Silence is included in the category of nonsonorant segments. Rates from Table 4-1 are shown in comparable format. Parenthized values give number of boundaries in each class.

| | DS | ML | Total<br>DS + ML | From Ch.4<br>Table 4-1 |
|---|---|---|---|---|
| nonsonorant/nonsorant | 24% (90) | 35% (97) | 30% (187) | 60% (20) |
| sonorant/sonorant | 42% (69) | 39% (69) | 41% (138) | 69% (32) |
| sonorant/nonsonorant | 63% (166) | 47% (168) | 55% (334) | 87% (72) |

It is clear that although the absolute error rates in the current experiment are substantially different than those obtained in the preliminary experminent of Chapt. 4, the ordering of errors by boundary category is identical: Boundaries most prone to error are nonsonorant/nonsonorant boundaries, those least prone to error are sonorant/nonsonorant boundaries, with sonorant/sonorant boundaries intermediate.

To identify the classes of segments contributing to the high error rate at nonsonorant/nonsonorant boundaries, Table 5-7 further analyzes that boundary type.

TABLE 5-7. Analysis of segmenter performance at nonsonorant/nonsonorant boundaries. Boundary types are listed in order of numerical importance.

| | | speaker DS | | speaker ML | | TOTAL | |
|---|---|---|---|---|---|---|---|
| left | right | N | %Correc | N | %Correc | N | %Correc |
| closure | burst | 42 | 4.8 % | 44 | 18.2 % | 86 | 11.6 % |
| fricative | closure | 9 | 22.2 % | 10 | 20.0 % | 19 | 21.1 % |
| fricative | fricative | 9 | 0.0 % | 9 | 44.4 % | 18 | 22.2 % |
| burst | fricative | 6 | 16.7 % | 9 | 33.3 % | 15 | 26.7 % |
| fricative | silence | 6 | 100.0 % | 8 | 100.0 % | 14 | 100.0 % |
| closure | fricative | 5 | 40.0 % | 6 | 16.7 % | 11 | 27.3 % |
| silence | fricative | 5 | 100.0 % | 5 | 100.0 % | 10 | 100.0 % |
| closure | closure | 3 | 33.3 % | 3 | 0.0 % | 6 | 16.7 % |
| silence | closure | 3 | 100.0 % | 3 | 100.0 % | 6 | 100.0 % |
| burst | silence | 2 | 0.0 % | 0 | | 2 | 0.0 % |
| TOTAL | | 90 | 24.4 % | 97 | 35.1 % | 187 | 29.9 % |

From Table 5-7 it is seen that segmentation performance is not uniformly poor within the nonsonorant/nonsonorant boundary class: In particular, the segmenter does well in identifying boundaries of the types silence/fricative and fricative/silence. The high performance on silence/closure boundaries requires special interpretation: These were utterance-initial stop closure segments whose onset is indeterminate. The hand-segmentation convention, therefore, was to make initial closures very short so that they would always be correct unless boundaries were placed after the onset of the burst.

The most important factor in the poor performance of the automatic segmenter on nonsonorant/nonsonorant boundaries is its performance at closure/burst boundaries. We hypothesize that the difficulty here results from the short transitional region provided in the hand segmentation between the end of the stop closure and the beginning of the stop burst. Since burst onsets are well defined in time, short transitional regions are appropriate. To check this hypothesis, average transition region times and standard deviations are given in Table 5-8 for nonsonorant/nonsonorant boundaries (excluding boundaries with silence).

TABLE 5-8. Average durations (ms) and standard deviations of transitional regions in hand-segmented models.

| | | Speaker DS | | | Speaker ML | | |
|---|---|---|---|---|---|---|---|
| left | right | N | AveDur | StdDev | N | AveDur | StdDev |
| closure | burst | 42 | 4.9 | 14.6 | 44 | 5.0 | 13.4 |
| fricative | closure | 9 | 13.2 | 19.0 | 10 | 7.9 | 12.6 |
| fricative | fricative | 9 | 20.6 | 30.2 | 9 | 43.1 | 46.5 |
| burst | fricative | 6 | 20.0 | 43.1 | 9 | 20.8 | 36.6 |
| closure | fricative | 5 | 19.0 | 18.7 | 6 | 8.1 | 10.0 |
| closure | closure | 3 | 34.4 | 38.9 | 3 | 23.3 | 24.1 |

A strong relationship is observed between percent correct segmentation and average transition region duration for closure/burst, fricative/closure, and closure/fricative boundaries. For example, the synthesize-and-warp algorithm misses 88 percent of the closure/burst boundaries, which have an average transition region of 5 ms. Since the analysis window has an effective length of 12.8 ms and the window advance varies from 5 to 6.4 ms, time resolution of the system appears inadequate for location of such boundaries. One possible solution would be to use a shorter analysis window and/or a shorter frame advance.

The relationship between transitional region duration and segmenter performance breaks down when boundaries are compared which involve large differences in the degree to which their left and right-hand phones are acoutically dissimilar. For example, the closure/fricative boundary has a high degree of acoustic dissimilarity, while that of burst/fricative boundary is low. In Tables 5-7 and 5-8, we see that even though DS´s transitional regions are essentially equal in average duration for these two classes, segmentation of closure/fricative is substantially better. We conclude that two components contribute to segmentation performance. The stronger of the two appears to be a priori probability of correct segmentation, which is proportional to transition region duration. A secondary component is related to acoustic similarity.

Table 5-9 gives an analysis by boundary type of the sonorant/sonorant class introduced earlier in Table 5-6.

TABLE 5-9. Analysis of Sonorant/Sonorant Boundary Performance. N is the total number of boundaries in the class. %Correc is the percentage of those boundaries correctly located by the automatic segmenter. Earl is the number of errors in which the automatically determined boundary was too early. Late is the number of errors in which the automatically determined was too late. EAve is the average magnitude of the early errors (ms). LAve is the average magnitude of the late errors (ms). Tot Err is the average magnitude error over all N occurrences of the specified boundary (ms). AveTran is the average transition region duration (ms).

Speaker DS

| left | right | N | %Correc | Earl | Late | EAve | LAve | Tot Err | AveTran |
|---|---|---|---|---|---|---|---|---|---|
| liquid | vowel | 20 | 35.0 % | 6 | 7 | 11.0 | 18.6 | 9.8 | 28.7 |
| vowel | liquid | 13 | 38.5 % | 3 | 5 | 14.4 | 9.0 | 6.8 | 38.5 |
| vowel | nasal | 11 | 45.5 % | 0 | 6 | 0.0 | 29.8 | 16.2 | 29.9 |
| glide | vowel | 10 | 30.0 % | 3 | 4 | 4.2 | 11.6 | 5.9 | 16.0 |
| nasal | vowel | 5 | 60.0 % | 1 | 1 | 15.1 | 55.5 | 14.1 | 43.9 |
| vowel | vowel | 2 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 55.8 |
| vowel | glide | 2 | 0.0 % | 0 | 2 | 0.0 | 7.8 | 7.8 | 30.6 |
| liquid | glide | 2 | 50.0 % | 1 | 0 | 33.2 | 0.0 | 16.6 | 50.2 |
| liquid | nasal | 2 | 50.0 % | 0 | 1 | 0.0 | 45.4 | 22.7 | 9.5 |
| nasal | glide | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 45.1 |
| glide | liquid | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 75.1 |

Speaker ML

| left | right | N | %Correc | Earl | Late | EAve | LAve | Tot Err | AveTran |
|---|---|---|---|---|---|---|---|---|---|
| liquid | vowel | 20 | 40.0 % | 5 | 7 | 26.4 | 9.8 | 10.0 | 28.6 |
| vowel | liquid | 14 | 28.6 % | 8 | 2 | 27.9 | 26.5 | 19.7 | 26.4 |
| vowel | nasal | 11 | 54.5 % | 2 | 3 | 27.4 | 17.9 | 9.9 | 12.5 |
| glide | vowel | 10 | 40.0 % | 5 | 1 | 11.2 | 7.3 | 6.3 | 16.8 |
| nasal | vowel | 5 | 40.0 % | 2 | 1 | 14.2 | 15.2 | 8.7 | 32.9 |
| vowel | vowel | 2 | 50.0 % | 0 | 1 | 0.0 | 7.1 | 3.6 | 68.1 |
| vowel | glide | 2 | 50.0 % | 0 | 1 | 0.0 | 1.8 | 0.9 | 35.9 |
| liquid | glide | 2 | 50.0 % | 1 | 0 | 14.9 | 0.0 | 7.5 | 12.5 |
| liquid | nasal | 2 | 0.0 % | 1 | 1 | 2.2 | 36.6 | 19.4 | 13.5 |
| glide | liquid | 1 | 0.0 % | 0 | 1 | 0.0 | 9.3 | 9.3 | 8.5 |

| TOTAL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| left | right | N | %Correc | Earl | Late | EAve | LAve | Tot Err | AveTran |
| liquid | vowel | 40 | 37.5 % | 11 | 14 | 18.0 | 14.2 | 9.9 | 28.6 |
| vowel | liquid | 27 | 33.3 % | 11 | 7 | 24.2 | 14.0 | 13.5 | 32.2 |
| vowel | nasal | 22 | 50.0 % | 2 | 9 | 27.4 | 25.8 | 13.0 | 21.2 |
| glide | vowel | 20 | 35.0 % | 8 | 5 | 8.6 | 10.7 | 6.1 | 16.4 |
| nasal | vowel | 10 | 50.0 % | 3 | 2 | 14.5 | 35.4 | 11.4 | 38.4 |
| vowel | vowel | 4 | 75.0 % | 0 | 1 | 0.0 | 7.1 | 1.8 | 61.9 |
| vowel | glide | 4 | 25.0 % | 0 | 3 | 0.0 | 5.8 | 4.4 | 33.3 |
| liquid | glide | 4 | 50.0 % | 2 | 0 | 24.1 | 0.0 | 12.0 | 31.4 |
| liquid | nasal | 4 | 25.0 % | 1 | 2 | 2.2 | 41.0 | 21.1 | 11.5 |
| glide | liquid | 2 | 50.0 % | 0 | 1 | 0.0 | 9.3 | 4.7 | 41.8 |
| nasal | glide | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 45.1 |

Excluding of the four vowel/vowel boundaries, there is only a weak correlation between segmentation performance and average transition region length for the sonorant/sonorant boundary. Acoustic dissimilarity effects are also smaller than in the nonsonorant/nonsonorant case, but our distance measure appears to be more sensitive to the differences between vowels and nasals than to those between vowels and liquids or between vowels and glides as evidenced by better segmentation performance on the former.

Table 5-10 gives a breakdown of sonorant/nonsonorant and nonsonorant/sonorant boundaries.

TABLE 5-10. Analysis of sonorant/nonsonorant and nonsonorant/sonorant boundaries. Column labels as in Table 5-9.

Speaker DS

| left | right | N | %Correc | Earl | Late | EAve | LAve | Tot Err | AveTran |
|---|---|---|---|---|---|---|---|---|---|
| vowel | fricative | 36 | 58.3 % | 7 | 8 | 17.5 | 9.1 | 5.4 | 31.8 |
| fricative | vowel | 32 | 62.5 % | 10 | 2 | 12.9 | 4.6 | 4.3 | 24.3 |
| vowel | closure | 26 | 50.0 % | 1 | 12 | 15.9 | 18.3 | 9.0 | 27.0 |
| burst | vowel | 20 | 75.0 % | 5 | 0 | 7.9 | 0.0 | 2.0 | 80.5 |
| burst | liquid | 12 | 83.3 % | 2 | 0 | 4.1 | 0.0 | 0.7 | 43.8 |
| fricative | liquid | 8 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 38.0 |
| liquid | closure | 7 | 42.9 % | 0 | 4 | 0.0 | 13.2 | 7.5 | 21.6 |
| nasal | closure | 6 | 16.7 % | 0 | 5 | 0.0 | 6.9 | 5.8 | 14.3 |
| nasal | fricative | 3 | 33.3 % | 2 | 0 | 43.6 | 0.0 | 29.0 | 17.2 |
| fricative | glide | 3 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 76.7 |
| liquid | fricative | 3 | 33.3 % | 1 | 1 | 13.4 | 0.9 | 4.8 | 33.4 |
| liquid | silence | 2 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 2815.7 |
| closure | nasal | 2 | 50.0 % | 1 | 0 | 2.5 | 0.0 | 1.3 | 34.2 |
| burst | glide | 2 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 82.8 |
| closure | glide | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 27.2 |
| closure | liquid | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 8.5 |
| silence | vowel | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 454.9 |
| silence | liquid | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 498.8 |

Speaker ML

| left | right | N | %Correc | Earl | Late | EAve | LAve | Tot Err | AveTran |
|---|---|---|---|---|---|---|---|---|---|
| vowel | fricative | 36 | 30.6 % | 13 | 12 | 9.5 | 6.0 | 5.4 | 19.0 |
| fricative | vowel | 33 | 57.6 % | 11 | 3 | 6.7 | 8.6 | 3.0 | 13.2 |
| vowel | closure | 26 | 38.5 % | 7 | 9 | 4.9 | 12.8 | 5.8 | 13.9 |
| burst | vowel | 20 | 70.0 % | 6 | 0 | 14.4 | 0.0 | 4.3 | 45.1 |
| burst | liquid | 12 | 58.3 % | 4 | 1 | 8.4 | 0.3 | 2.8 | 32.6 |
| fricative | liquid | 8 | 62.5 % | 1 | 2 | 5.7 | 11.3 | 3.5 | 8.7 |
| liquid | closure | 7 | 14.3 % | 3 | 3 | 7.4 | 8.1 | 6.6 | 8.1 |
| nasal | closure | 6 | 16.7 % | 3 | 2 | 16.6 | 10.6 | 11.8 | 11.3 |
| nasal | fricative | 4 | 0.0 % | 3 | 1 | 20.5 | 15.6 | 19.3 | 4.6 |
| fricative | glide | 4 | 50.0 % | 0 | 2 | 0.0 | 12.0 | 6.0 | 16.6 |
| liquid | fricative | 3 | 33.3 % | 2 | 0 | 24.5 | 0.0 | 16.3 | 12.5 |
| burst | glide | 3 | 66.7 % | 1 | 0 | 2.5 | 0.0 | 0.8 | 46.6 |
| liquid | silence | 2 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 2639.3 |
| closure | nasal | 2 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 59.1 |
| silence | vowel | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 324.0 |
| silence | liquid | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 240.0 |

TOTAL

| left | right | N | %Correc | Earl | Late | EAve | LAve | Tot Err | AveTran |
|---|---|---|---|---|---|---|---|---|---|
| vowel | fricative | 72 | 44.4 % | 20 | 20 | 12.3 | 7.2 | 5.4 | 25.4 |
| fricative | vowel | 65 | 60.0 % | 21 | 5 | 9.7 | 7.0 | 3.7 | 18.6 |
| vowel | closure | 52 | 44.2 % | 8 | 21 | 6.3 | 15.9 | 7.4 | 20.4 |
| burst | vowel | 40 | 72.5 % | 11 | 0 | 11.4 | 0.0 | 3.1 | 62.8 |
| burst | liquid | 24 | 70.8 % | 6 | 1 | 6.9 | 0.3 | 1.7 | 38.2 |
| fricative | liquid | 16 | 81.3 % | 1 | 2 | 5.7 | 11.3 | 1.8 | 23.3 |
| liquid | closure | 14 | 28.6 % | 3 | 7 | 7.4 | 11.0 | 7.1 | 14.9 |
| nasal | closure | 12 | 16.7 % | 3 | 7 | 16.6 | 8.0 | 8.8 | 12.8 |
| nasal | fricative | 7 | 14.3 % | 5 | 1 | 29.7 | 15.6 | 23.4 | 10.0 |
| fricative | glide | 7 | 71.4 % | 0 | 2 | 0.0 | 12.0 | 3.4 | 42.3 |
| liquid | fricative | 6 | 33.3 % | 3 | 1 | 20.8 | 0.9 | 10.5 | 22.9 |
| burst | glide | 5 | 80.0 % | 1 | 0 | 2.5 | 0.0 | 0.5 | 61.1 |
| liquid | silence | 4 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 2727.5 |
| closure | nasal | 4 | 75.0 % | 1 |  | 2.5 | 0.0 | 0.6 | 46.6 |
| silence | vowel | 2 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 411.4 |
| silence | liquid | 2 | 100.0 % | 0 |  | 0.0 | 0.0 | 0.0 | 347.5 |
| closure | glide | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 27.2 |
| closure | liquid | 1 | 100.0 % | 0 | 0 | 0.0 | 0.0 | 0.0 | 8.5 |

With a few exceptions, all transition types in the categories sonorant/nonsonorant and nonsonorant/sonorant are segmented with above-average accuracy. Exceptions are nasal/fricative (14.3 percent), nasal/closure (16.7 percent), liquid/closure (28.6 percent). All these have short average transition regions.

One striking effect evident in Table 5-10 is the asymmetry of performance in segmenting a sonorant followed by closure as opposed to segmenting closure followed by the same sonorant. In every case, performance is at least twice as high when the closure occurs first. Since transitions from stops to sonorants are more rapid than those from sonorants to stops, this result contradicts the general finding that the segmentation algorithm has more difficulty locating rapid events. The result appears to be an artifact of the hand segmentation. Since the hand segmentation was done on spectrograms, the segmentation marker indicating the beginning of stop closure was sometimes placed too early, after the energy of the sonorant had decayed below the grey-level threshold of the spectrogram, but before it had completely decayed. This became evident upon subsequent inspection of the hand segmentation as projected onto the waveform.

## 5.7 An Alternative Evaluation of Alignment Performance

In Sect. 5.2 we saw that the alignment algorithm described does not reliably locate segment boundaries in natural speech. How good a job does it do at finding segment centers? Another way of evaluating the alignment procedure is to check whether or not the temporal center of the automatically derived segment lies within the nuclear region of the segment as defined by the hand segmentation. This is an easier task, and indeed the algorithm performs better at it. Results are shown in Table 5-11.

TABLE 5-11. Performance of the alignment algorithm as a segment center locater for different types of segments. N is total number of segments in named class. %Correc is the percentage of correctly located centers. Earl is the number of errors resulting from placing segment center too early. Late is number of errors having segment center too late. EAve is average magnitude of early errors (ms). LAve is the average magnitude of late errors (ms). Tot Err is the average magnitude of error over all N segments in class (ms). ADur is the average duration (ms) of the nuclear region. Classes are listed in descending order of %Correc.

| | N | %Correc | Earl | Late | EAve | LAve | Tot Err | ADur |
|---|---|---|---|---|---|---|---|---|
| vowels DS | 90 | 85.6 % | 7 | 6 | 43.9 | 55.4 | 11.0 | 77.0 |
| vowels ML | 91 | 93.4 % | 6 | 0 | 52.5 | 0.0 | 3.5 | 87.7 |
| vowels DS+ML | 181 | 89.5 % | 13 | 6 | 74.8 | 55.4 | 7.2 | 82.4 |
| closures DS | 54 | 79.6 % | 3 | 8 | 54.1 | 64.6 | 12.6 | 46.9 |
| closures ML | 55 | 90.9 % | 2 | 3 | 34.9 | 58.0 | 4.4 | 55.0 |
| closures DS+ML | 109 | 85.3 % | 5 | 11 | 46.4 | 62.8 | 8.5 | 51.0 |
| sonorants DS | 152 | 80.9 % | 15 | 14 | 67.0 | 56.6 | 11.8 | 58.5 |
| sonorants ML | 153 | 87.6 % | 15 | 4 | 54.9 | 38.4 | 6.4 | 78.1 |
| sonorants DS+ML | 305 | 84.3 % | 30 | 18 | 61.0 | 52.6 | 9.1 | 73.3 |
| liquid DS | 36 | 77.8 % | 4 | 4 | 45.0 | 64.1 | 12.1 | 58.8 |
| liquids ML | 36 | 83.3 % | 5 | 1 | 64.6 | 61.6 | 10.7 | 71.1 |
| liquids DS+ML | 72 | 80.6 % | 9 | 5 | 55.9 | 63.6 | 11.4 | 65.0 |
| fricatives DS | 67 | 76.1 % | 14 | 2 | 38.8 | 64.5 | 10.0 | 63.6 |
| fricatives ML | 72 | 77.8 % | 10 | 6 | 47.6 | 44.6 | 10.3 | 67.6 |
| fricatives DS+M | 139 | 77.0 % | 24 | 8 | 42.4 | 49.5 | 10.2 | 65.7 |
| all segs. DS | 315 | 70.2 % | 45 | 49 | 45.6 | 50.6 | 14.4 | 55.4 |
| all segs. ML | 324 | 78.4 % | 41 | 29 | 39.5 | 34.6 | 8.1 | 62.6 |
| all segs. DS+ML | 639 | 74.3 % | 86 | 78 | 42.7 | 44.7 | 11.2 | 59.0 |
| nasals DS | 15 | 80.0 % | 1 | 2 | 25.6 | 53.8 | 8.9 | 64.9 |
| nasals ML | 15 | 66.7 % | 3 | 2 | 54.6 | 29.5 | 14.8 | 55.3 |
| nasals DS+ML | 30 | 73.3 % | 4 | 4 | 47.3 | 41.6 | 11.9 | 60.1 |
| glides DS | 11 | 54.5 % | 3 | 2 | 47.6 | 48.0 | 21.7 | 34.8 |
| glides ML | 11 | 81.8 % | 1 | 1 | 21.7 | 33.2 | 5.0 | 53.4 |
| glides DS+ML | 22 | 68.2 % | 4 | 3 | 41.1 | 43.1 | 13.4 | 44.1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| nonsonorants DS | 163 | 60.1 % | 30 | 35 | 34.9 | 48.2 | 16.8 | 43.1 |
| nonsonorants ML | 171 | 70.2 % | 26 | 25 | 30.6 | 34.0 | 9.6 | 48.6 |
| nonsonorants DS+ML | 334 | 65.3 % | 56 | 60 | 32.9 | 42.3 | 13.1 | 45.9 |
| bursts DS | 42 | 9.5 % | 13 | 25 | 26.2 | 41.7 | 32.9 | 5.6 |
| bursts ML | 44 | 31.8 % | 14 | 16 | 17.9 | 25.6 | 15.0 | 9.6 |
| bursts DS+ML | 86 | 20.9 % | 27 | 41 | 21.9 | 35.4 | 23.8 | 7.7 |

Just as DS appeared to be better segmented because the transitional regions in the hand-segmented model were longer, ML´s segment centers appear better located due to longer nuclear regions. There appears to be a rather strong correlation between average nuclear region duration and percent correct center location. The only notable exception is for closures, which yield a higher percentage of correct location than other segments having the same average length. This is explicable because they are acoustically distinct from other types of segments.

In this section we have seen that the alignment algorithm performs much better at locating segment centers than it does at locating segment boundaries. A possible approach to automatic segmentation would use the present algorithm to locate segments and some other, more segment specific approach, to find segment endpoints.

## 5.4 Additional Experiments

The Zip algorithm is a suboptimal version of the dynamic time warping algorithm. As such, we must take care to choose the diagonal length sufficiently large so as to include the optimal time alignment path. To check if this parameter is degrading alignment results, another experiment was run with twice the diagonal length (100 instead of 50 frames). All other aspects of the experiment remained constant. The outcome of the experiment was that the dynamic time warp paths generated using a diagonal of 100 frames were identical to those using 50 frames for all twenty sentences. We conclude that a diagonal of 50 frames is sufficiently large to find the optimum path.

As seen in Chapt. 4, the constrained warp performed substantially better than the unconstrained warp. However, if the warp is too constrained, performance will degrade. An experiment was performed in which the Zip slope constraint was relaxed slightly, without returning to a completely unconstrained warp. The main experiment constrained the slope to be between 1/3 and 3. In this experiment the slope was constrained to lie between 1/4 and 4. All other conditions remained identical to those of the main experiment (see Table 5-3).

The outcome resulted in slightly different warp paths and slightly different induced segmentations. The statistics were close to those in Table 5-5, with segmentation performance on speaker DS improving insignificantly from 48.0 percent to 48.3 percent, while performance on ML worsened from 41.9 percent to 40.1 percent. For both speakers the total error magnitude increased, but more for ML than for DS.

Another possible reason for poor performance of the alignment technique, especially in burst location, may be that the analysis window, 25.6 ms in the main experiment, is too long to give the required time resolution. In another experiment, the analysis window was halved to 12.8 ms, keeping the frame advance and all other conditions as in Table 5-3.

To focus specifically on the three major categories of boundary discussed above, Table 5-12 selectively compares performance on closure/burst, liquid/vowel, and fricative/vowel boundaries for the main experiment (1), relaxed slope constraints (2), and reduced window length (3).

TABLE 5-12. Comparative segmentation performance on three selected boundary types of (1) the algorithm in Table 5-3, (2) looser slope constraints, and (3) shorter analysis window.

Speaker DS

| Exp | left | right | N | %Correc | Earl | Late | EAve | LAve | TotDist |
|---|---|---|---|---|---|---|---|---|---|
| 1 | closure | burst | 42 | 4.8 % | 20 | 20 | 16.6 | 28.7 | 21.5 |
| 2 | closure | burst | 42 | 7.1 % | 20 | 19 | 16.7 | 29.5 | 21.3 |
| 3 | closure | burst | 42 | 11.9 % | 15 | 22 | 20.0 | 29.2 | 22.4 |
| 1 | liquid | vowel | 20 | 40.0 % | 5 | 7 | 16.3 | 25.0 | 12.8 |
| 2 | liquid | vowel | 20 | 40.0 % | 5 | 7 | 16.3 | 25.0 | 12.8 |
| 3 | liquid | vowel | 20 | 25.0 % | 6 | 9 | 13.0 | 16.0 | 11.1 |
| 1 | fricative | vowel | 32 | 62.5 % | 10 | 2 | 12.9 | 4.6 | 4.3 |
| 2 | fricative | vowel | 32 | 68.8 % | 10 | 0 | 15.8 | 0.0 | 4.9 |
| 3 | fricative | vowel | 32 | 68.8 % | 8 | 2 | 17.6 | 15.9 | 5.4 |

Speaker ML

| Exp | left | right | N | %Correc | Earl | Late | EAve | LAve | TotDist |
|---|---|---|---|---|---|---|---|---|---|
| 1 | closure | burst | 44 | 18.2 % | 19 | 17 | 11.6 | 11.1 | 9.3 |
| 2 | closure | burst | 44 | 18.2 % | 19 | 17 | 16.7 | 11.1 | 11.5 |
| 3 | closure | burst | 44 | 18.2 % | 17 | 19 | 14.4 | 11.3 | 10.5 |
| 1 | liquid | vowel | 20 | 40.0 % | 5 | 7 | 26.4 | 9.8 | 10.0 |
| 2 | liquid | vowel | 20 | 30.0 % | 5 | 9 | 28.6 | 10.8 | 12.0 |
| 3 | liquid | vowel | 20 | 45.0 % | 4 | 7 | 29.5 | 8.4 | 8.8 |
| 1 | fricative | vowel | 33 | 57.6 % | 11 | 3 | 6.7 | 8.6 | 3.0 |
| 2 | fricative | vowel | 33 | 57.6 % | 12 | 2 | 7.2 | 14.3 | 3.5 |
| 3 | fricative | vowel | 33 | 48.5 % | 14 | 3 | 7.6 | 26.8 | 5.7 |

On speaker DS, relaxing the slope constraint had the desired effect: Segmentation performance on closure/burst boundaries improved markedly. This was not the case for speaker ML, however, where segmentation performance was not affected. Conversely, the liquid/vowel boundary for speaker DS was unaffected by the looser constraint, while on speaker ML, segmentation at this boundary worsened. Fricative/vowel boundaries for speaker DS are correctly segmented more often under relaxed constraints, but with larger errors for the boundaries that are missed. All these effects are small and probably not significant.

A larger speaker-dependent effect occured in the experiment with a reduced analysis window. The number of correctly located closure/burst boundaries for speaker DS more than doubled when the analysis window length was halved, although the average magnitude of the remaining errors increased, just as in the case of relaxed slope constraint. The shortened window had very little effect on ML's sentences. Conversely, segmentation of DS´s liquid/vowel transitions was worse with the shorter window. Since these sounds involve slowly changing spectra, the use of a longer window should be an advantage. However, for ML´s speech, the shorter window did not degrade liquid/vowel boundary segmentation. As was the case with the slope constraint, the best window length appears to depend upon the speaker.

## 5.5 Summary of Alignment Experiments

In Chapt. 1, we set forth five critical assumptions that must hold in order for the synthesize-and-warp technique to produce correct results consistently. With full knowledge that each of these assumptions held only partially, we proceeded to explore the performance consequences in Chapt. 2. By manipulating experimental conditions, it was possible to isolate the effects of forcing certain of the assumptions to hold. Thus, by employing rule-based segmentation of the synthetic model, it was possible to achieve correct model segmentation (Assumption 3). In Chapt. 4, use of a hand-segmented natural model eliminated errors arising from synthesis quality (Assumption 2). The resulting reduction in error rate from 22 percent to 13 percent indicates that the warp algorithm is indeed sensitive to deficiencies in synthesis quality. At the same time, it is clear from the residual 13 percent error rate that the remaining assumptions, particularly Assumptions 4 and 5, having to do with local distance and the warp algorithm, are also extremely important. In the main experiment (Chapt. 5), only Assumption 3 holds. The resulting 45 percent error indicates that the synthesize-and-warp technique in its present form cannot reliably perform the job of a human transcriber/segmenter.

Assumption 1 states that the phonetic transcription of the model corresponds to what the speaker actually pronounced. Since in the present system the phonetic transcription is deterministically derived from the input word string, Assumption 1 does not hold in general. In fact, the single segmentation error of largest magnitude (220.4 ms) occurred in sentence 5 when speaker DS inserted a pause after the word cape: "She flaps her cape, as she parades the street." Since the synthetic model did not contain a pause, desynchronization occurred, causing surrounding segments to be misplaced as well. Another example of failure of Assumption 1 is the use by MITalk of the preaspirated glide [hw] as the initial phoneme of which and whiff. This dialect feature is absent from the speech of both DS and ML. Thus, dialect variability tends to degrade the automatic segmentation. Finally, when two stop consonants occur together, our phonological rules predict that the first will not be released. In practice, speakers sometimes do release the first stop. In future work on direct rule-based segmentation of natural speech, optional phonological rules will be used to handle this type of variability.

Assumption 4 is difficult to control for, since no "correct" distance metric is known. The consequence of having Assumption 4 not hold is that local distances between corresponding segments of the test and model utterance exceed local distances between noncorresponding segments. Since the warp algorithm always finds the minimum distance path between the test and model utterances, the alignment, and therefore the induced segmentation of the test utterance, will be incorrect. Spectral distance measures often behave incorrectly when the two utterances are from different speakers. In speech recognition, this is called the ´normalization problem´. An analogous problem arises in this case when natural and synthetic speech are aligned in the synthesize-and-warp procedure.

In the dynamic time warping algorithm, slope constraints are placed on the path to exclude unlikely alignments. Assumption 5 states that these constraints are correctly formulated. In Chapt. 4 we saw that removing slope constraints completely has a devasting effect on segmentation (41 percent versus 22 percent error). This result underlines the inadequacy of the distance metric: Local distance alone is insufficient for avoiding gross misalignments. When slope constraints are used, however, small alterations in them do not affect segmentation performance in a significant way (Sect. 5.2). The introduction of slope constraints is only a partial remedy to overcome the limitations of the distance metric used. Further improvements are needed to ensure that the distance metric corresponds more closely with the peroeptual similarity of the corresponding segments.

## 5.6 Conclusions and Possible Solutions

The reason that the synthesize-and-warp procedure does not perform well enough to be practical for automatic labelling of speech databases is that of the five critical assumptions, only Assumption 3 holds consistently. It may be possible to have Assumption 2 hold by using an iterative synthesize-and-warp procedure in which the spectral features of the model are adjusted to match those of the test utterance. To make Assumption 5 hold requires modification of the time-warping algorithm itself. Since the segmentation and labelling of the model utterance is known, warp constraints may be defined as a function of segment type. A possible solution to the problem of phonological variability (Assumption 1) may be to associate with each segment of the model a set of phonological rules, accessible to the warp algorithm. Such rules could specify, for example, that certain segments may be skipped without penalty.

An alternative to the iterative synthesize-and-warp approach, which avoids the five critical assumptions altogether, is suggested by the success of the rule-based segmenter on synthetic speech: Modify the rule-based segmentation algorithm to segment natural speech. Two additional difficulties are present in the rule-based segmentation of natural speech that were not present in the rule-based segmentation of synthetic speech: (1) natural speech is far more variable than synthetic speech and (2) estimation of acoustic parameters from the waveform is errorful. As is true for the iterative synthesize-and-warp procedure described above, the use of a more complex phonology is necessary, specifying which segmental substitutions, deletions, and insertions are likely. However, the rule-based environment is ideally suited to the expression of an elaborate phonology.

Most importantly, a rule-based approach does not encounter the problem of distance metrics (Assumption 4) since no model utterance is used. Instead, we are required to state an acoustic event that coincides with each boundary type, thus focussing directly on the segmentation problem. Acoustic events may be arbitrarily complex, involving conjunction and disjunction of simpler events. The problem of spectral variation can be handled in the context of the rule-based segmenter by specifying events in relative terms. For example, instead of defining the event corresponding to the onset of voicing as was done for synthetic speech, i.e., periodic excitation exceeds some fixed threshold, voicing onset would be defined as an increase of periodicity by a specified amount or proportion within a specified time.

Since perfect warp algorithms and perfect distance metrics do not exist, rule-based segmentation of natural speech potentially offers a more precise segmentation than methods based on warping. Each time a boundary is detected, the required event is known to have occurred. The user is thus assured that the agreed upon definition of a boundary type has been instantiated by an acoustic event in the signal. If an expected event does not occur and its absence cannot be accounted for by a phonological rule, then the boundary is flagged for subsequent human intervention. Experience gained in the rule-based segmentation of synthetic speech indicates that errors tend to be discrete in nature. If a rule is missing or incorrect, a gross segmentation error occurs. Unlike the time-warp-based segmentation, minor errors, corresponding to a fraction of a segment's duration, rarely occur. Gross errors are generally easier to detect and repair than minor errors.

This investigation indicates that the synthesize-and-warp technique does not perform sufficiently well to be used as a speech analysis tool. Until distance metrics, time warping, and synthetic speech quality are improved, a better approach may be the direct, rule-based segmentation of natural speech. By choosing a direct, rule-based approach, Assumptions 2, 3, 4, and 5 are unnecessary. Assumption 1, concerning variable phonological rules, is most conveniently addressed within the rule-based framework. Subsequent research will explore this direction.

The implications of these results for speaker-independent speech recognition are that the distance metric and time-warping procedure require improvement. Better distance metrics would emphasize phonetic differences as opposed to interspeaker differences. More realistic slope constraints in the time-warping procedure would reflect observed temporal variability as a function of segment type.

## REFERENCES

Allen, J., S. Hunnicutt, R. Carlson, and B. Granstrom. 1979. MITalk-79 The 1979 MIT text-to-speech system. In J. Wolf and D.H. Klatt, eds., Speech Communication Papers presented at the 97th Meeting of the Acoustical Society of America, pp. 507-510.

Bridle, J.S. and R.M. Chamberlain. 1983. Automatic labelling of speech using synthesis-by-rule and non-linear time-alignment. Speech Communication 2 (2-3), 187-189.

Chamberlain, R.M. and J.S. Bridle 1983. "Zip: A Dynamic Programming Algorithm for Time-aligning Two Indefinately Long Utterances," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Boston, 816-819.

Cohen, J.R. 1981. Segmenting speech using dynamic programming Journ. Acoust. Soc. Am. 69(5), 1430-1438.

Fant, G. 1973. Speech Sounds and Features. Cambridge, MA The MIT Press.

Gill, G.S., H. Goldberg, R Reddy, and B. Yegnanarayana, 1978. A recursive segmentation procedure for continuous speech. Technical report, Department of Computer Science, Carnegie-Mellon University.

Goldberg, H.G. 1975. Segmentation and labeling of speech. A comparative performance evaluation. Ph.D. Thesis, Deptartment of Computer Science, Carnegie-Mellon University.

Hunt, M.J., M. Lennig, and P. Mermelstein. 1983. Use of dynamic programming in a syllable-based continuous speech recognition system. D. Sankoff and J. Kruskal, eds., Time Warps, String Edits, and Macromolecules· the Theory and Practice of Sequence comparison, New York Addison-Wesley, pp. 163-187.

IEEE. 1969. IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. on Audio and Electroacoustics AU-17(3).

Klatt, D.H. 1980. Software for a cascade/parallel formant synthesizer. Journ. Acoust. Soc. Amer. 67(3), 971-995.

Kozhevnikov, V.A. and L.A. Chistovich. 1965. Speech: Articulation and Perception. Washington: Joint Publications Research Service.

Le Saint-Milon, J. and M. Stella. 1983. Extraction automatique de diphones par programmation dynamique pour la synthèse de la parole. Speech Communication 2 (2-3), 196-198.

Lennig, M. 1983. Automatic alignment of natural speech with a corresponding transcription. Speech Communication 2(2-3), 190-192.

Lennig, M. and Brassard, J.-P. In preparation. A computer-readable phonetic alphabet for English and French.

Mermelstein, P. 1975a. A phonetic-context controlled strategy for segmentation and phonetic labeling of speech. IEEE Trans. on Acoustics, Speech and Signal Processing ASSP-23(1), 79-82.

Mermelstein, P. 1975b. Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58(4), 880-883.

Mermelstein, P. 1978. Recognition of monosyllabic words in continuous sentences using composite word templates. Proc. of the 1978 IEEE Int´l. Conf. on Acoustics, Speech and Signal Processing, pp. 708-711.

Néel, F., M. Eskenazi, and J.J. Mariani 1983. Cadrage automatique pour la constitution de dictionnaires d´entités phonétiques. Speech Communication 2(2-3), 193-195.

Rosenberg, A.E. 1971. Effect of glottal pulse shape on the quality of natural vowels. Journ. Acoust. Soc. Amer. 49(2), 583-590.

Sargent, D.C. 1982. Rhythmic cues aid lip readers. IEEE Spectrum 19(4), 46-49.

Sargent, D.C. and A. Malcolm. 1979. The presentation of continuous speech with synchronous printed text. Proc. of the 1979 IEEE Int´l. Conf. on Acoustics, Speech and Signal Processing, pp. 471-474.

Shoup, June E. 1980. Phonological aspects of speech recognition. Wayne A. Lea, ed., Trends in Speech Recognition, Englewood Cliffs, NJ: Prentice-Hall, p. 127.

Wagner, M. 1981. Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms. Proc. of the 1981 IEEE Int´l. Conf. on Acoustics, Speech and Signal Processing, pp. 1156-1159.

# APPENDIX A PHONETICALLY BALANCED SENTENCES USED IN THIS STUDY

1. The goose was brought straight from the old market.
2. The sink is the thing in which we pile dishes.
3. A whiff of it will cure the most stubborn cold.
4. The facts don't always show who is right.
5. She flaps her cape as she parades the street.
6. The loss of the cruiser was a blow to the fleet.
7. Loop the braid to the left and then over.
8. Plead with the lawyer to drop the lost cause.
9. Calves thrive on tender spring grass.
10. Post no bills on this office wall.

# APPENDIX B: PHONETIC SYMBOLS USED BY MITALK-79

| | | | | |
|---|---|---|---|---|
| IY | meet | WH | which | |
| IH | mit | LX | pal | (postvocalic |
| EY | mate | EL | title | (syllabic) |
| EH | met | Q | fasten | (glottal stop) |
| AE | mat | P | pat | |
| AA | pot | B | bat | |
| AO | salt | M | mat | |
| OW | mold | T | tag | |
| UH | book | D | did | |
| UW | mood | N | none | |
| AH | but | KP | kipper | palatal |
| ER | worker | GP | give | palatal |
| AY | bite | TQ | latin | glottalized |
| OY | boy | K | comb | |
| AW | house | G | gone | |
| YU | use | NG | ring | |
| IXR | pier | CH | church | |
| EXR | pear | J | jug | |
| AXR | march | DX | bottle | (flap) |
| OXR | more | F | far | |
| UXR | moor | V | very | |
| AX | pompous (schwa) | TH | thistle | |
| IX | impunity (barred i) | DH | then | |
| AXP | (nonsyllabic schwa) | S | sink | |
| W | witch | Z | zinc | |
| Y | yellow | SH | shrink | |
| R | rat | ZH | camouflage | |
| L | lit | EM | logarithm | (syllabic) |
| H | hat | EN | button | (syllabic) |
| HX | the hurrah (voiced) | | | |

# FIGURES FOR

# CHAPTER 2

Fig. 2-1a. Synthetic speech with MITalk segmentation and labelling.

Sentence 1: The goose was b(rought straight from the old market).

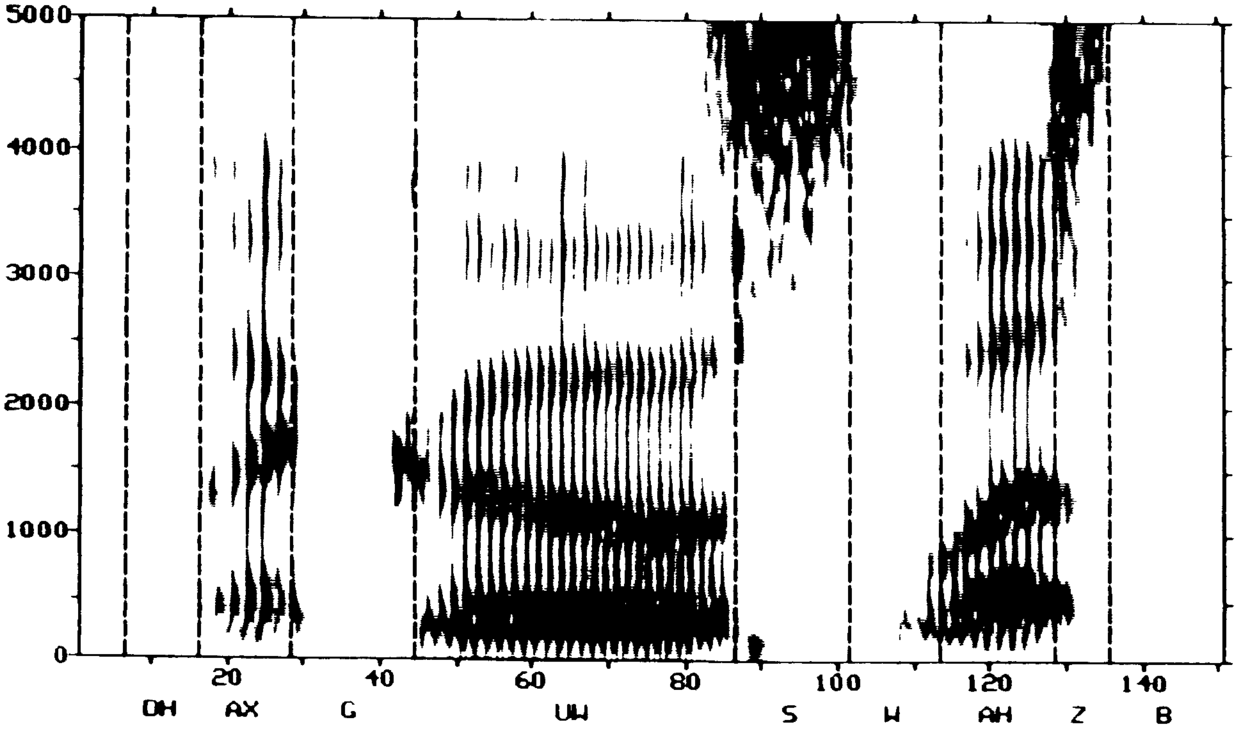


Fig. 2-1b. Synthetic speech with MITalk segmentation and labelling.

Sentence 1: (The goose wa)s brought strai(ght from the old market).

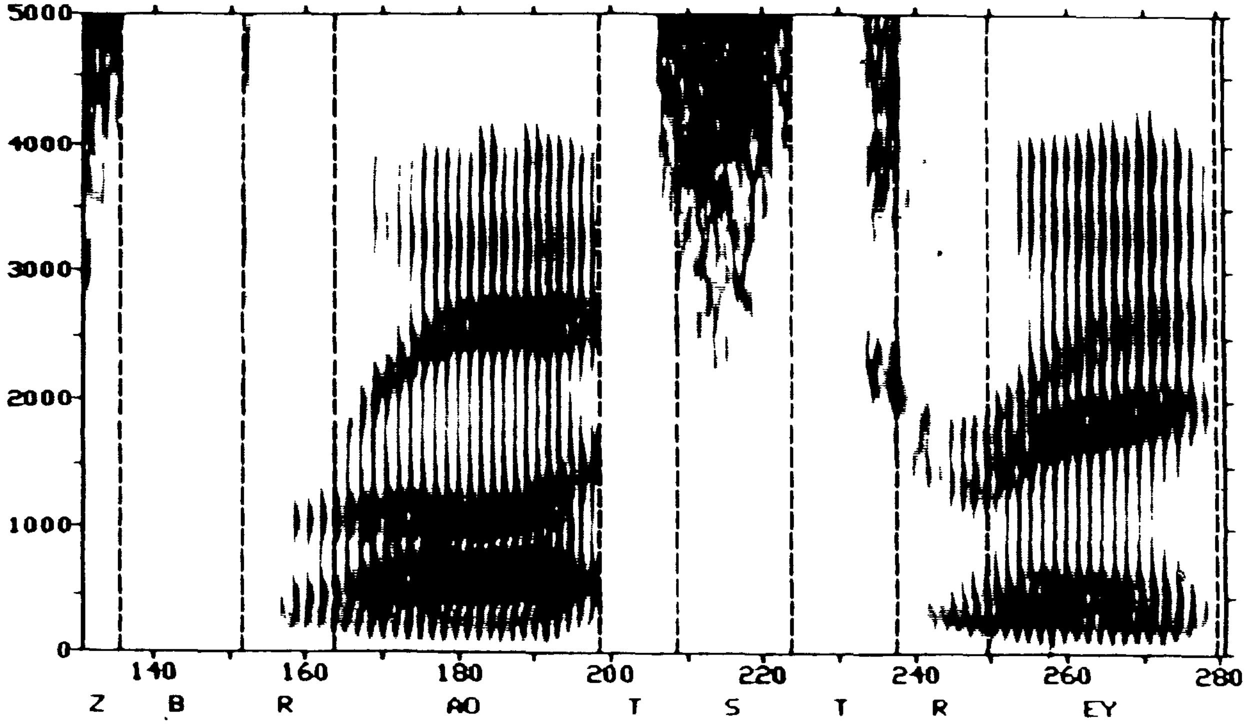


Fig. 2-1c. Synthetic speech with MITalk segmentation and labelling.

Sentence 1: (The goose was brought st)raight from the o(ld market).

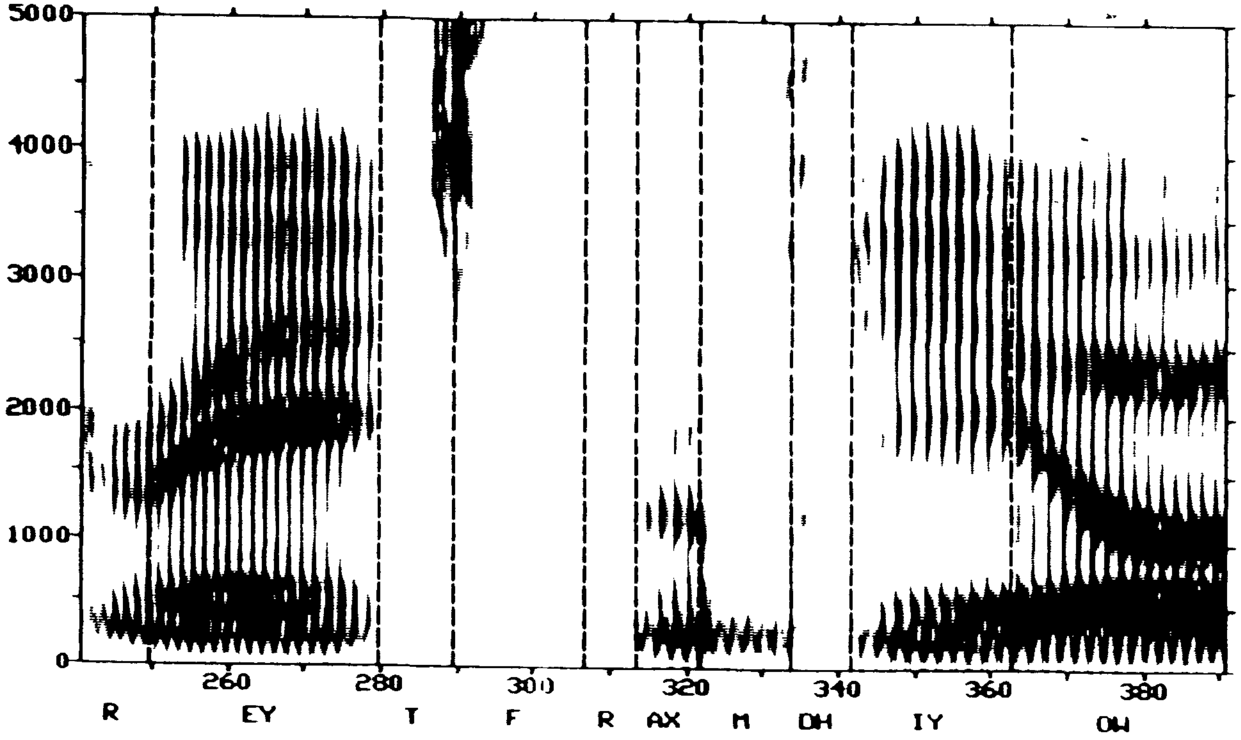


Fig. 2-1d. Synthetic speech with MITalk segmentation and labelling.

Sentence 1: (The goose was brought straight from the) old market.

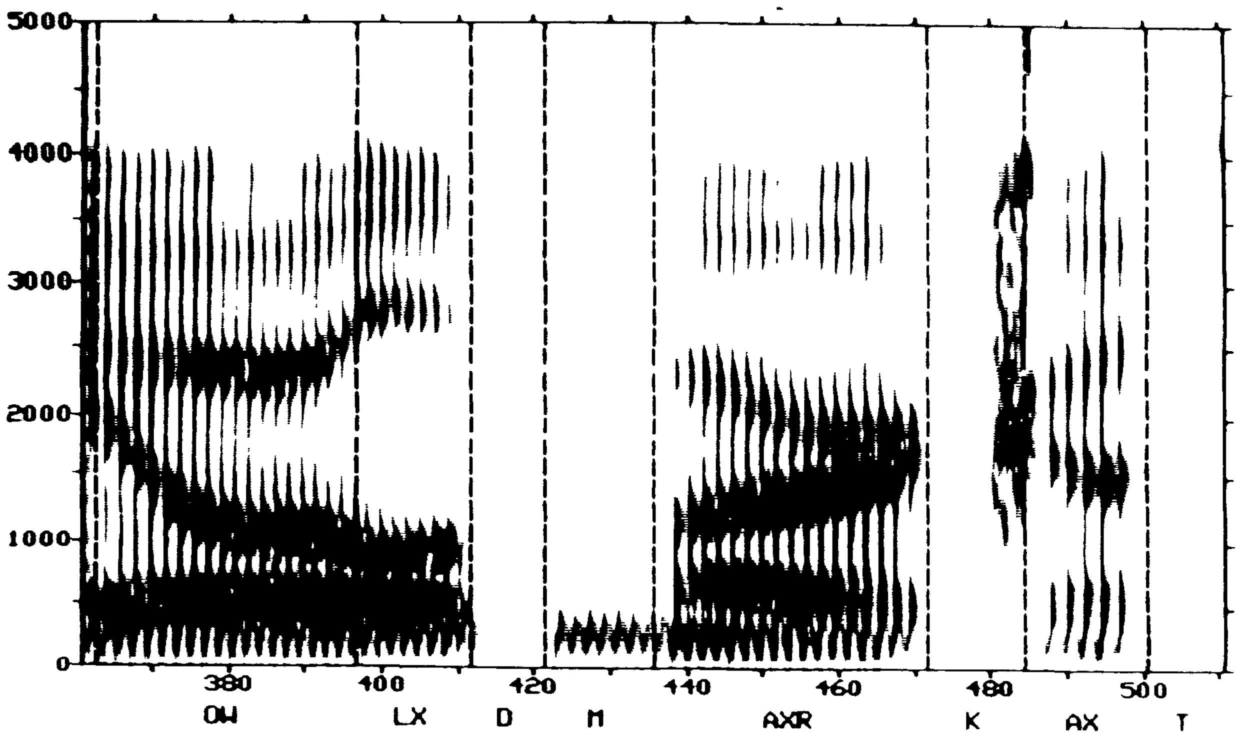


Fig. 2-1e. Synthetic speech with MITalk segmentation and labelling.

Sentence 1: (The goose was brought straight from the old mark)et.

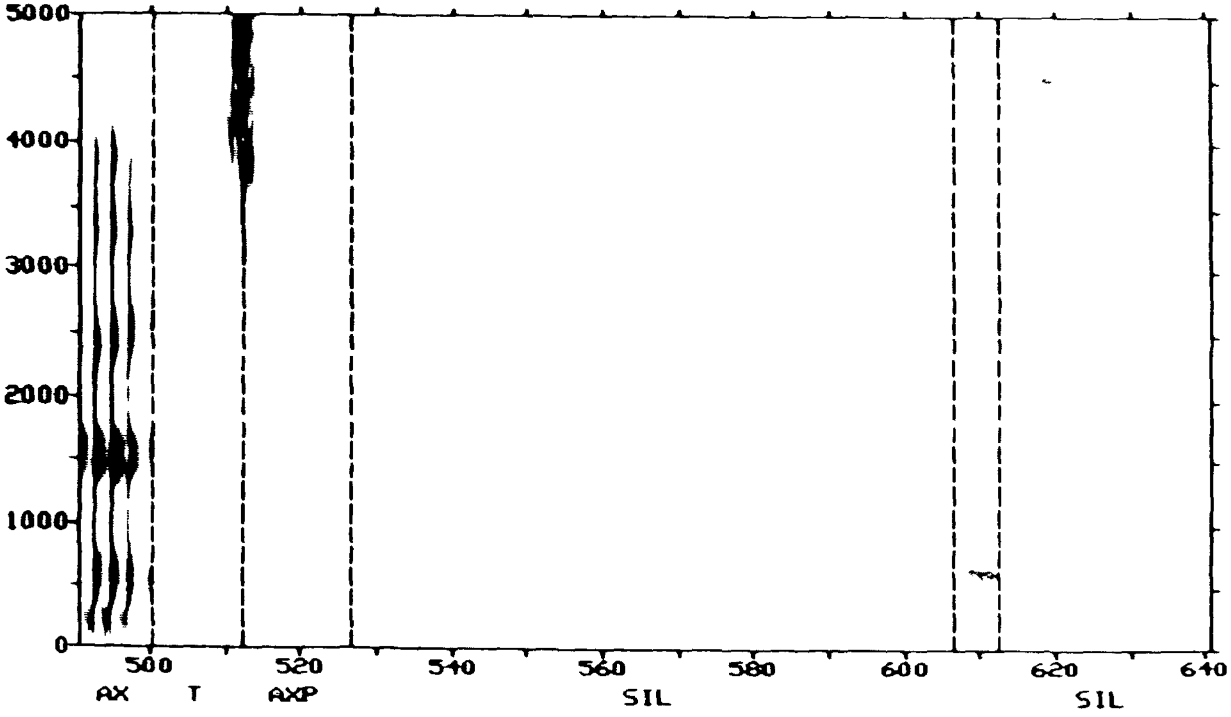


Fig. 2-2a. Synthetic speech with MITalk segmentation and labelling.

Sentence 2: The sink is the thi(ng in which we pile dishes).

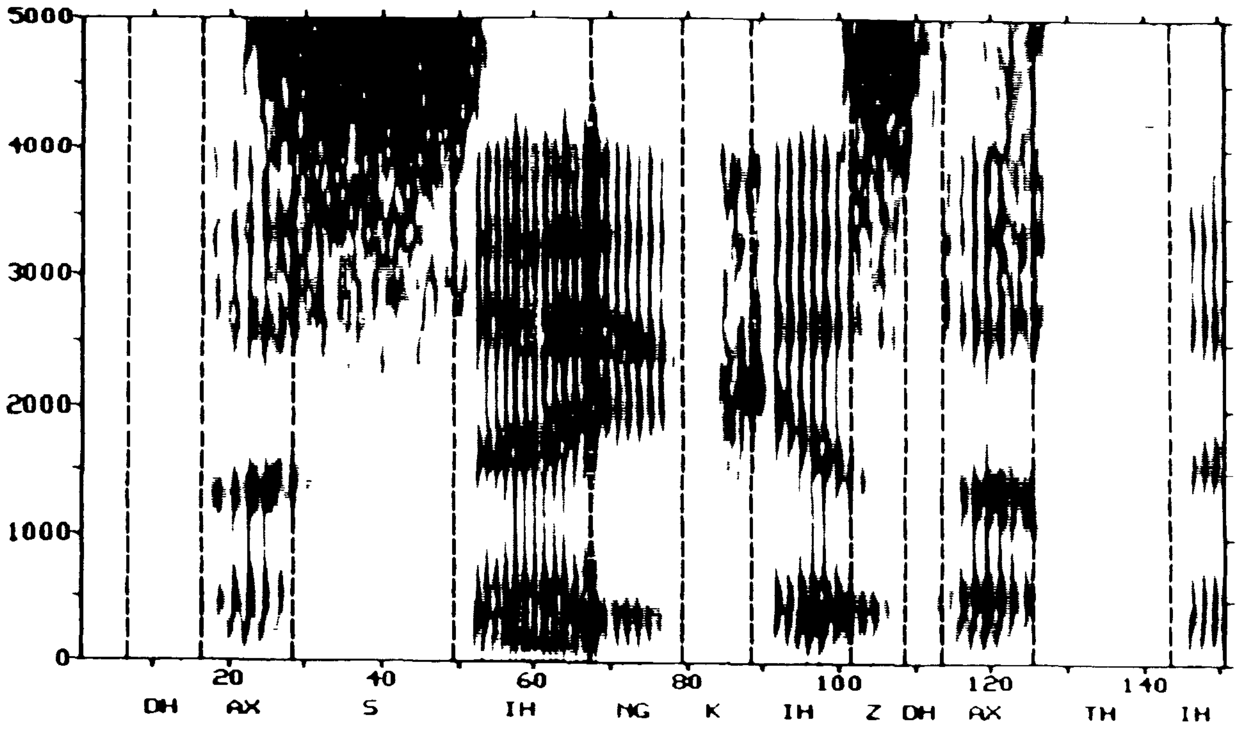


Fig. 2-2b. Synthetic speech with MITalk segmentation and labelling.

Sentence 2: (The sink is the th)ing in which we p(ile dishes).

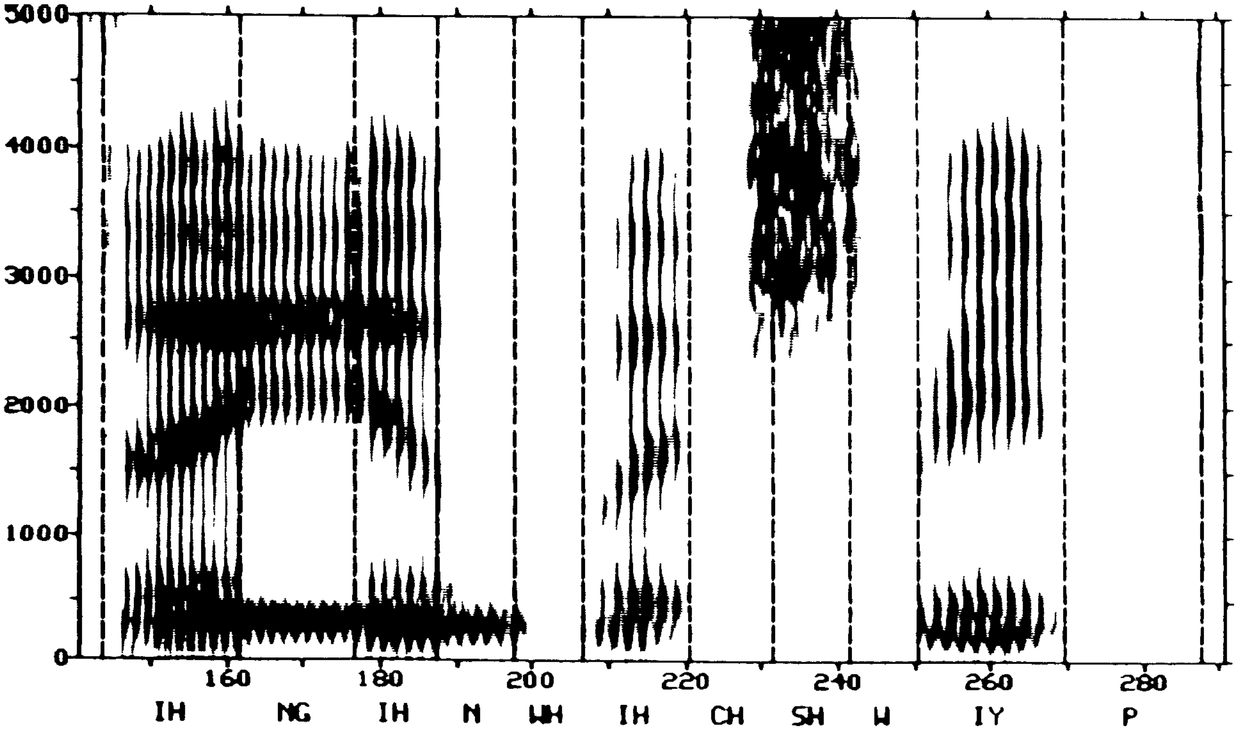


Fig. 2-2c. Synthetic speech with MITalk segmentation and labelling.

Sentence 2: (The sink is the thing in which we) pile dishe(s).

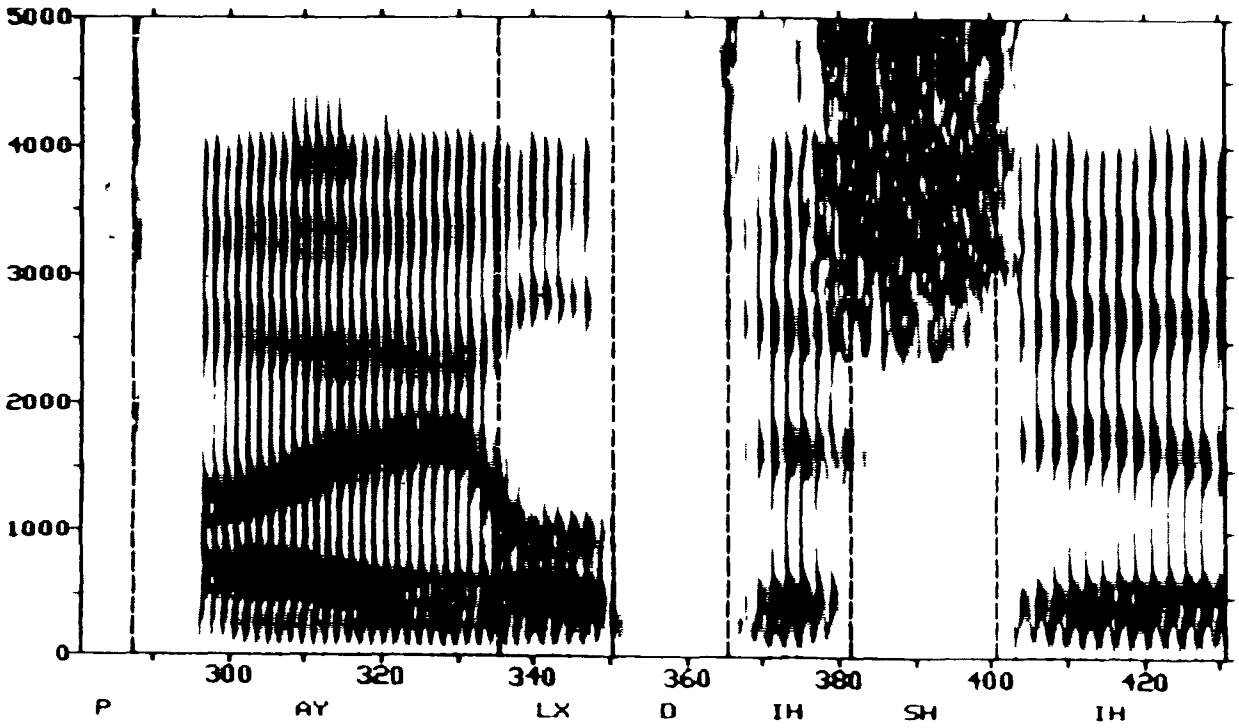


Fig. 2-2d. Synthetic speech with MITalk segmentation and labelling.

Sentence 2: (The sink is the thing in which we pile fish)es.

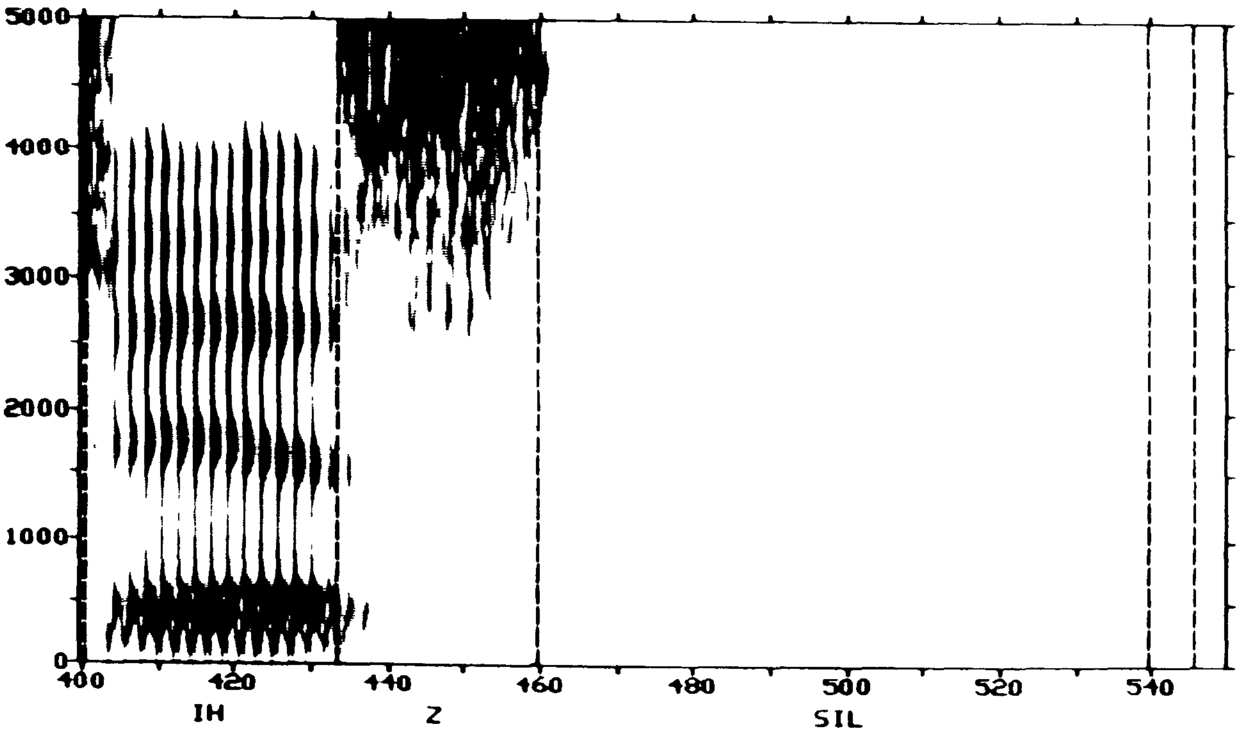


Fig. 2-3a. Synthetic speech with MITalk segmentation and labelling.

Sentence 3: A whiff of it will (cure the most stubborn cold).

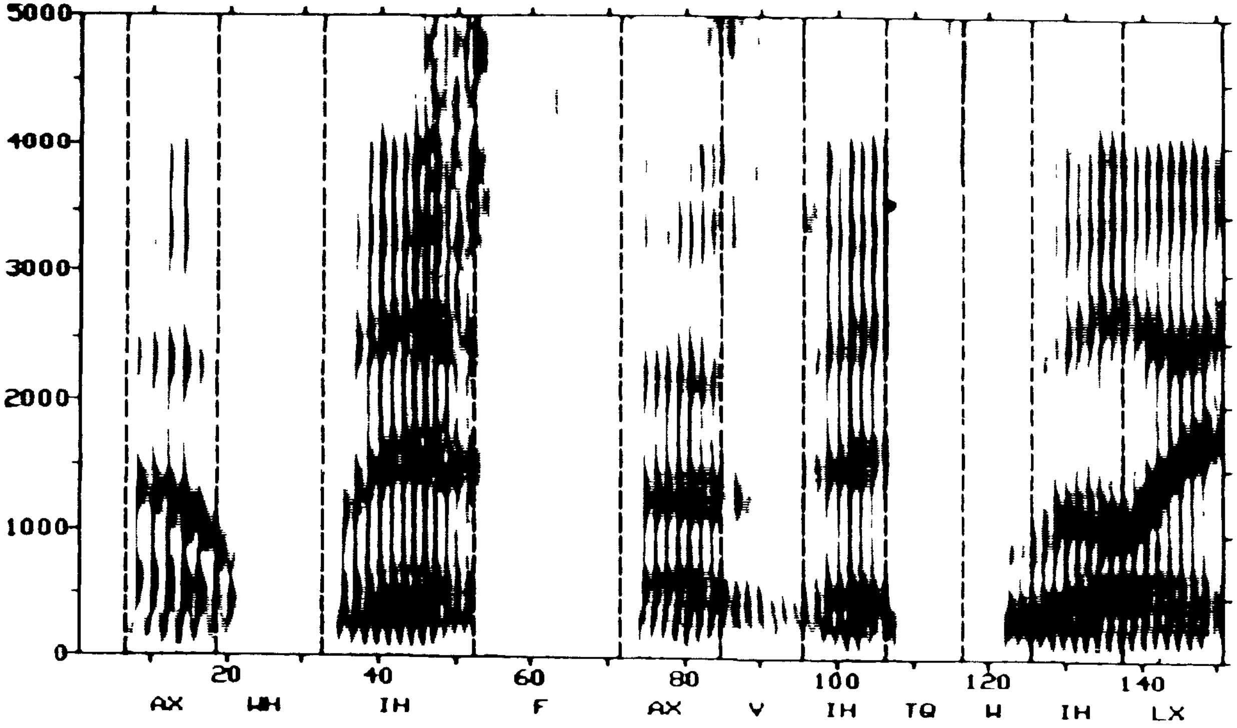


Fig. 2-3b. Synthetic speech with MITalk segmentation and labelling.

Sentence 3: (A whiff of it will cure the most stubborn cold).

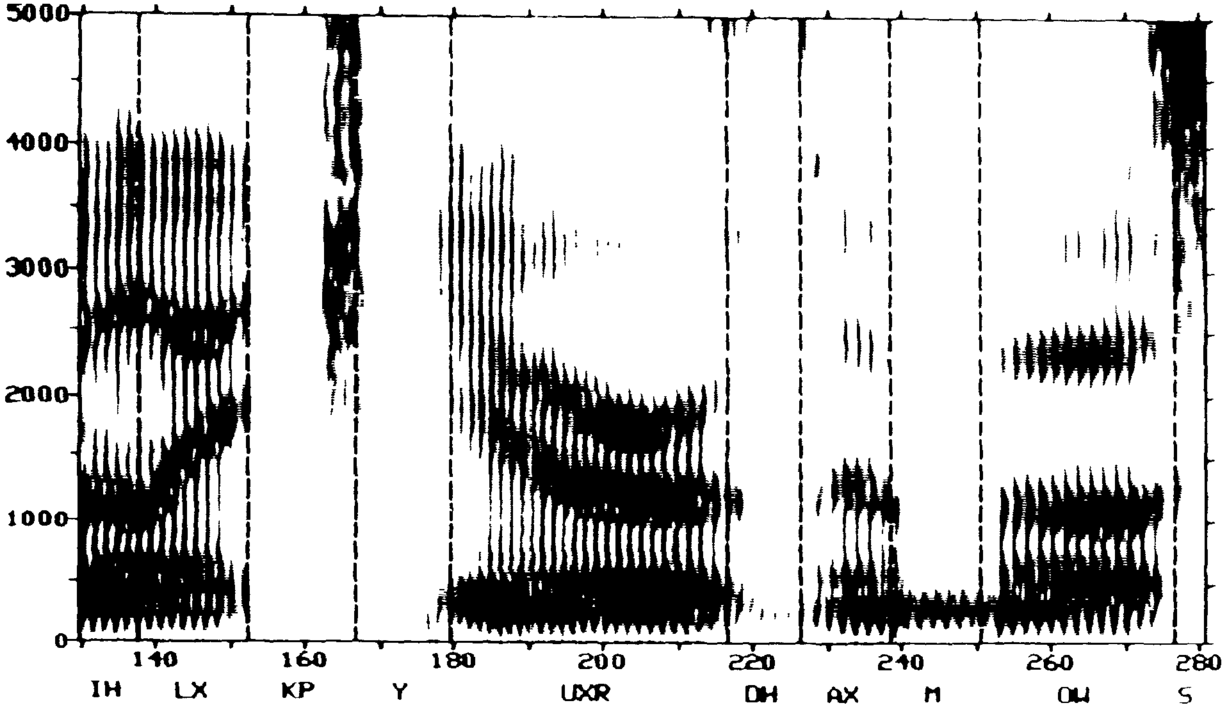


Fig. 2-3c. Synthetic speech with MITalk segmentation and labelling.

Sentence 3: (A whiff of it will cure the most stubborn co(ld).

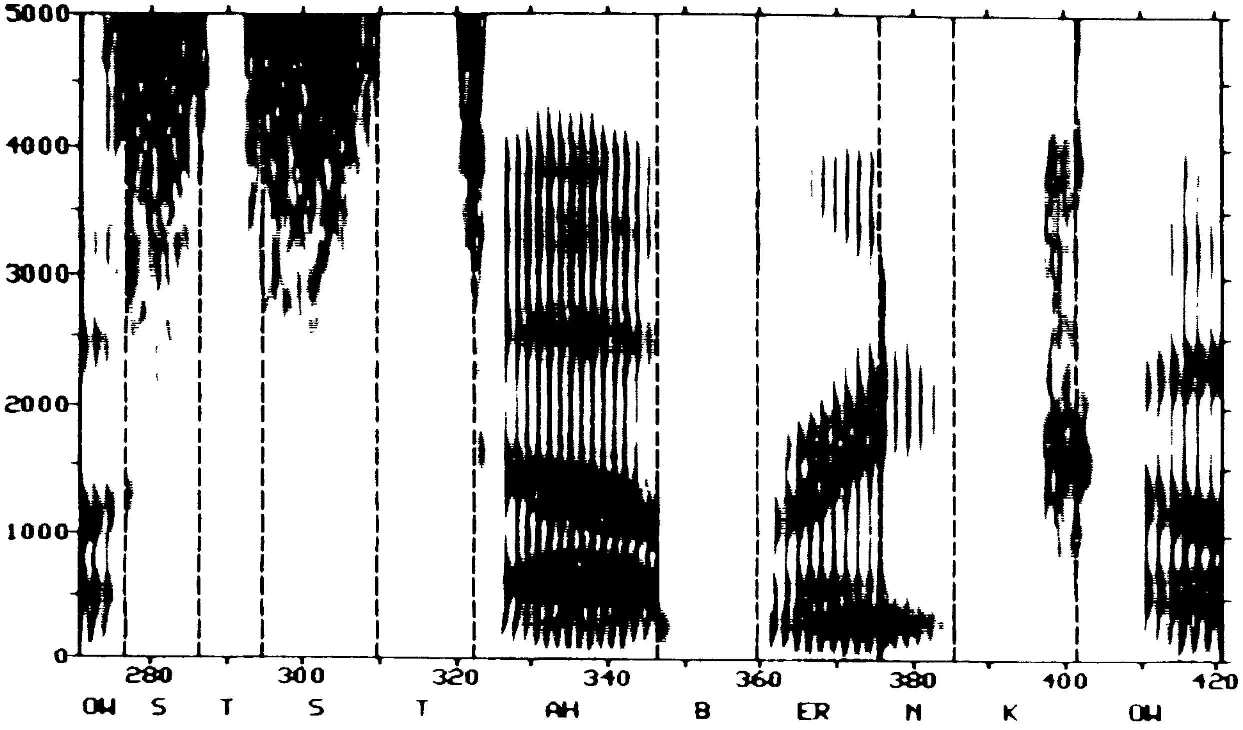


Fig. 2-3d. Synthetic speech with MITalk segmentation and labelling.

Sentence 3: (A whiff of it will cure the most stubborn c)old.

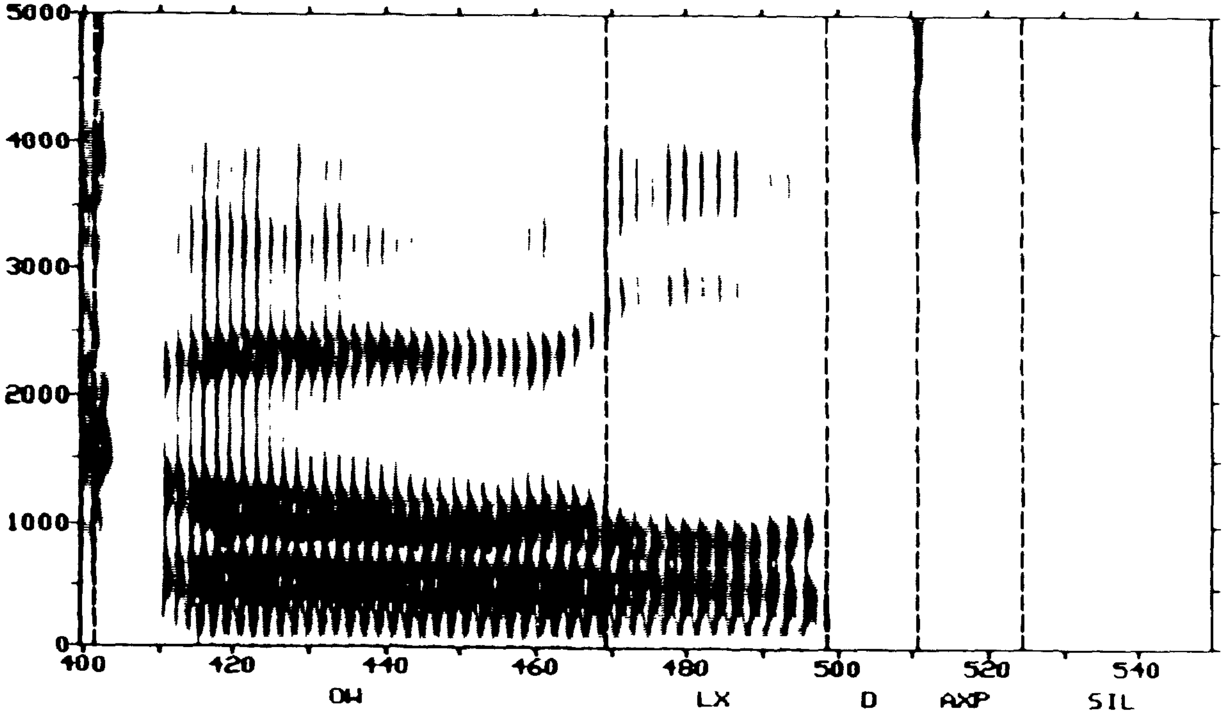


Fig. 2-4a. Synthetic speech with MITalk segmentation and labelling.

Sentence 4: The facts do(n´t always show who is right).

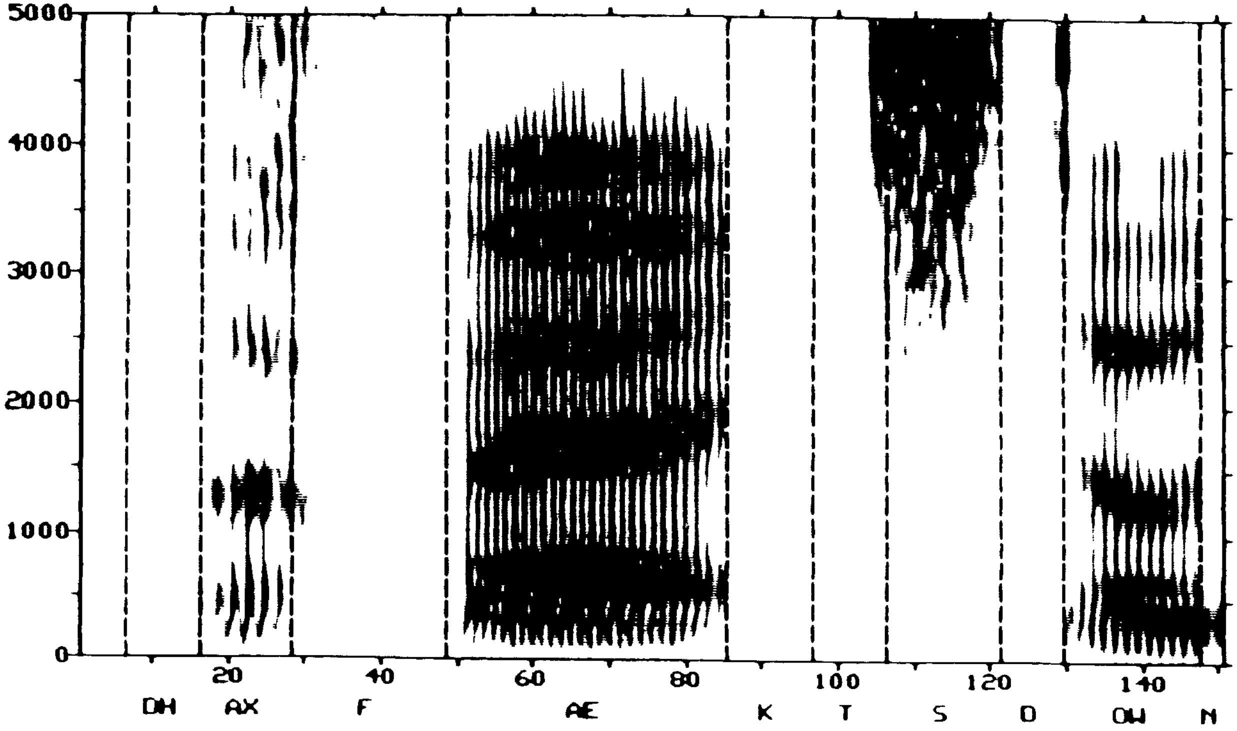


Fig. 2-4b. Synthetic speech with MITalk segmentation and labelling.

Sentence 4: (The facts d)on´t always show (who is right).

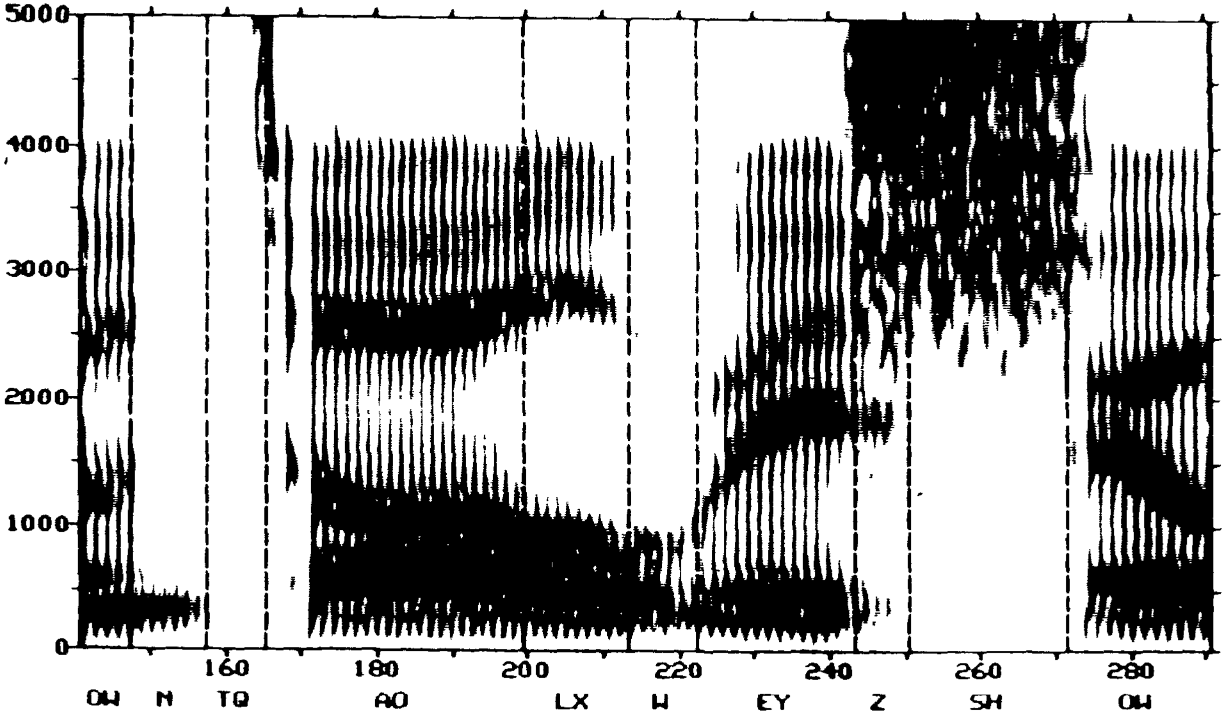


Fig. 2-4c. Synthetic speech with MITalk segmentation and labelling.

Sentence 4: (The facts don´t always sh)ow who is right.

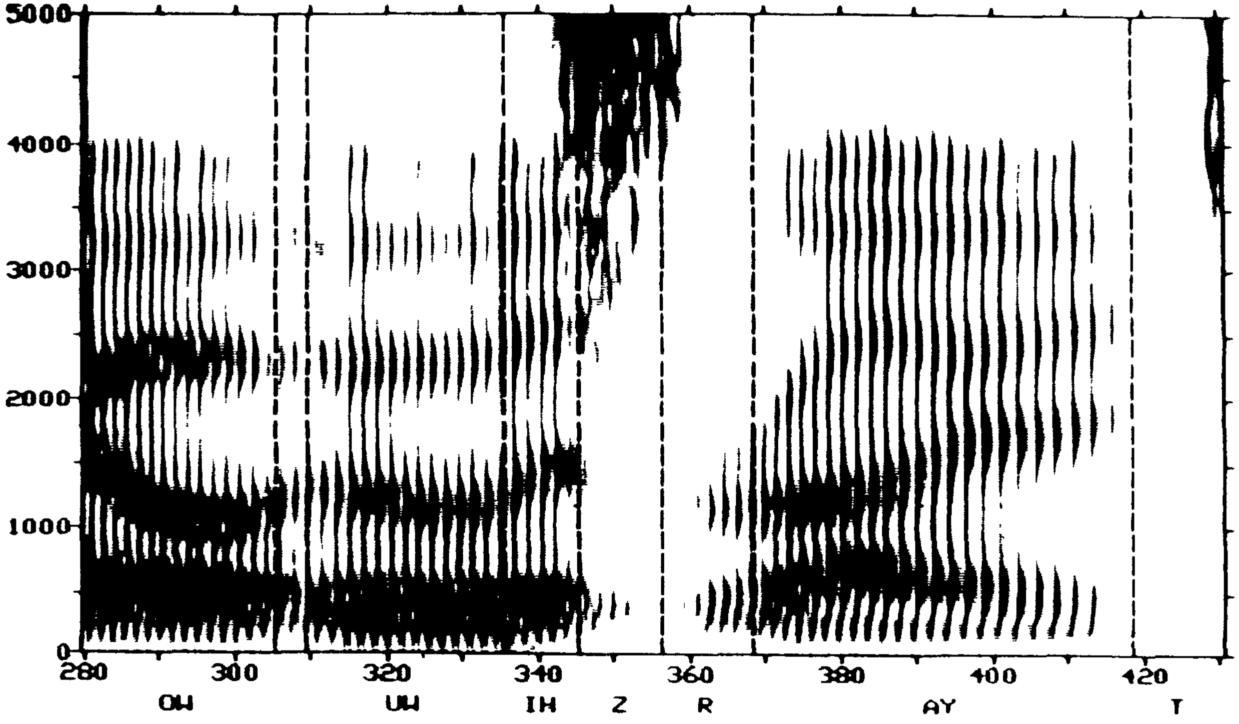


Fig. 2-4d. Synthetic speech with MITalk segmentation and labelling.

Sentence 4: (The facts don´t always show who is r)ight.

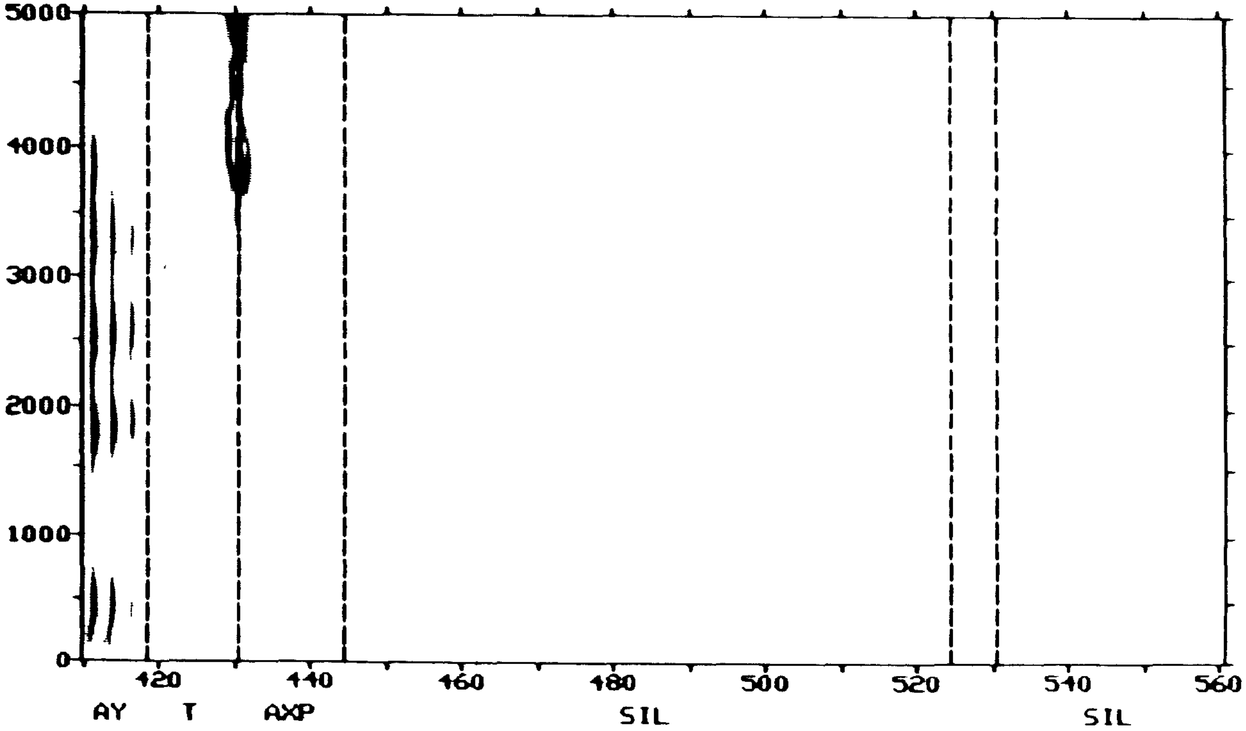


Fig. 2-5a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 1: The goose was b(rought straight from the old market).

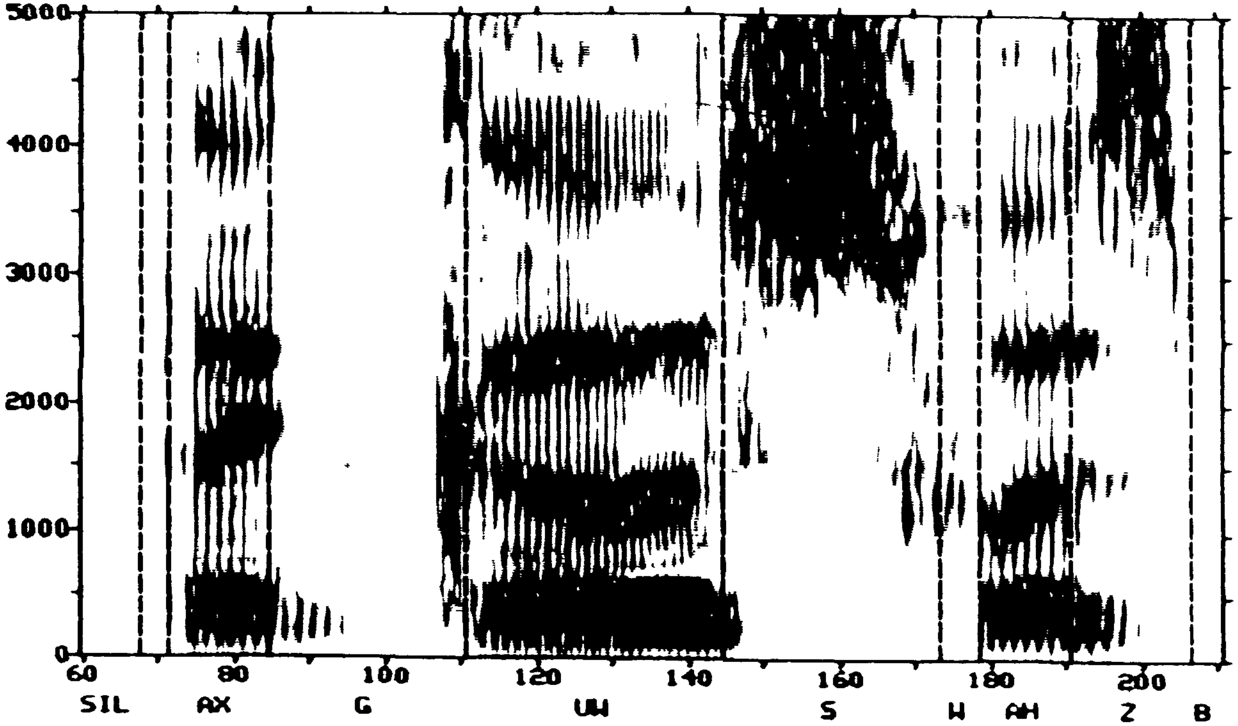


Fig. 2-5b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 1: (The goose wa)s brought strai(ght from the old market).

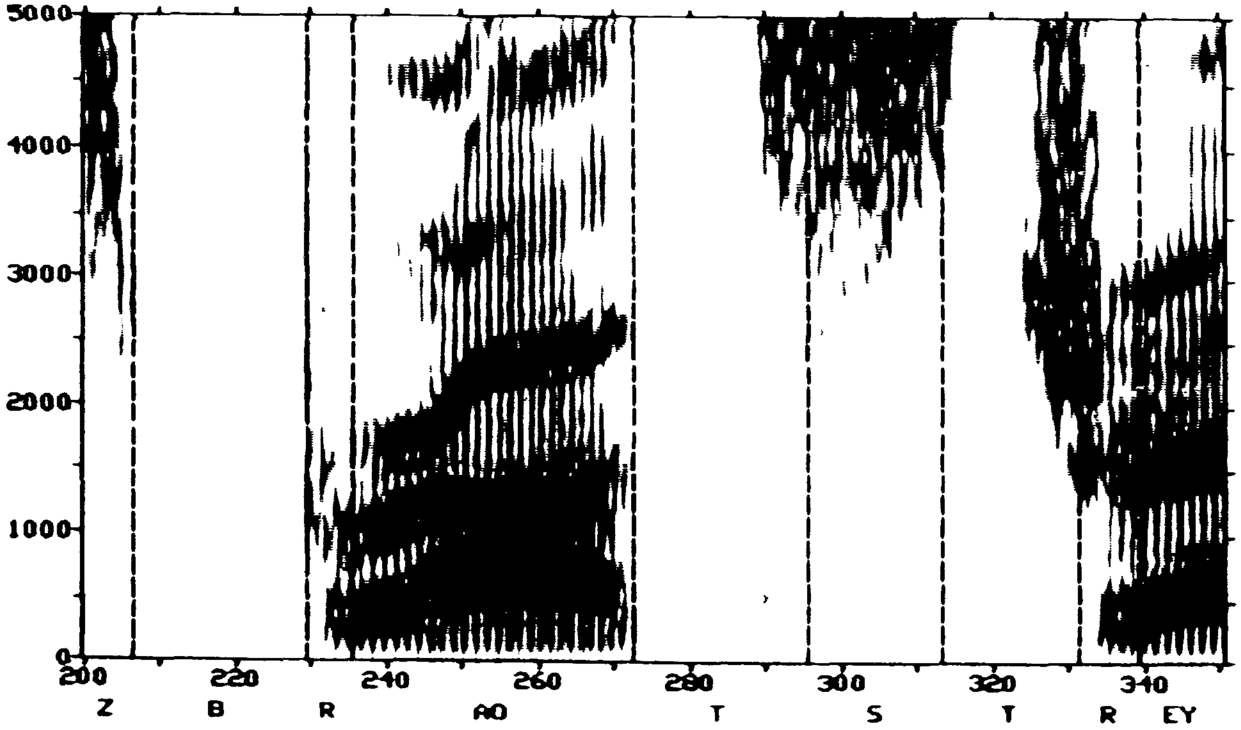


Fig. 2-5c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 1: (The goose was brought str)aight from the o(ld market).

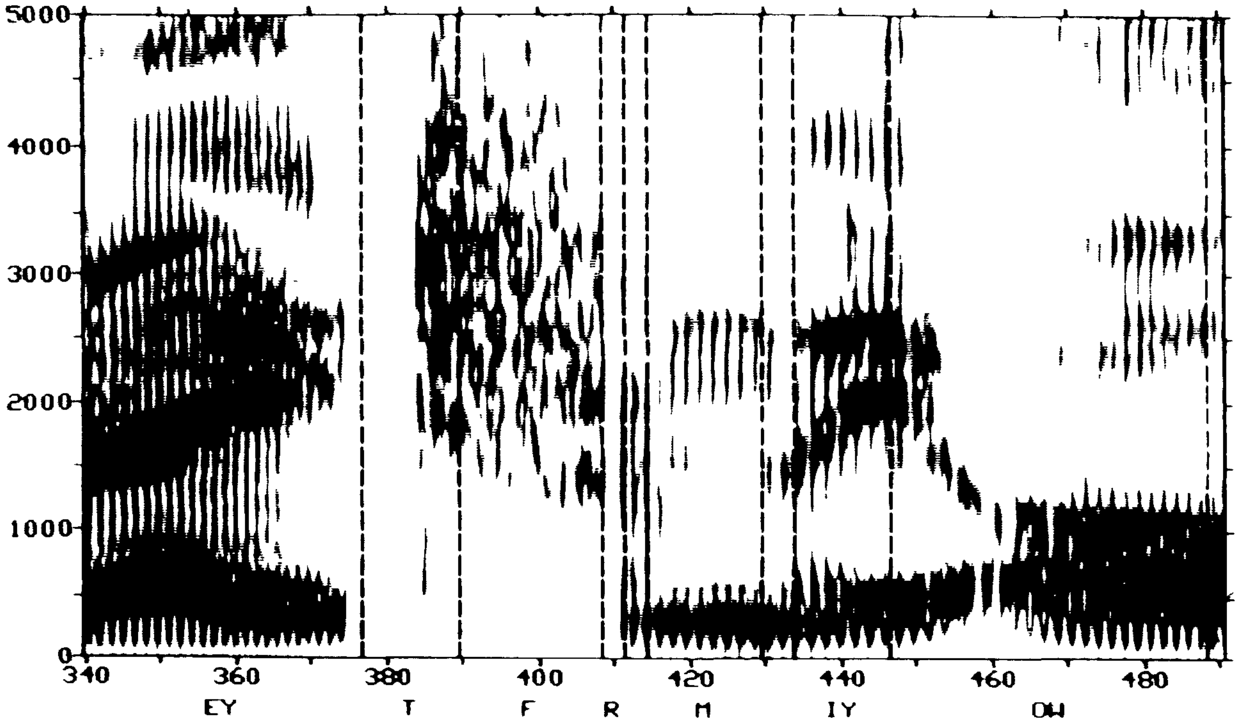


Fig. 2-5d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 1: (The goose was brought straight from the) old marke(t).

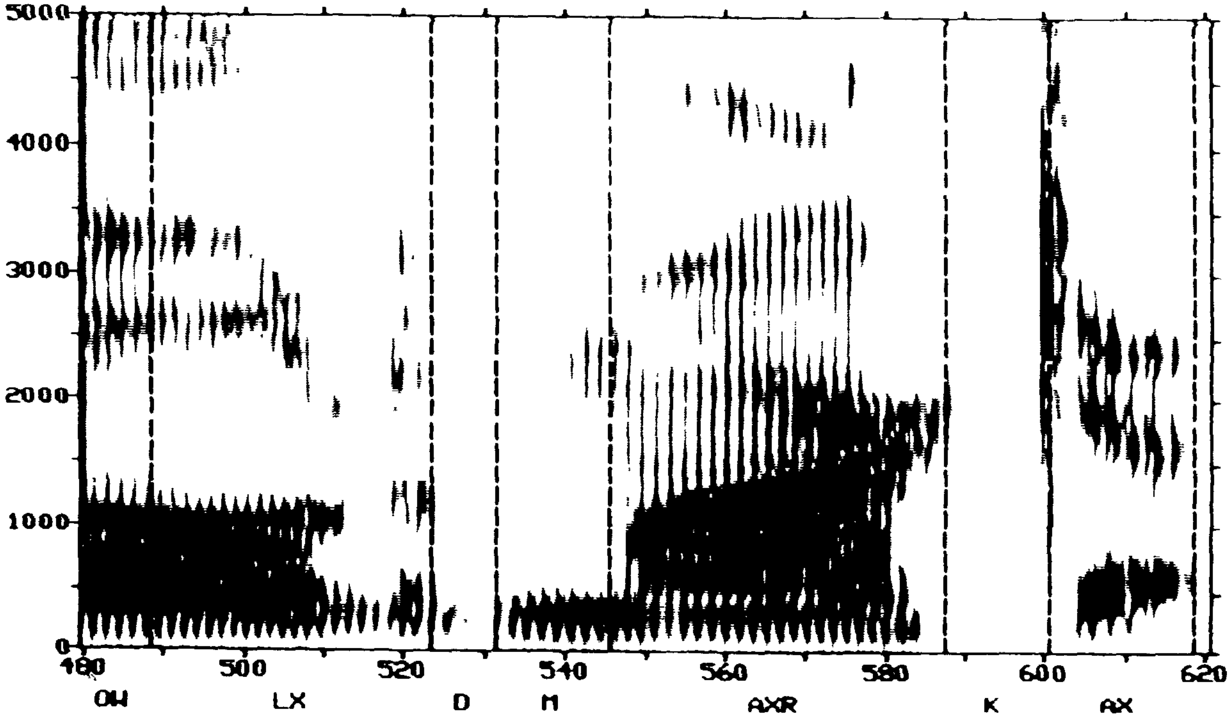


Fig. 2-5e. Speaker ML with induced MITalk segmentation and labelling.

Sentence 1: (The goose was brought straight from the old mark)et.

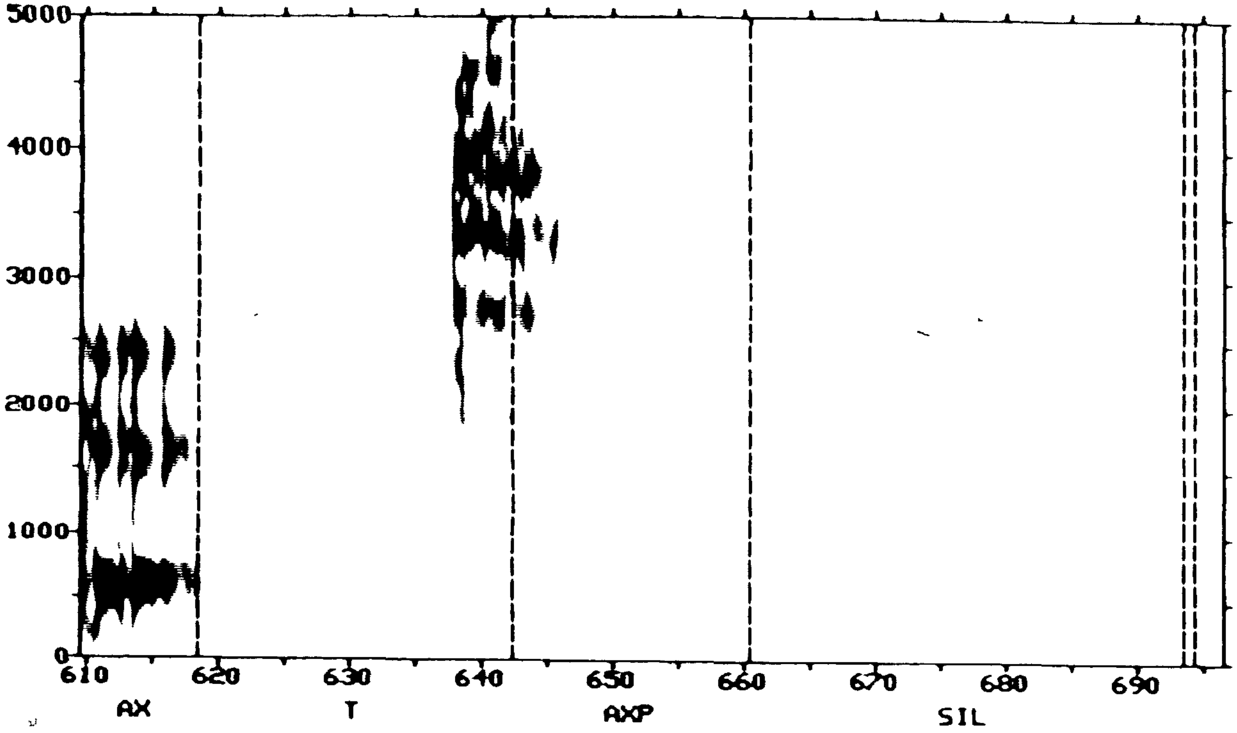


Fig. 2-6a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 2: The sink is the (thing in which we pile dishes).

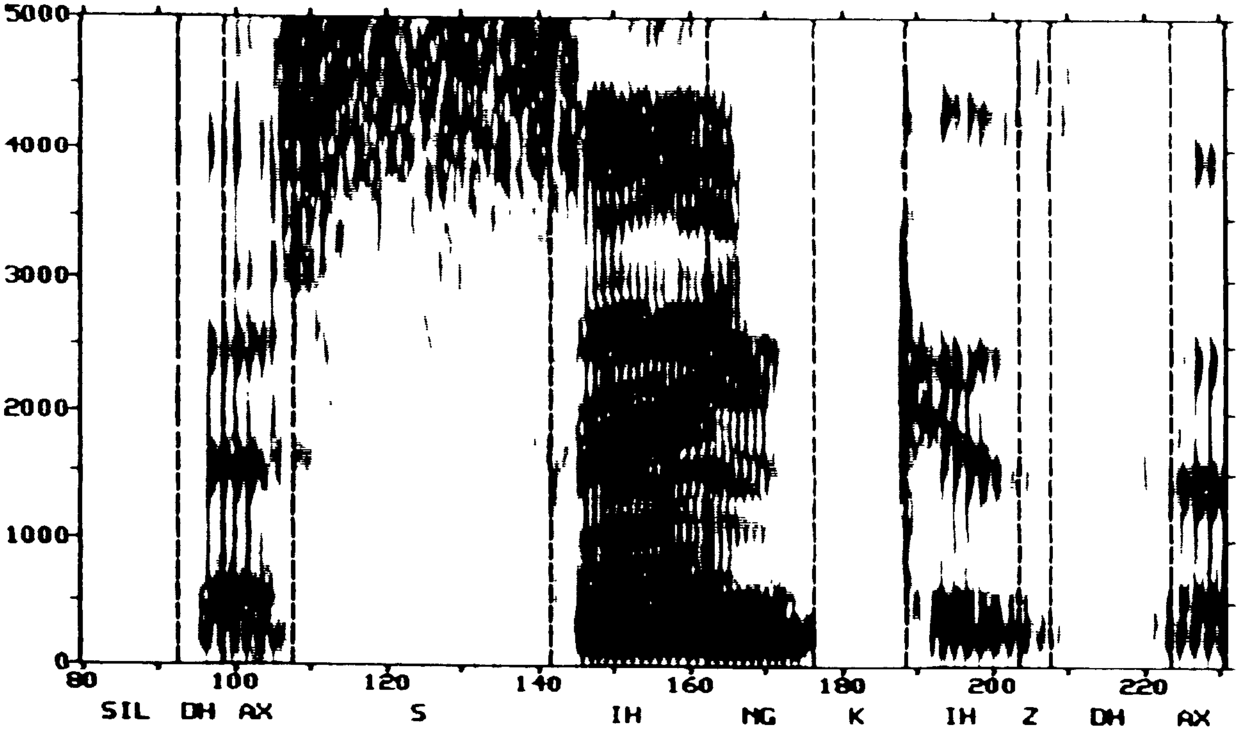


Fig. 2-6b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 2: (The sink is th)e thing in which w(e pile dishes).

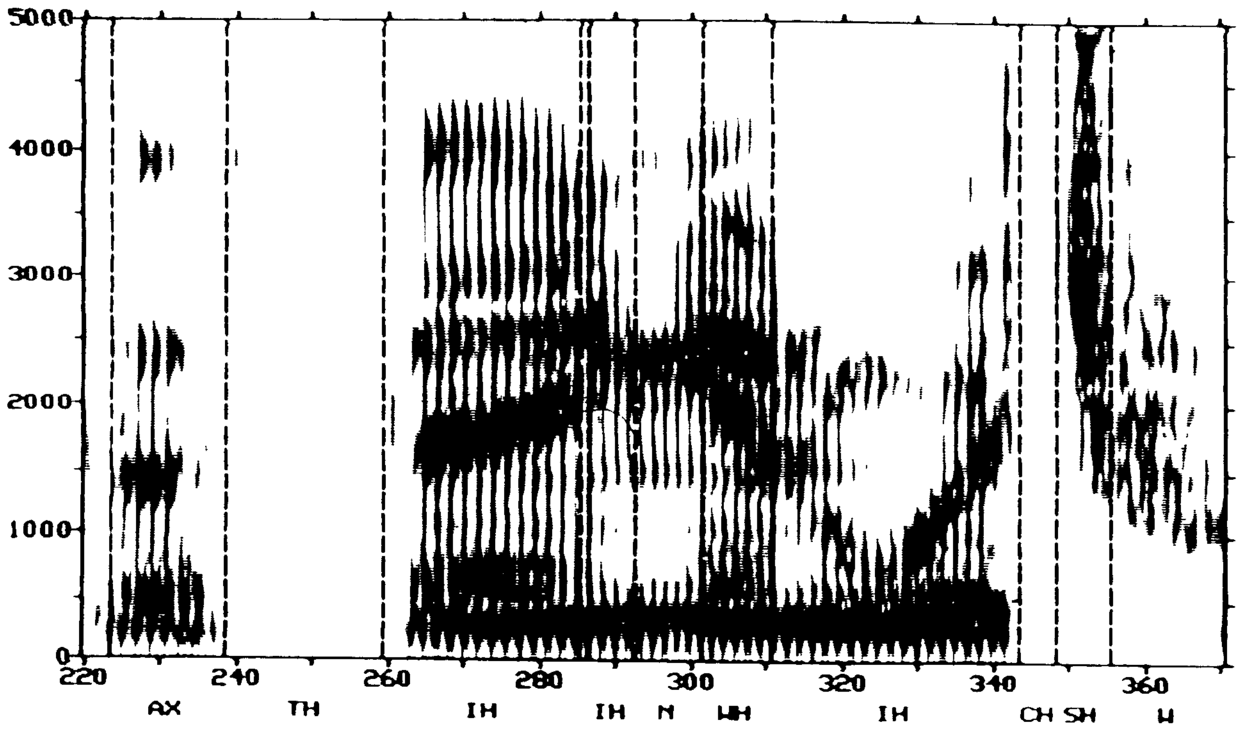


Fig. 2-6c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 2: (The sink is the thing in which) we pile di(shes).

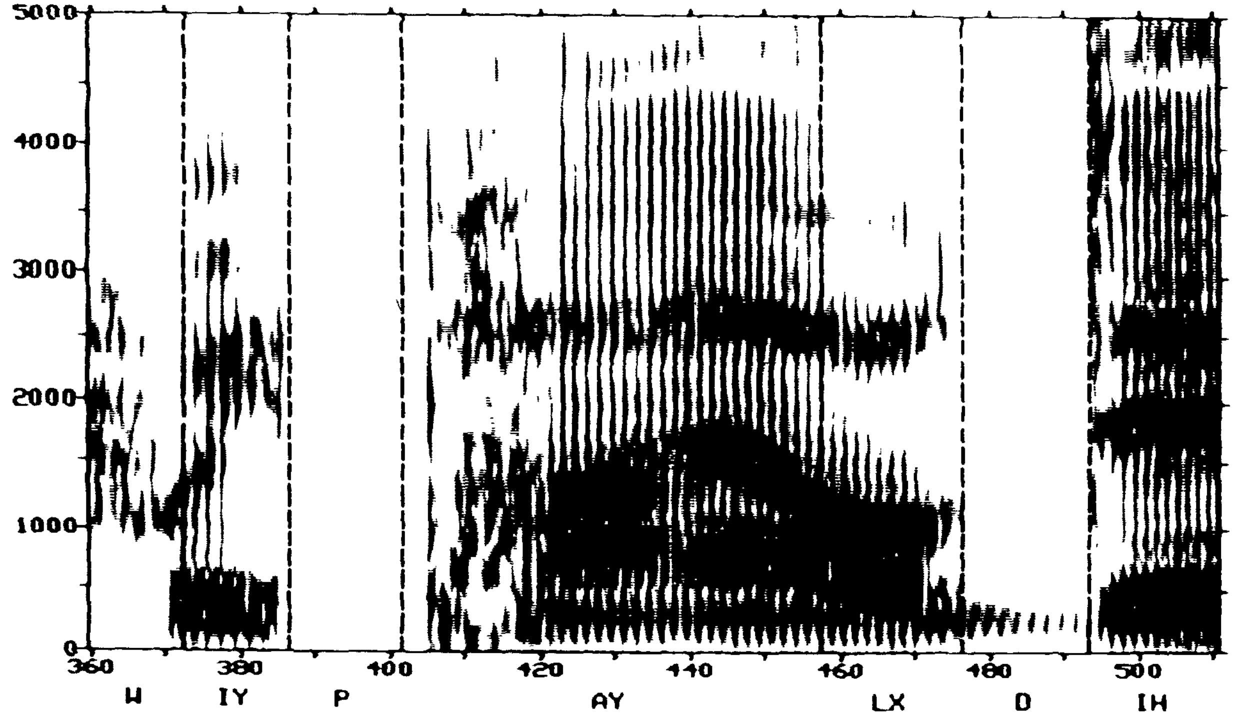


Fig. 2-6d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 2: (The sink is the thing in which we pile d)ishes.

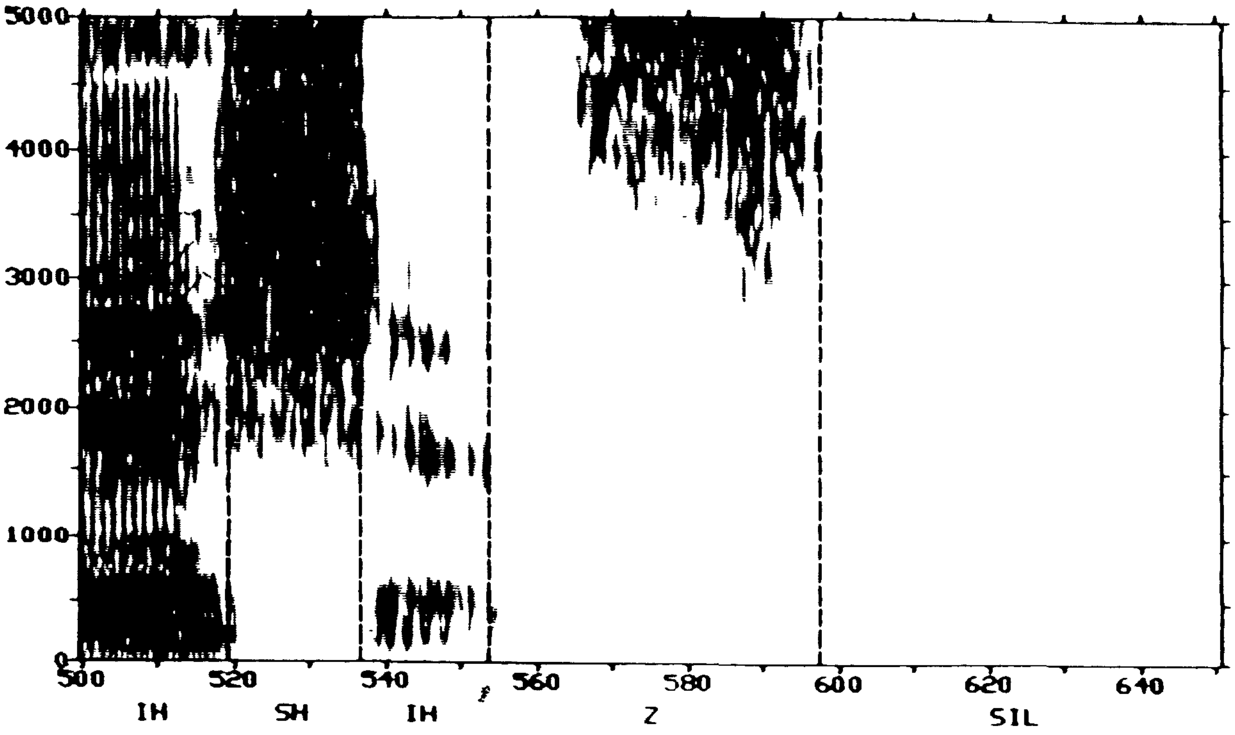


Fig. 2-7a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 3: A whiff of it w(ill cure the most stubborn cold).

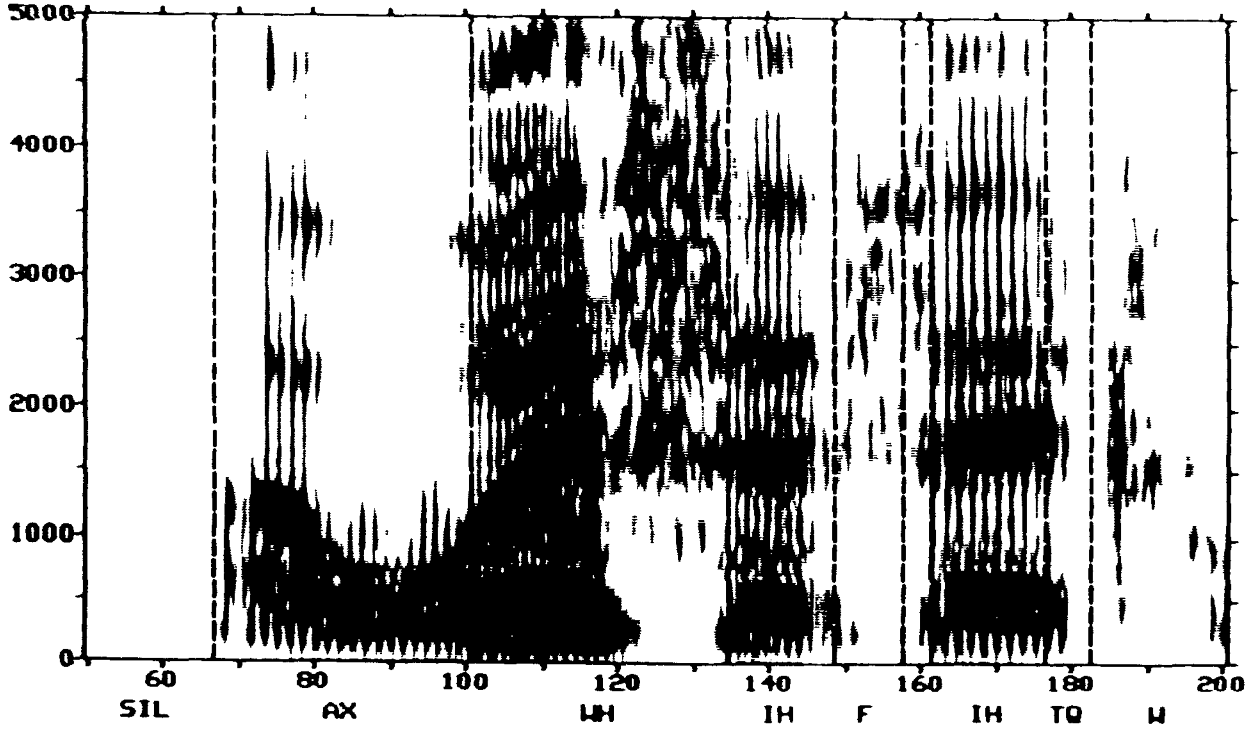


Fig. 2-7b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 3: (A whiff of it) will cure the mo(st stubborn cold).

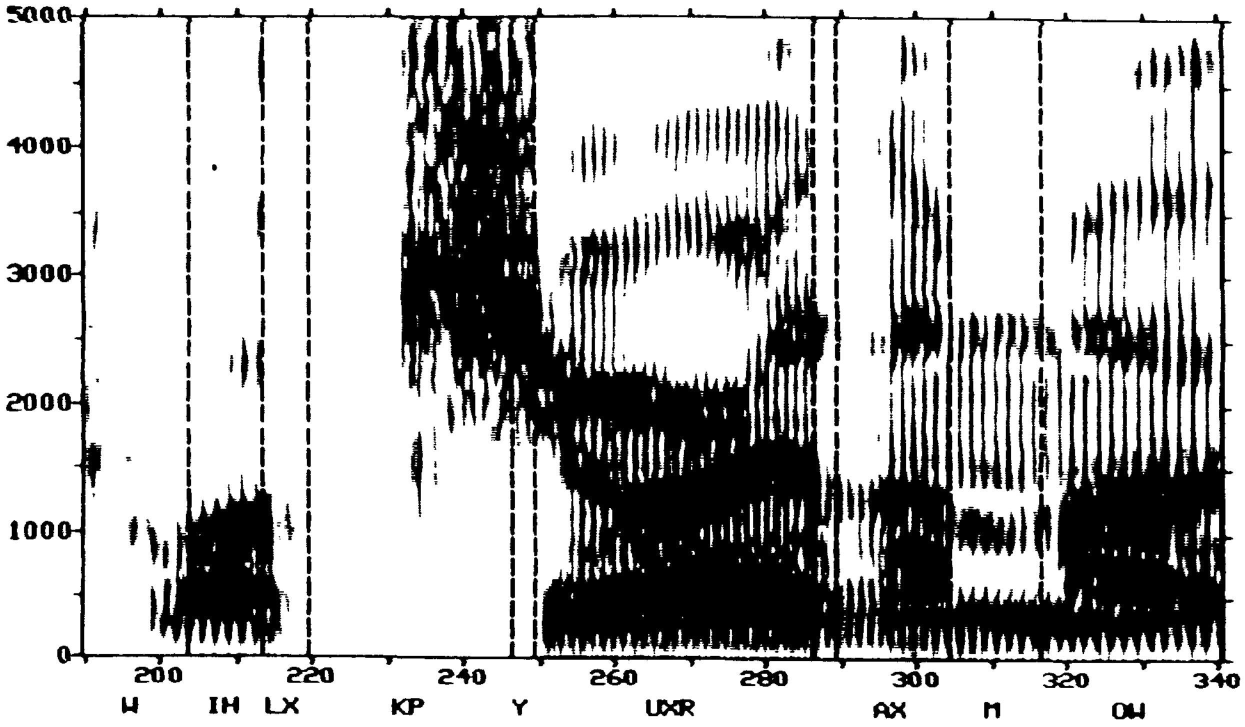


Fig. 2-7c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 3: (A whiff of it will cure the m)ost stubborn c(old).

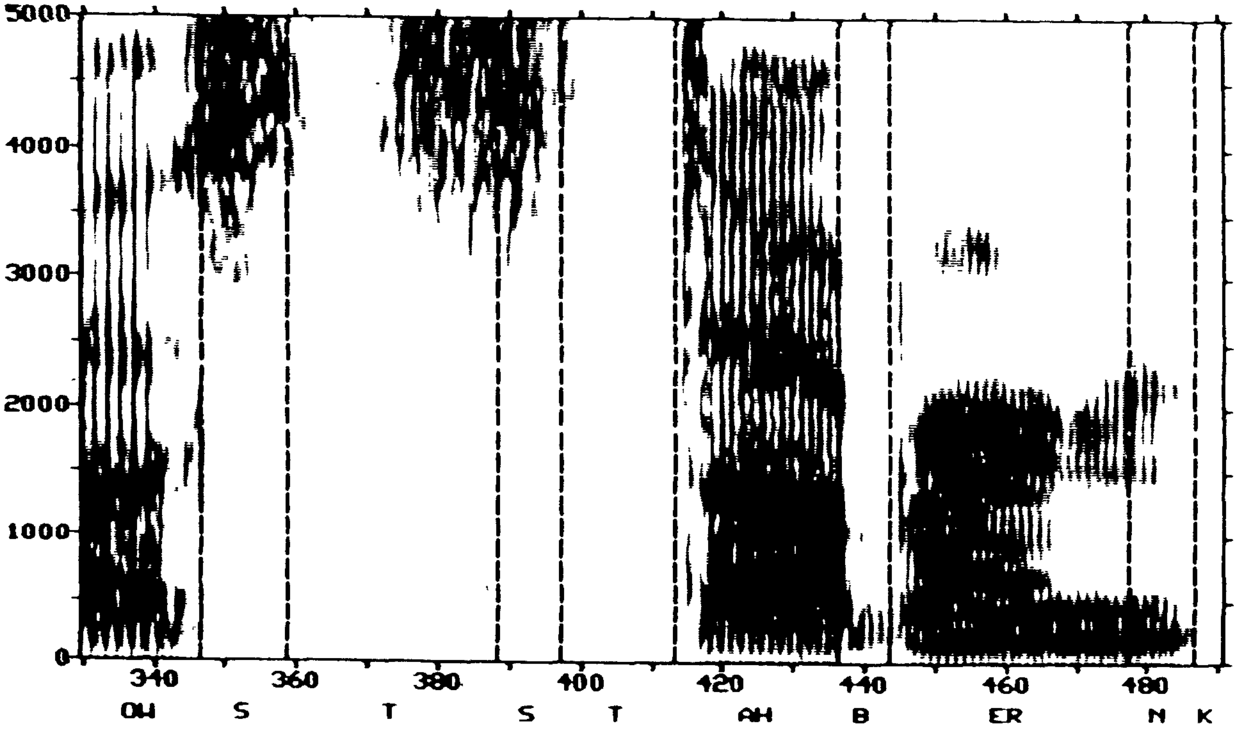


Fig. 2-7d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 3: (A whiff of it will cure the most stubbor)n cold.

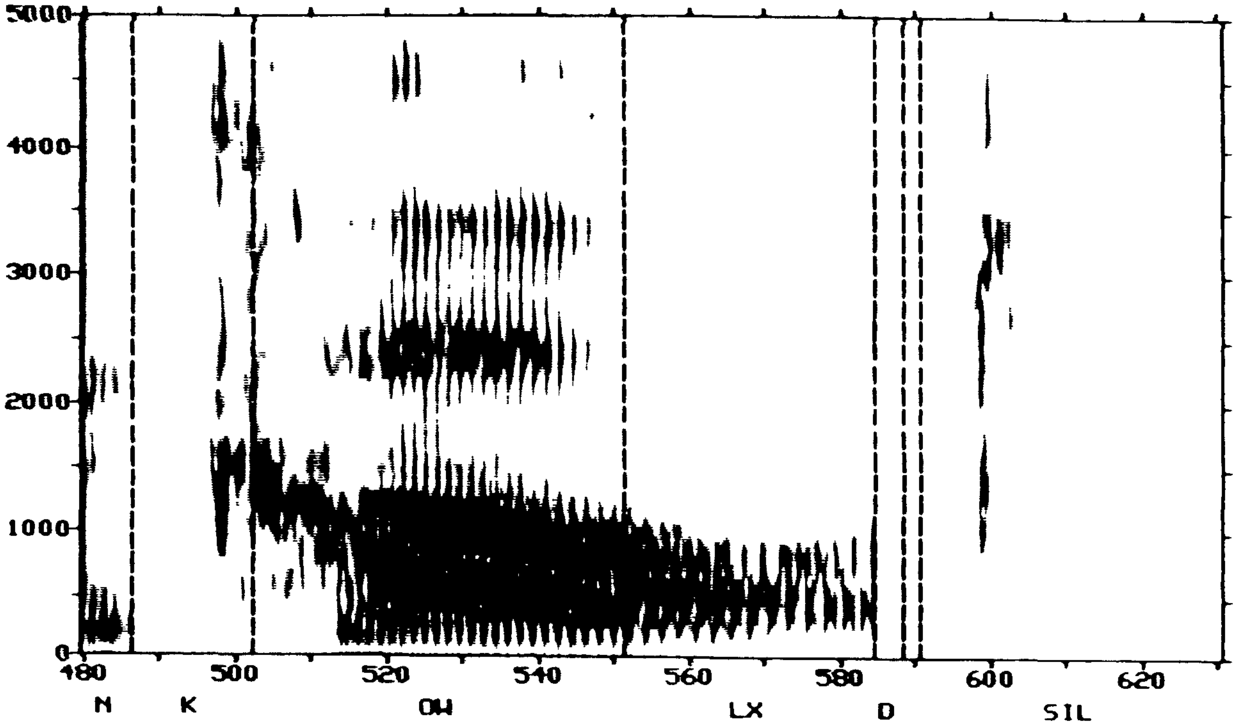


Fig. 2-8a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 4: The facts d(on´t always show who is right).

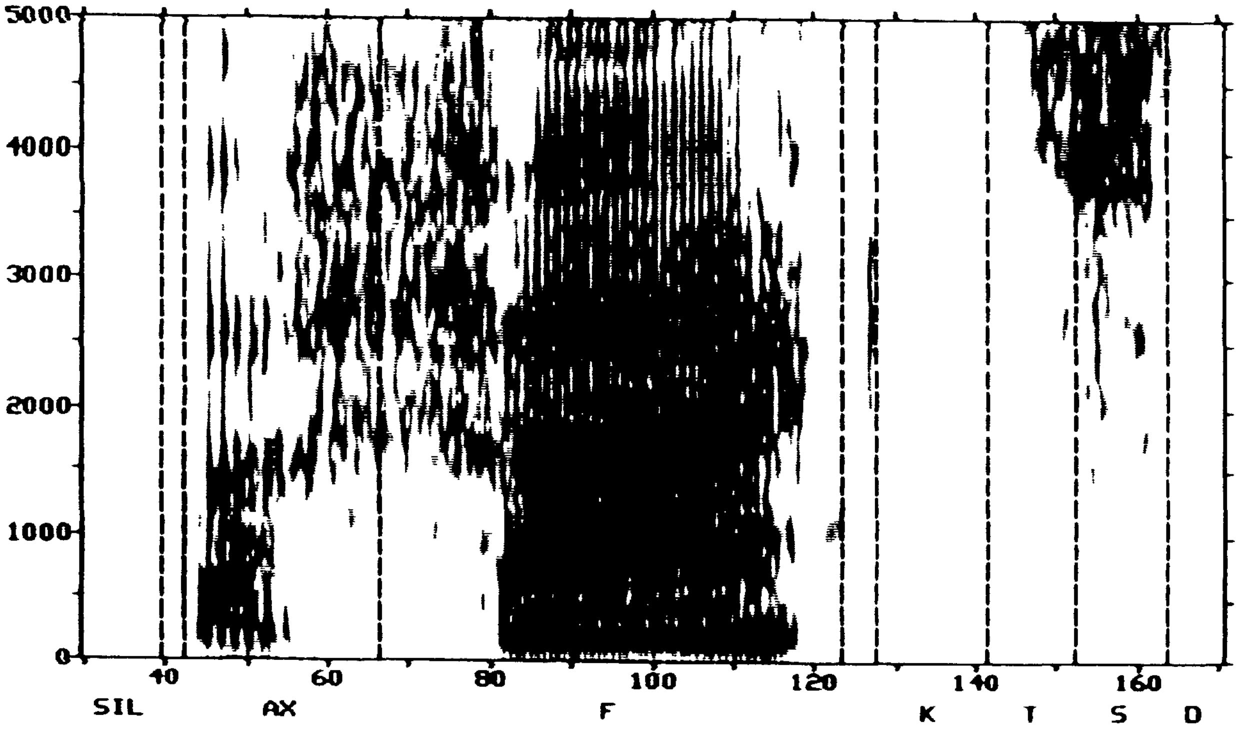


Fig. 2-8b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 4: (The fact)s don´t always sh(ow who is right).

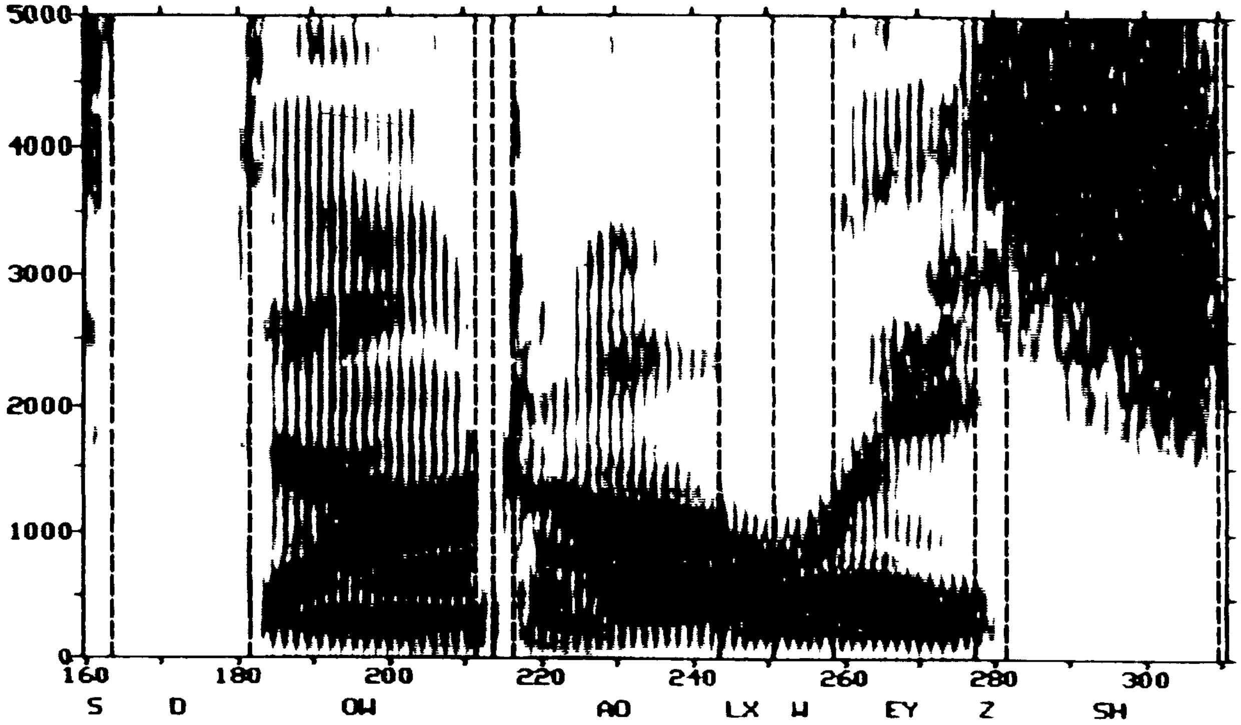


Fig. 2-8c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 4: (The facts don´t always) show who is ri(ght).

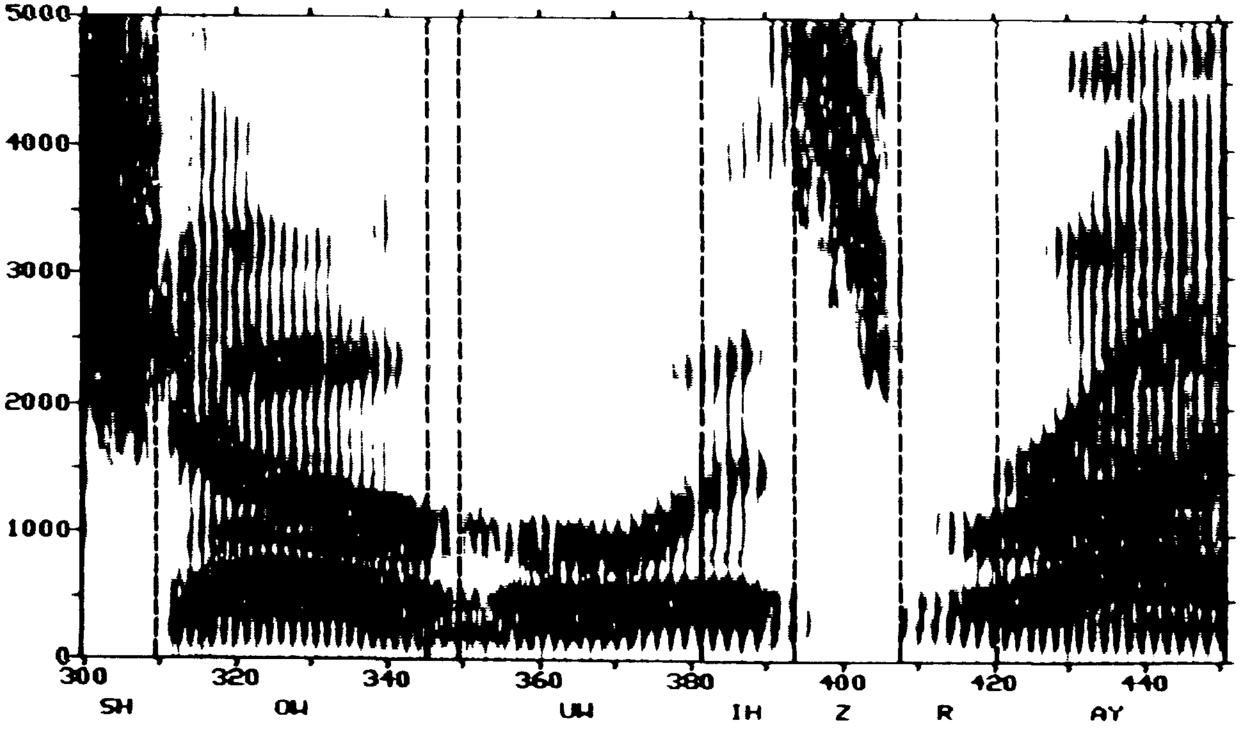


Fig. 2-8d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 4: (The facts don´t always show who i)s right.

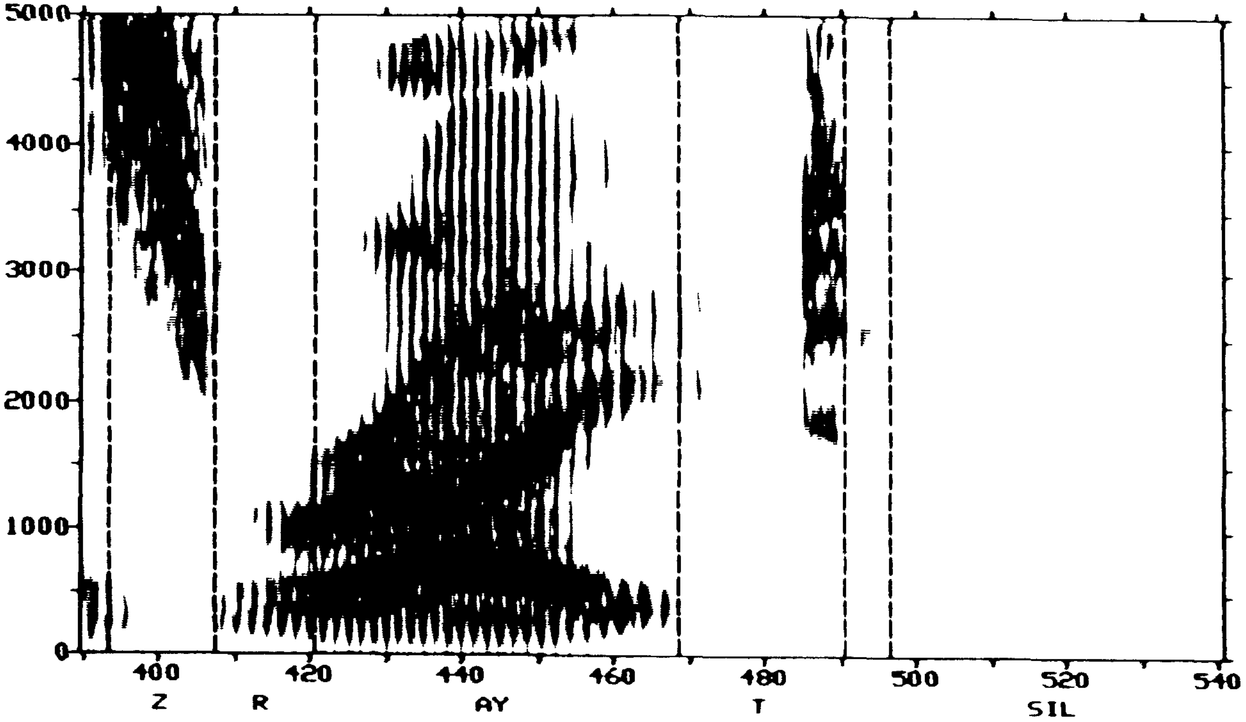


Fig. 2-9a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 5: She flaps (her cape as she parades the street).

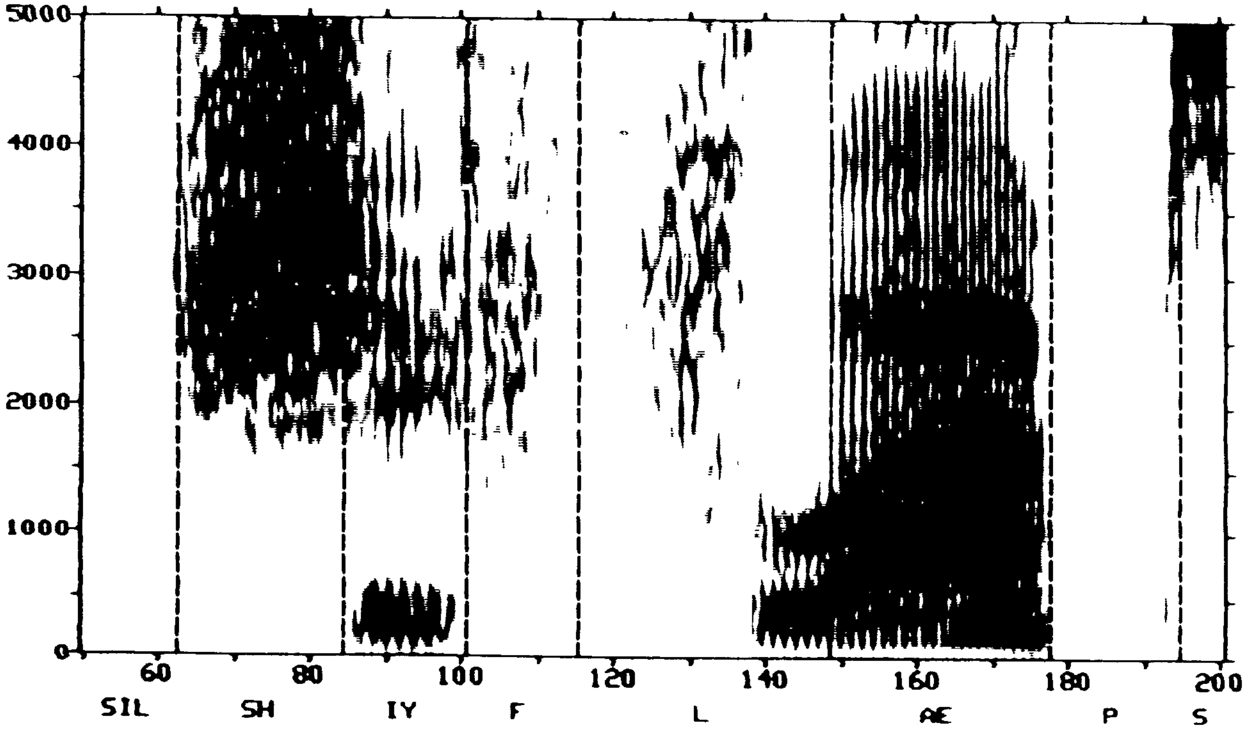


Fig. 2-9b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 5: (She fla)ps her cape as (she parades the street).

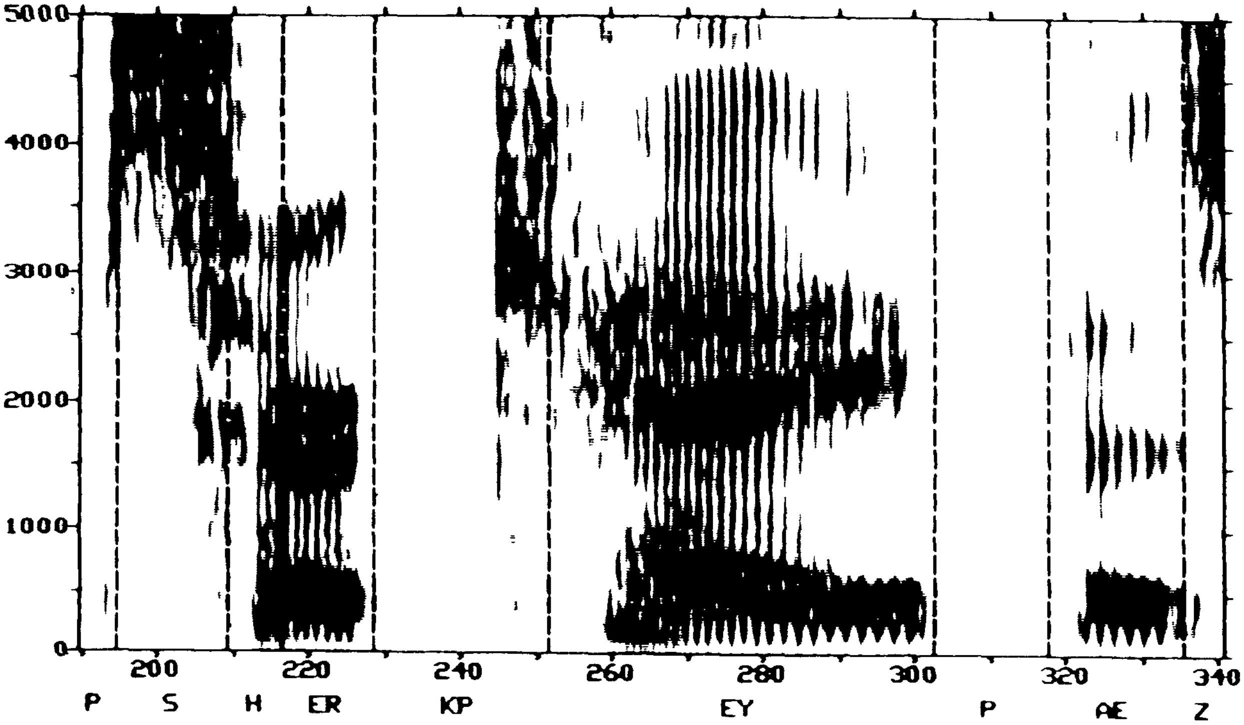


Fig. 2-9c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 5: (She flaps her cape) as she parade(s the street).

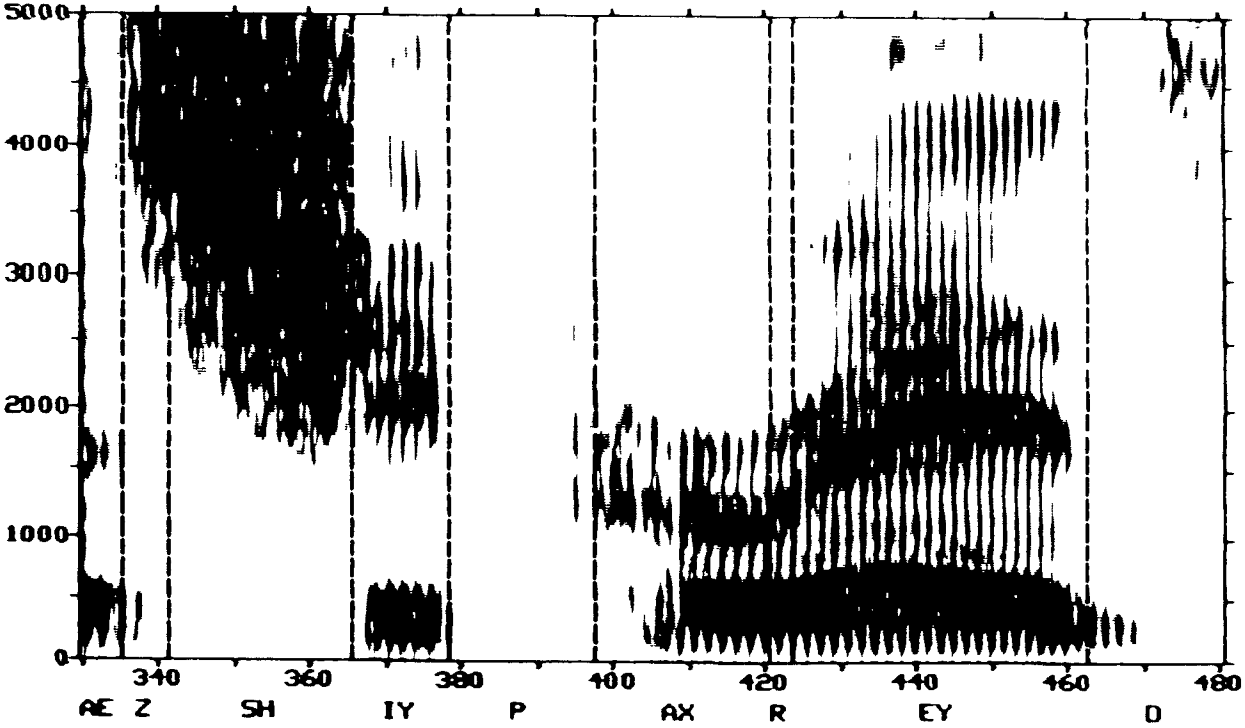


Fig. 2-9d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 5: (She flaps her cape as she para)des the street.

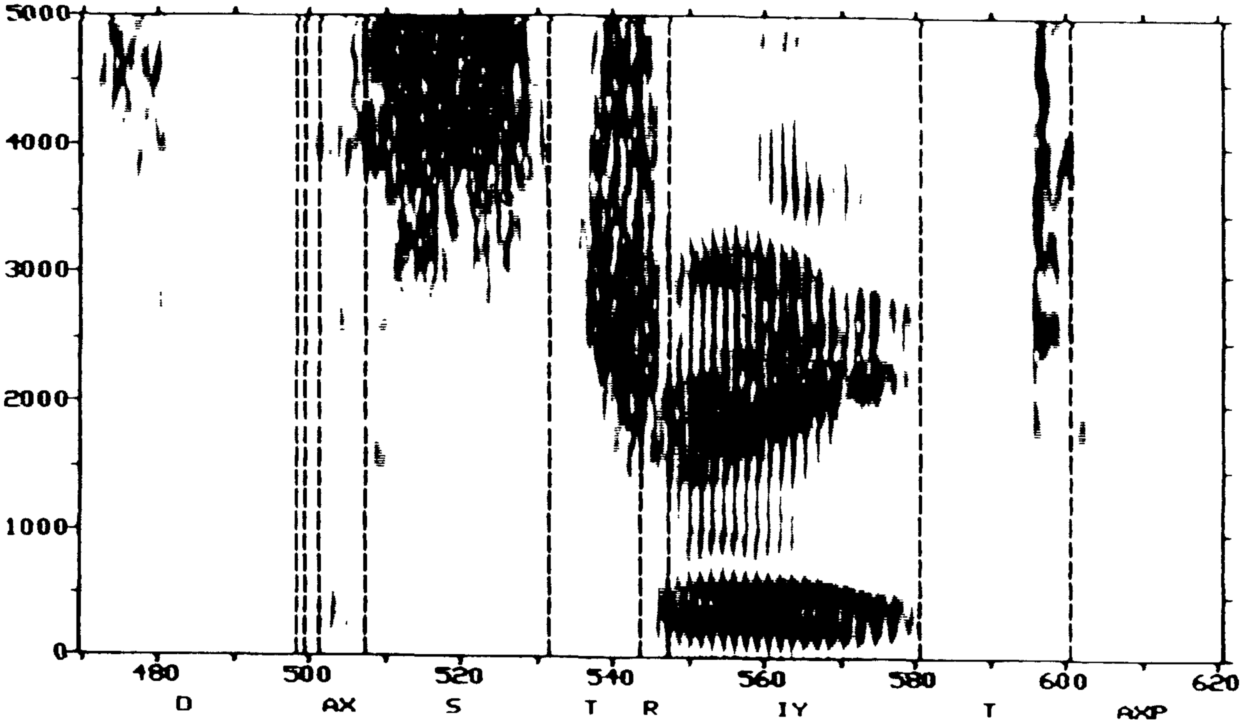


Fig. 2-10a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 6: The loss of th(e cruiser was a blow to the fleet).

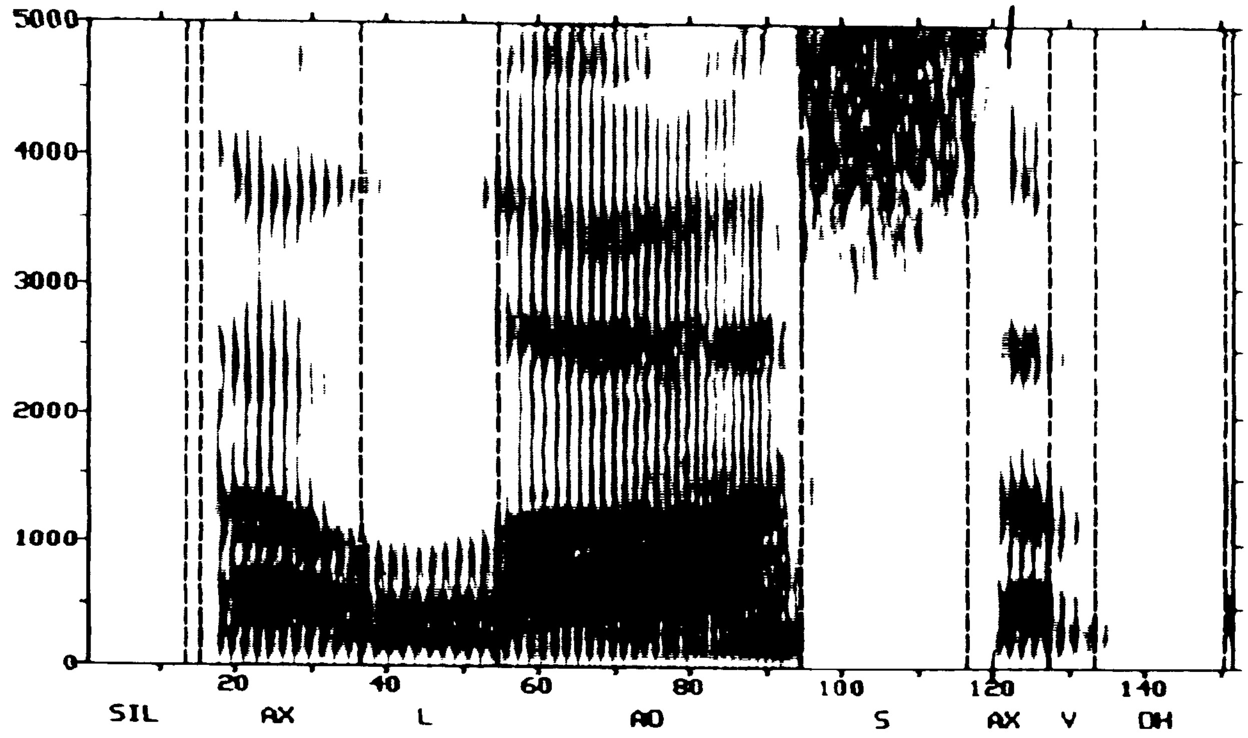


Fig. 2-10b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 6: (The loss of) the cruiser wa(s a blow to the fleet).

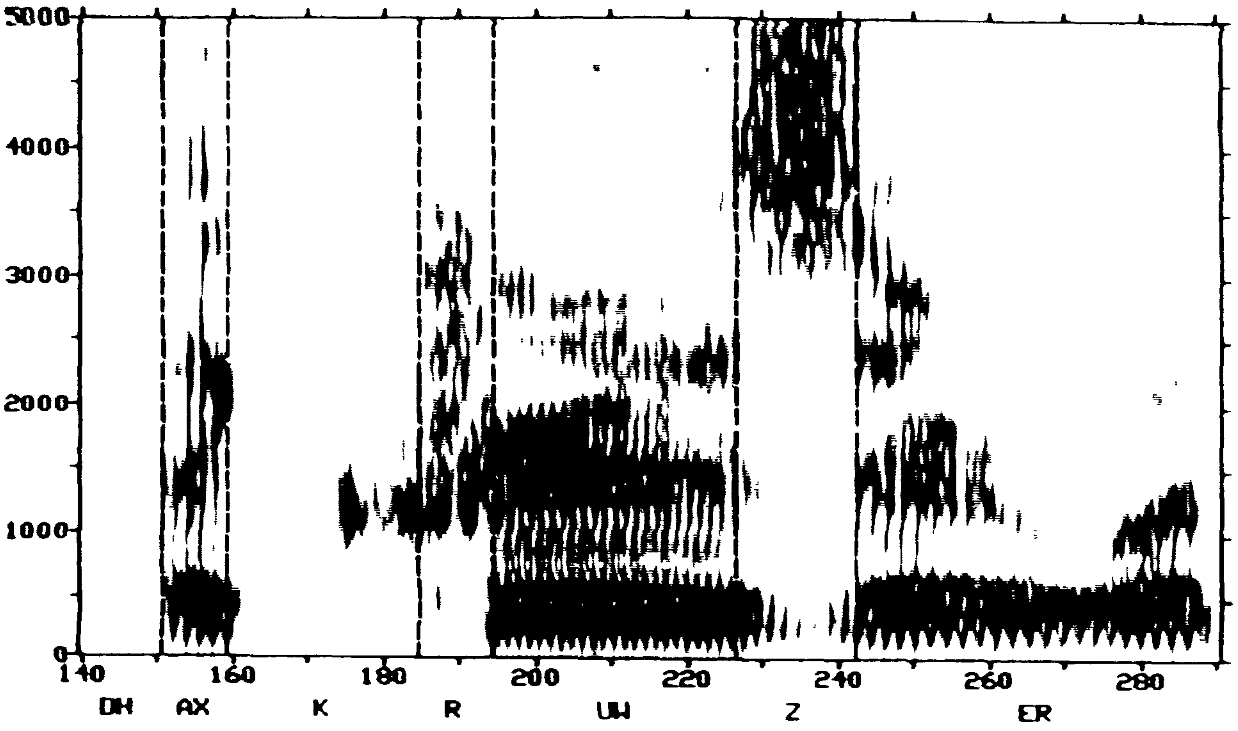


Fig. 2-10c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 6: (The loss of the cruis)er was a blow to (the fleet).

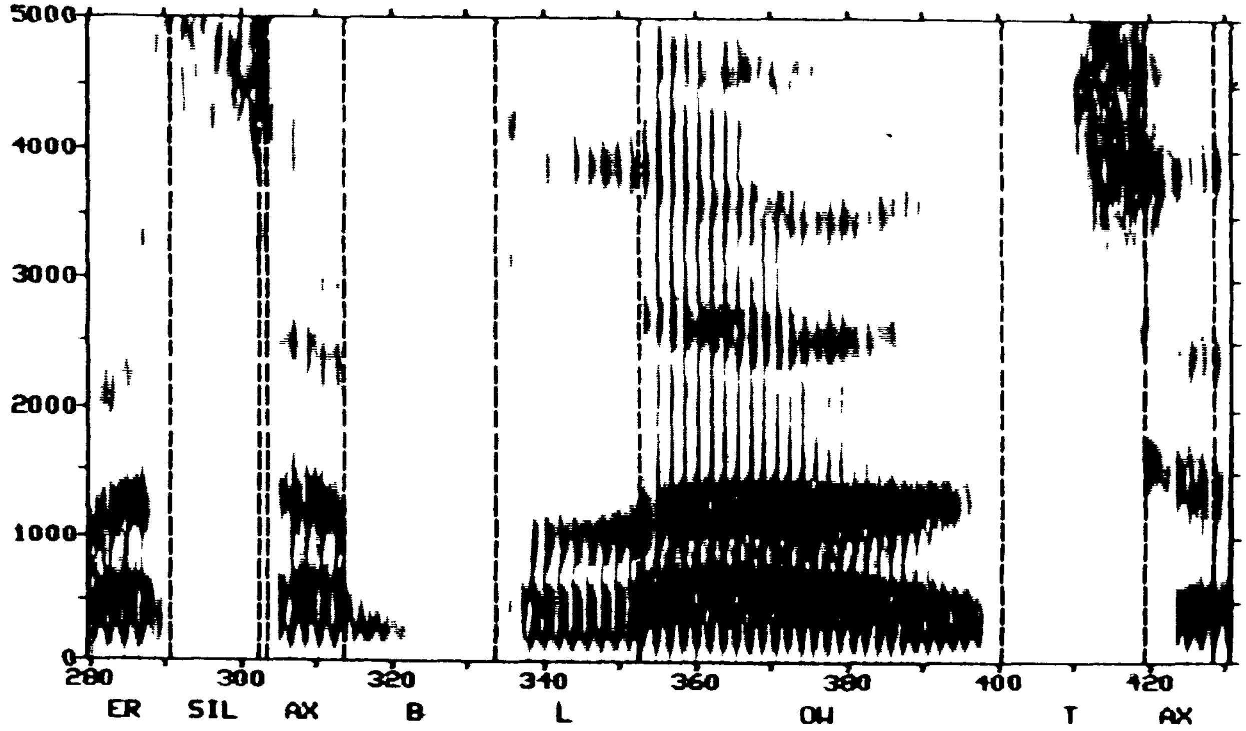


Fig. 2-10d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 6: (The loss of the cruiser was) a blow to the fleet.

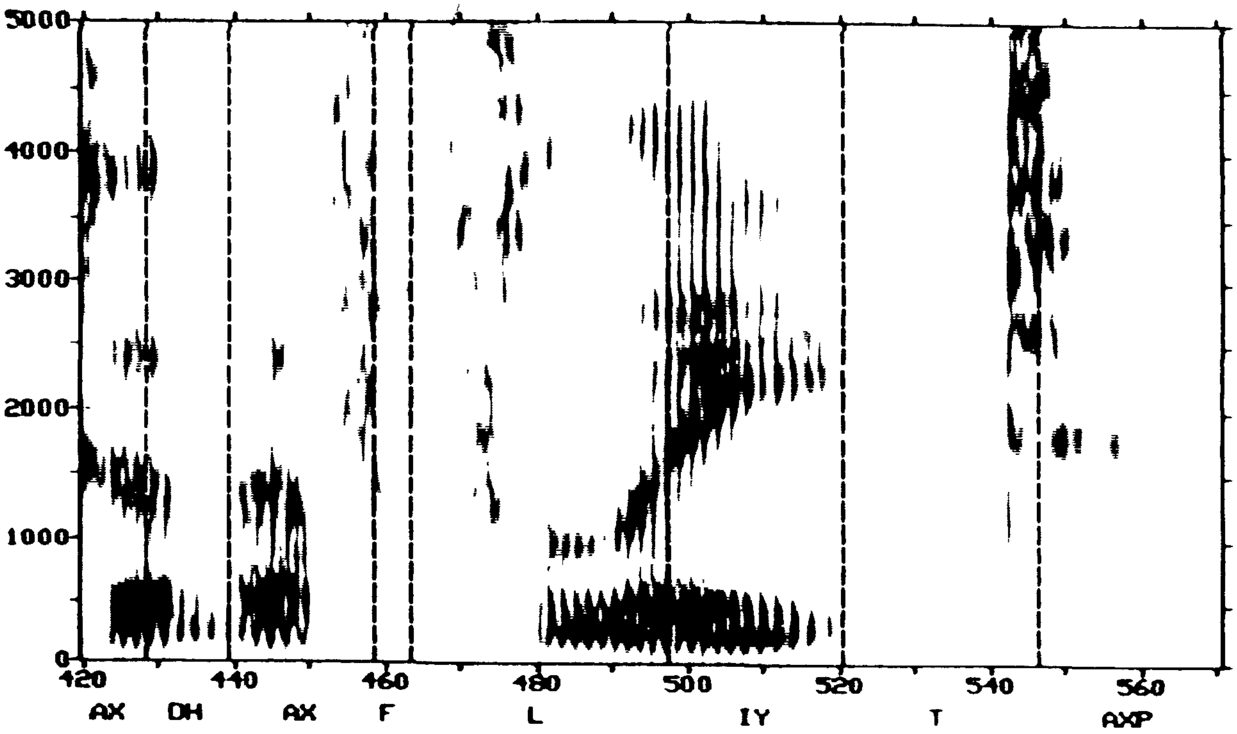


Fig. 2-11a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 7: Loop the brai(d to the left and then over).

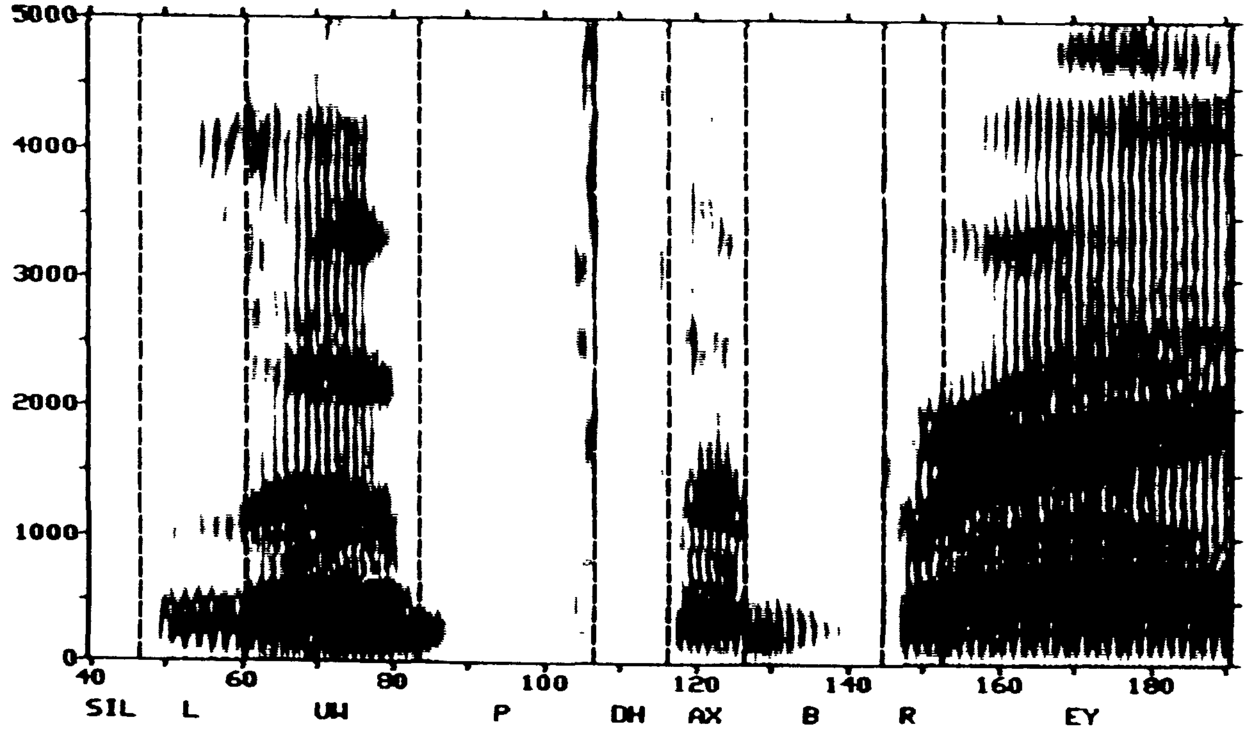


Fig. 2-11b. Speaker ML with induced MI"alk segmentation and labelling.

Sentence 7: (Loop the br)aid to the lef(t and then over).

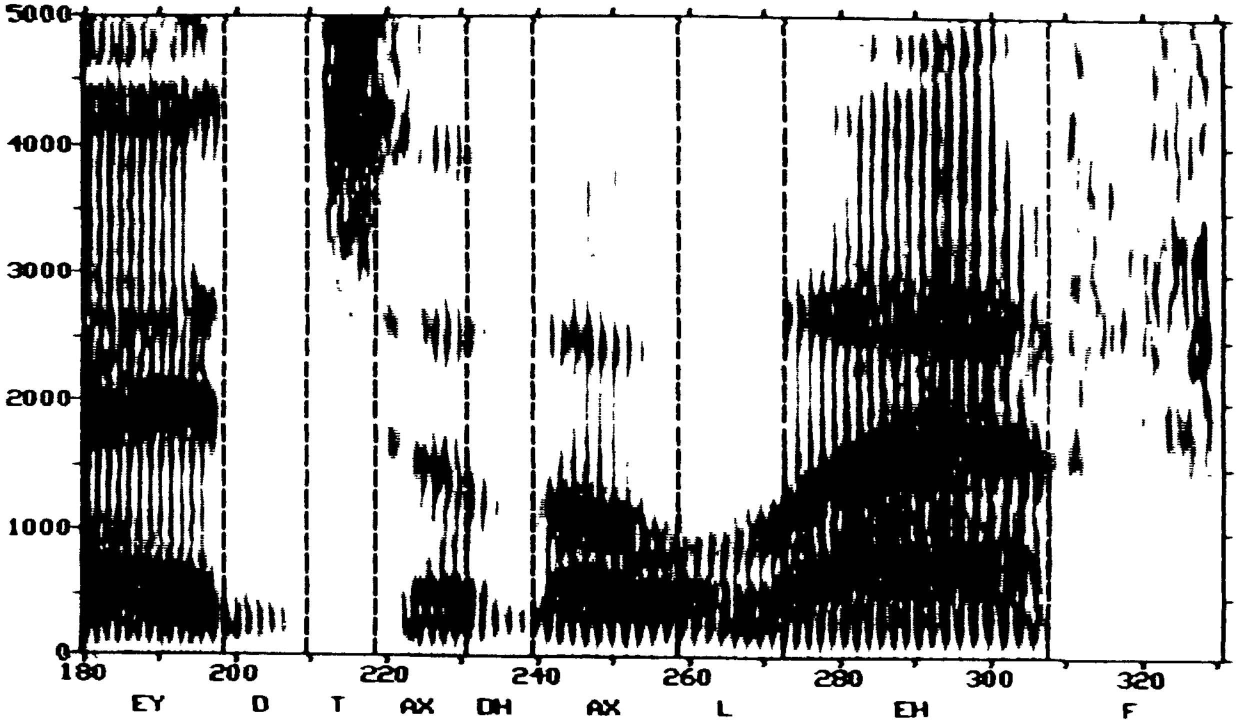


Fig. 2-11c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 7: (Loop the braid to the le)ft and then o(ver).

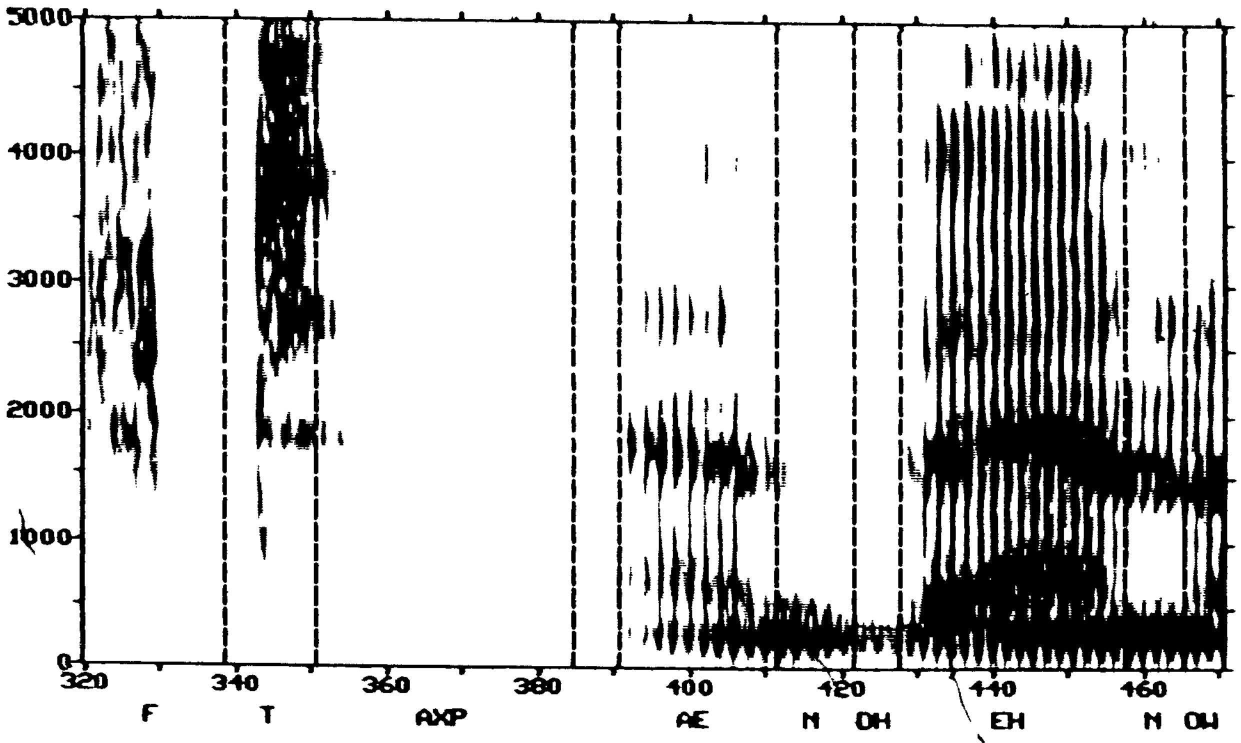


Fig. 2-11d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 7: (Loop the braid to the left and th)en over.

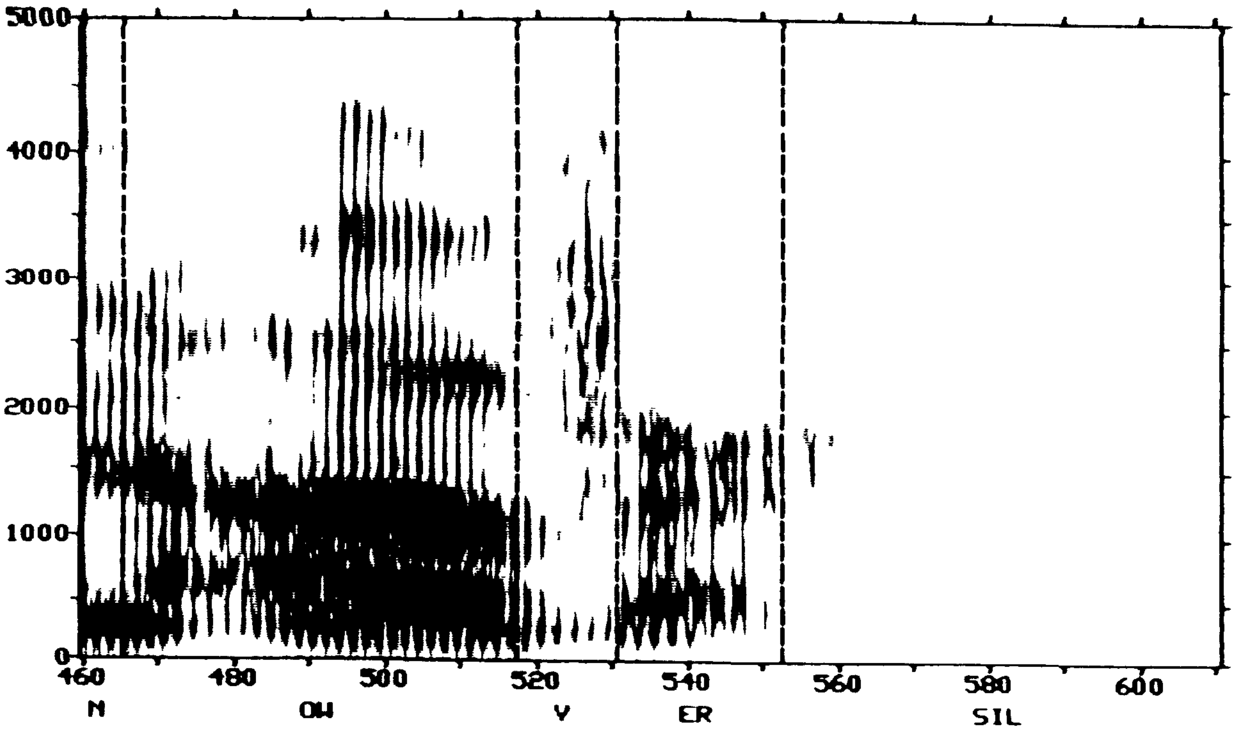


Fig. 2-12a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 8: Plead with the l(awyer to drop the lost cause).

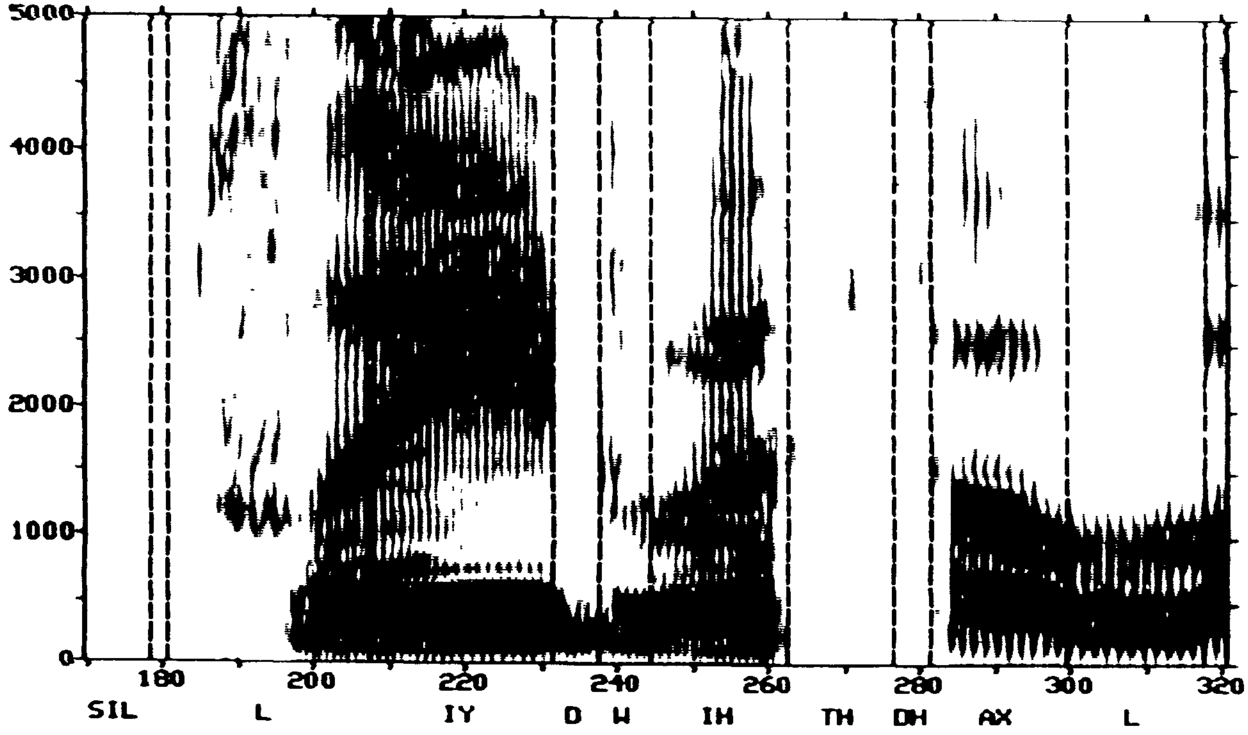


Fig. 2-12b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 8: (Plead with the) lawyer to dro(p the lost cause).

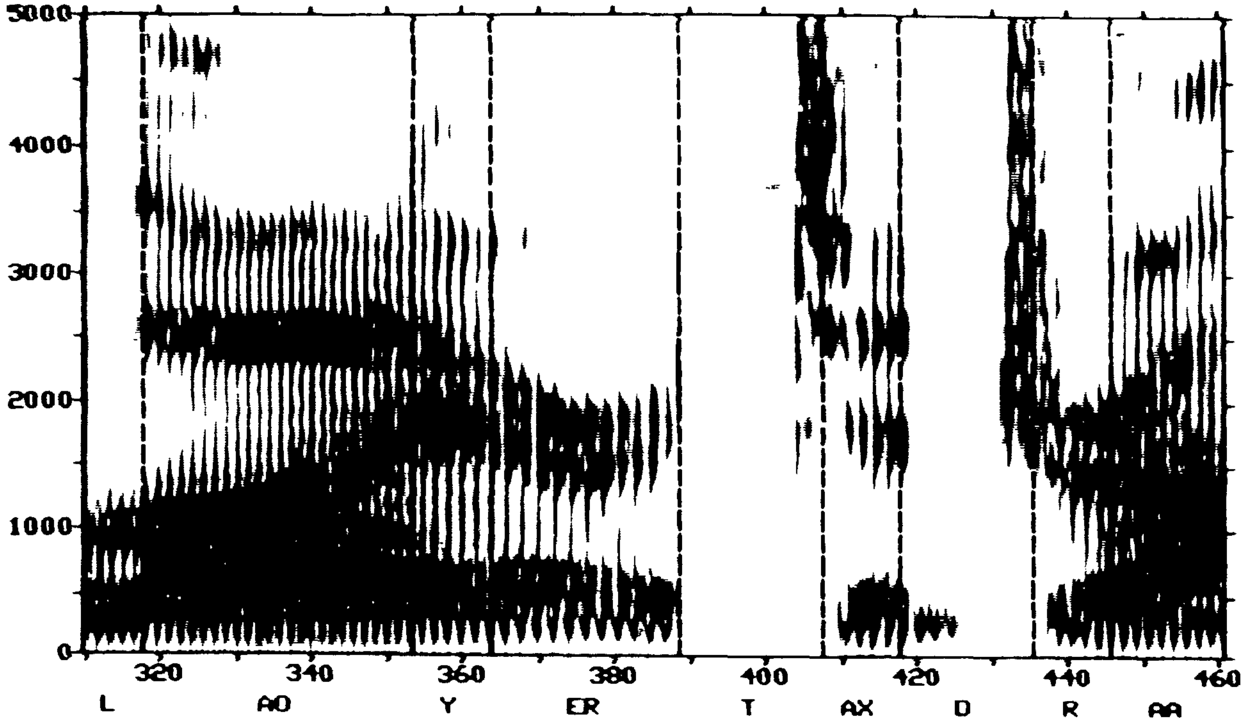


Fig. 2-12c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 8: (Plead with the lawyer to dr)op the lost cau(se).

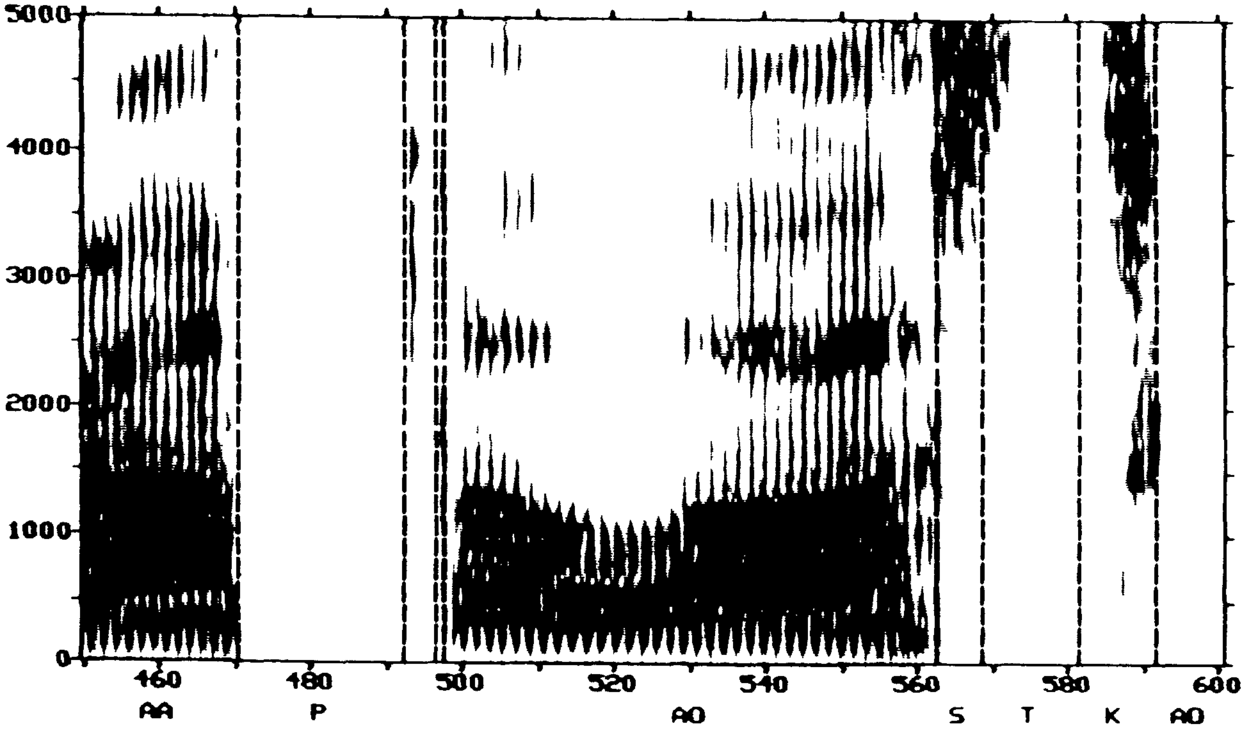


Fig. 2-12d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 8: (Plead with the lawyer to drop the los)t cause.

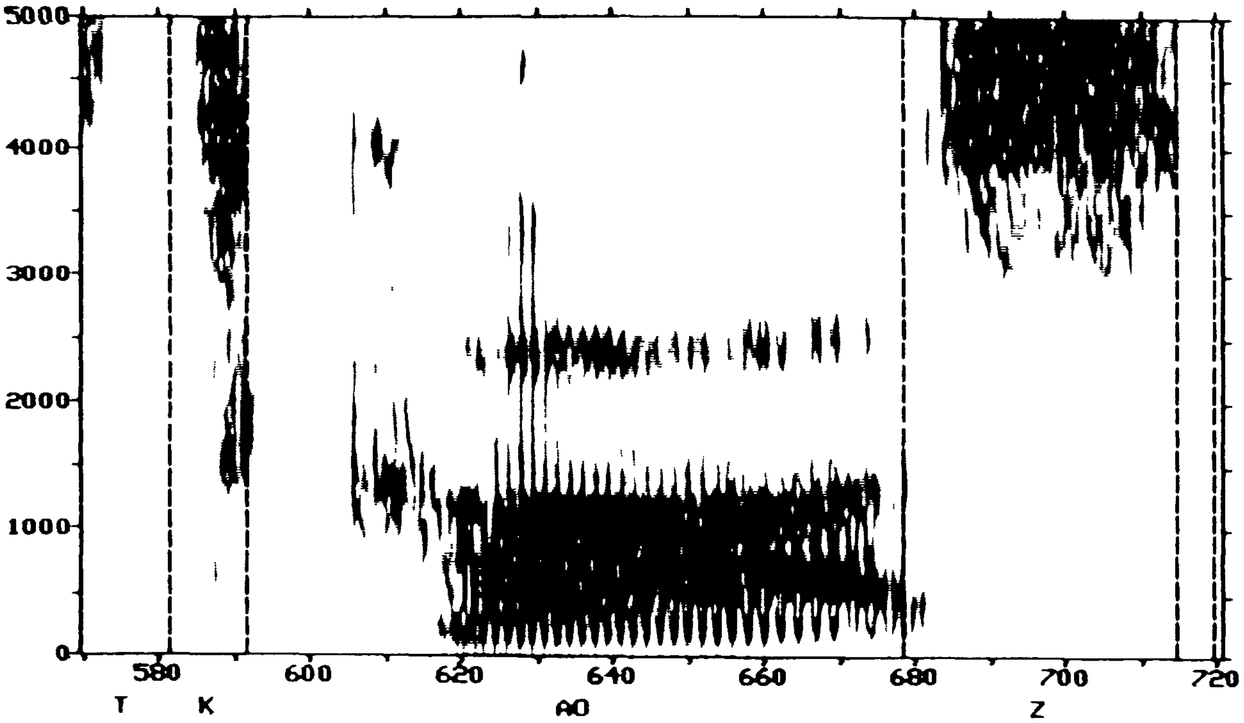


Fig. 2-13a. Speaker ML with induced MITalk segmentation and labelling.

Sentence 9: Calves thr(ive on tender spring grass).

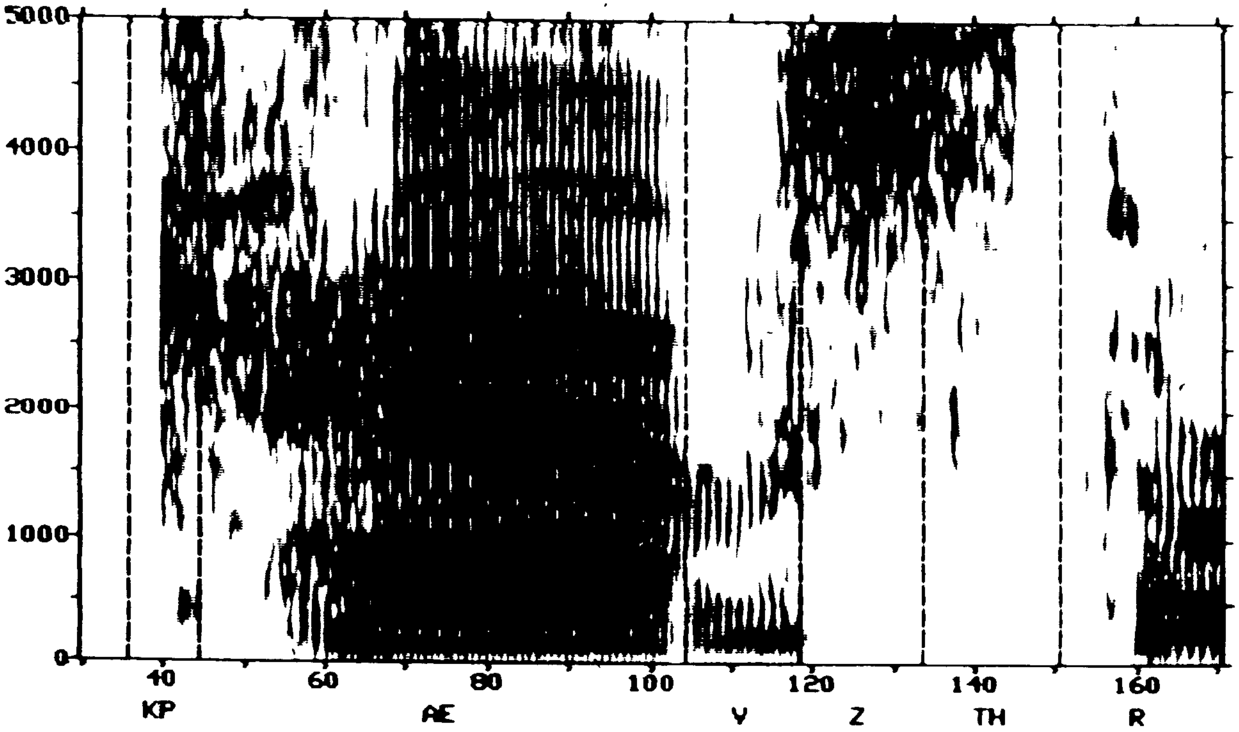


Fig. 2-13b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 9: (Calves th)rive on te(nder spring grass).

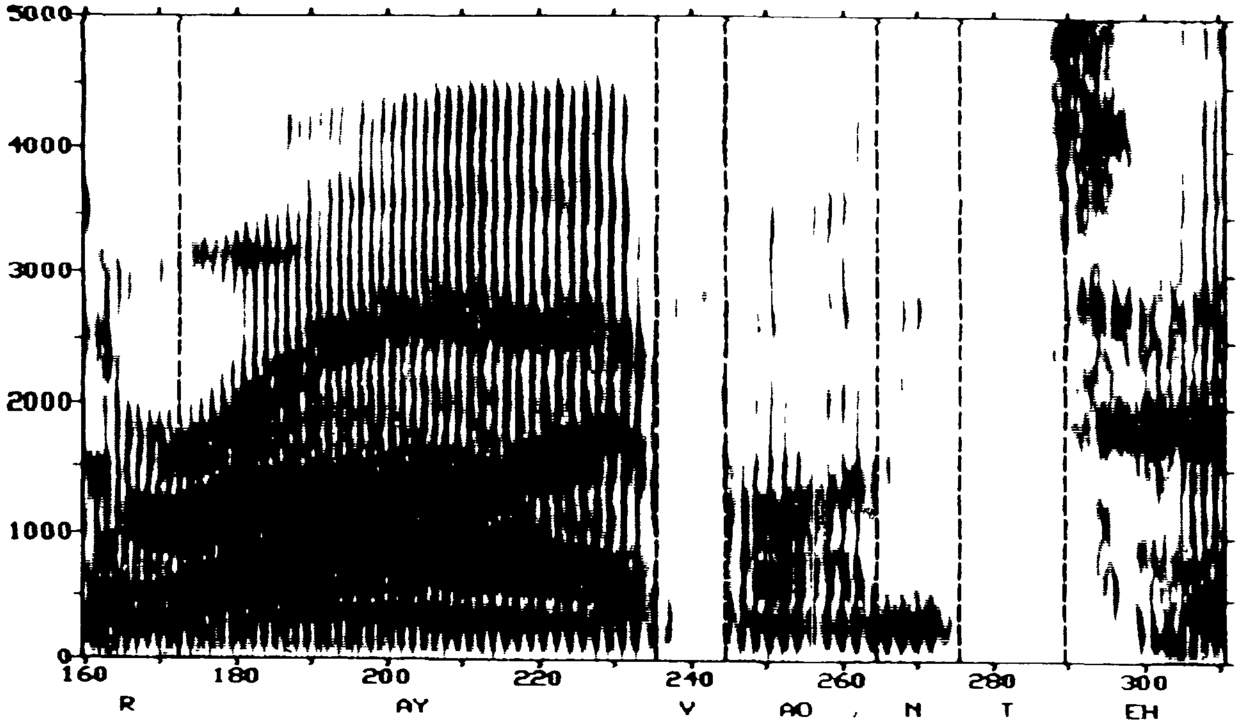


Fig. 2-13c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 9: (Calves thrive on t)ender spri(ng grass).

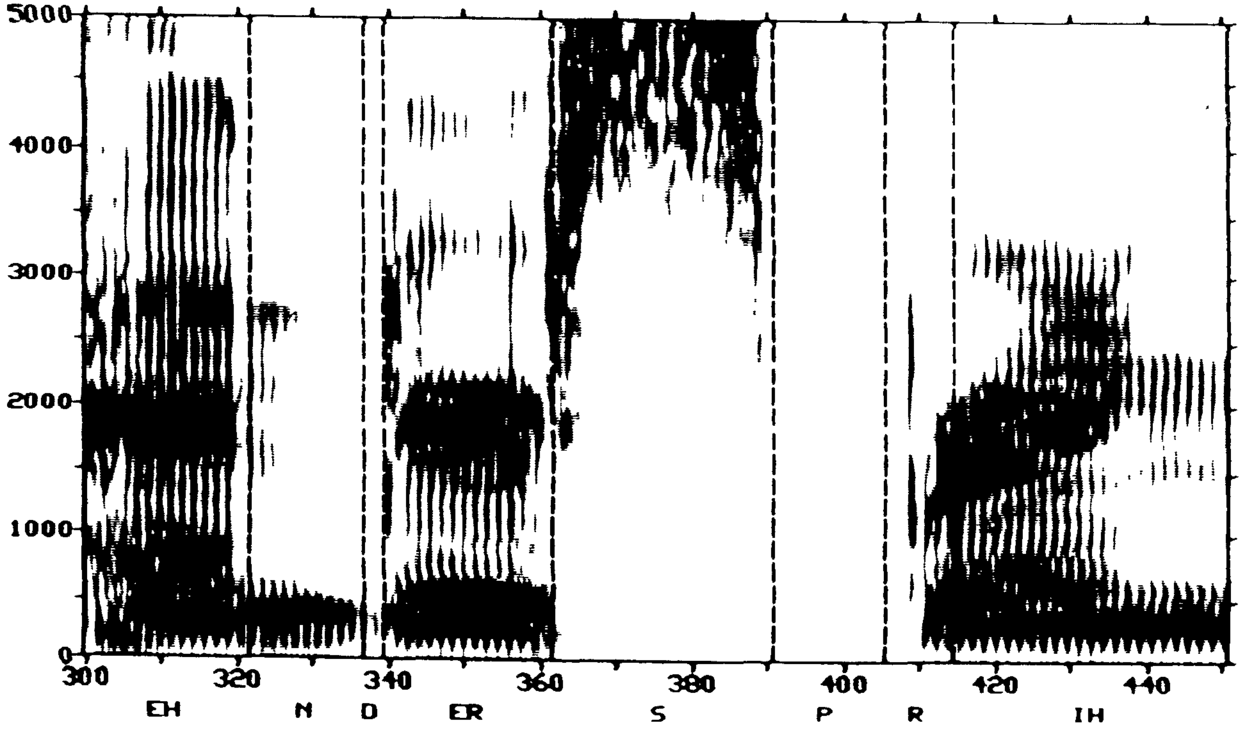


Fig. 2-13d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 9: (Calves thrive on tender spr)ing grass.

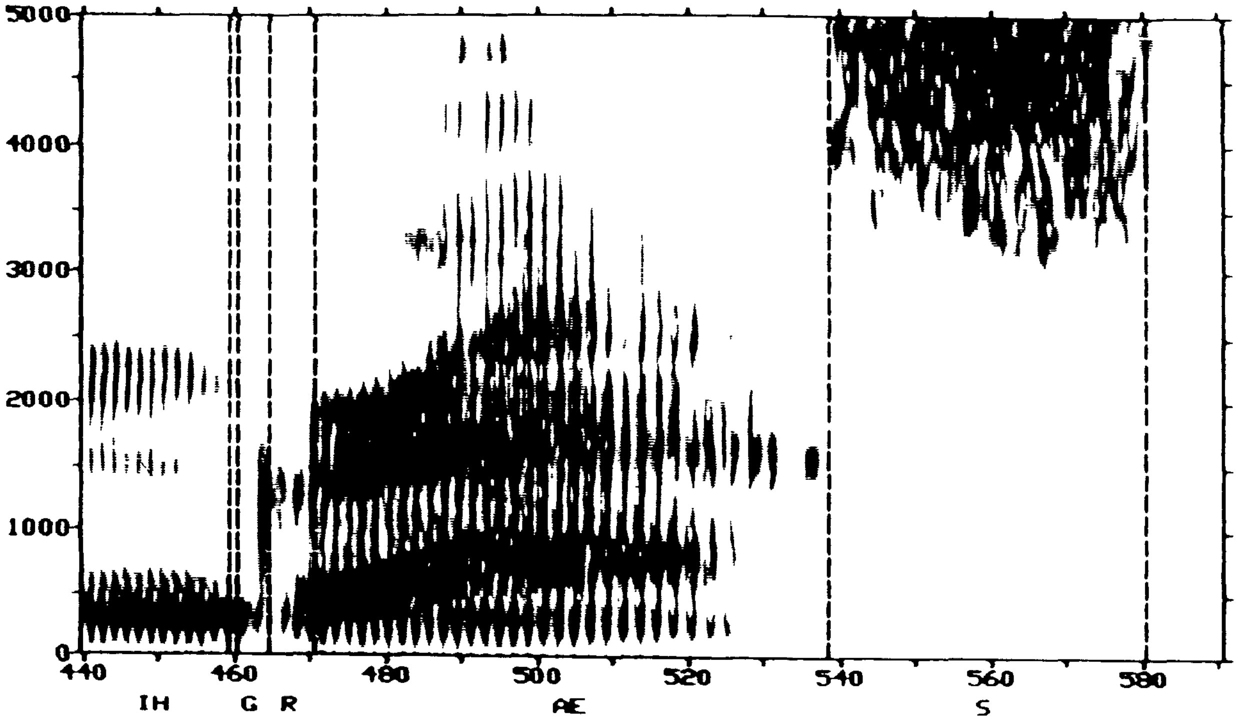


Fig. 2-14a. Speaker ML with induced MITalk segmentation and labellıng.

Sentence 10: Post no bi(lls on this office wall).

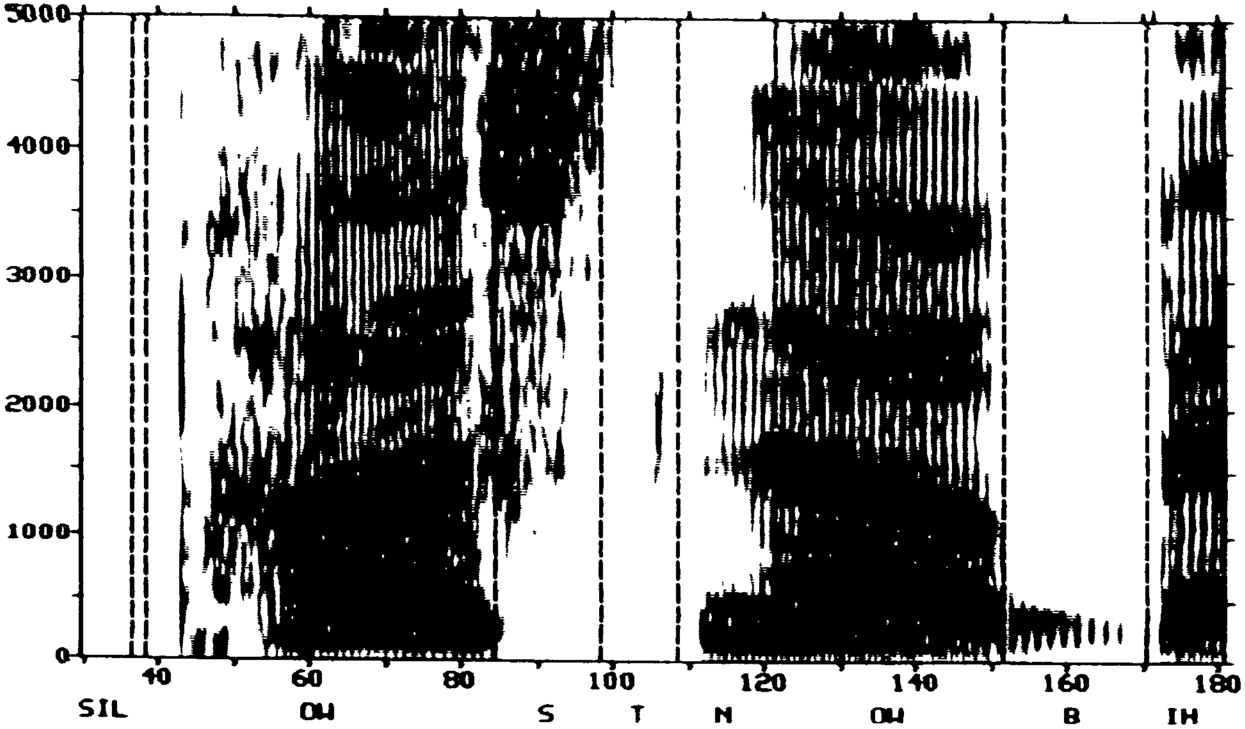


Fig. 2-14b. Speaker ML with induced MITalk segmentation and labelling.

Sentence 10: (Post no b)ills on this (office wall).

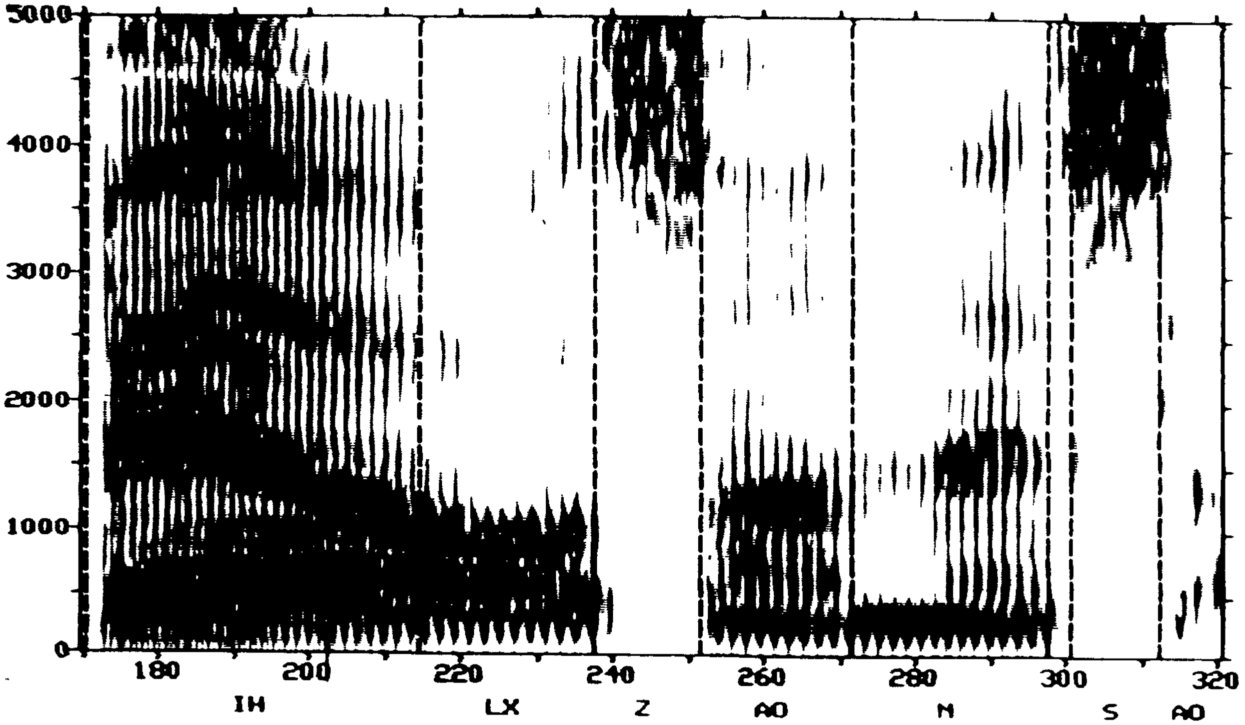


Fig. 2-14c. Speaker ML with induced MITalk segmentation and labelling.

Sentence 10: (Post no bills on this) office wa(ll).

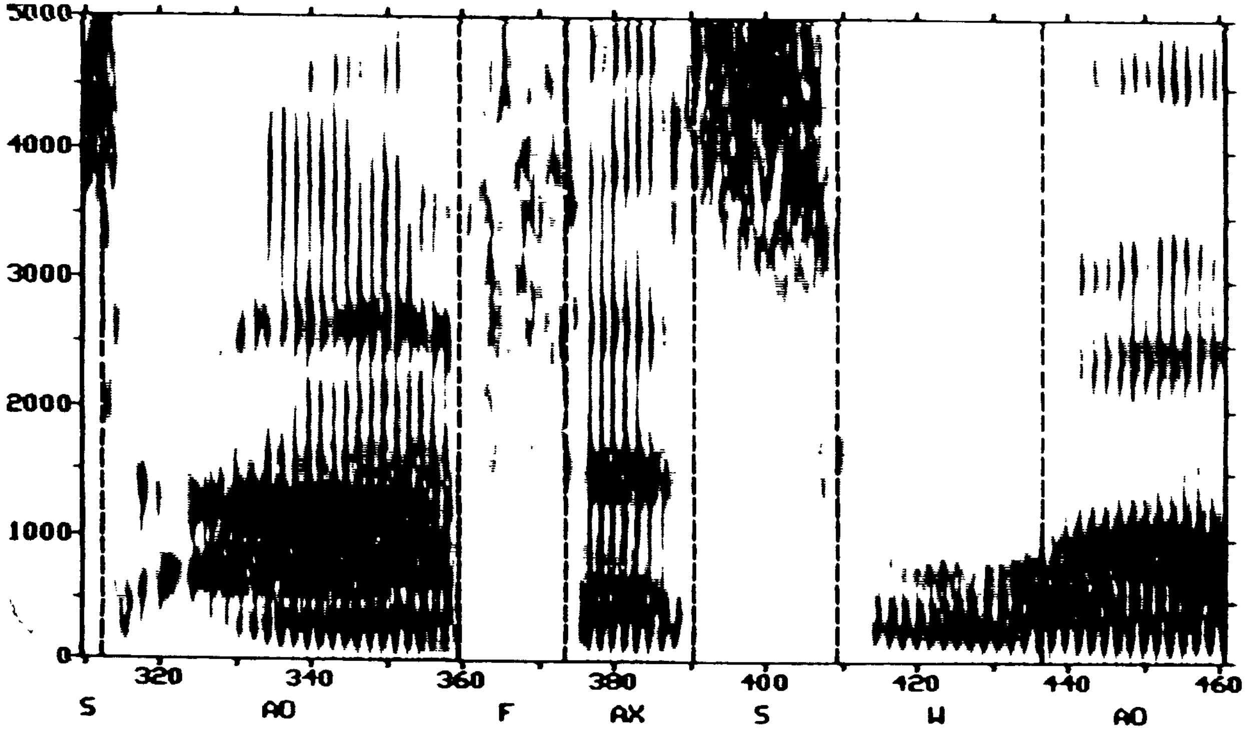
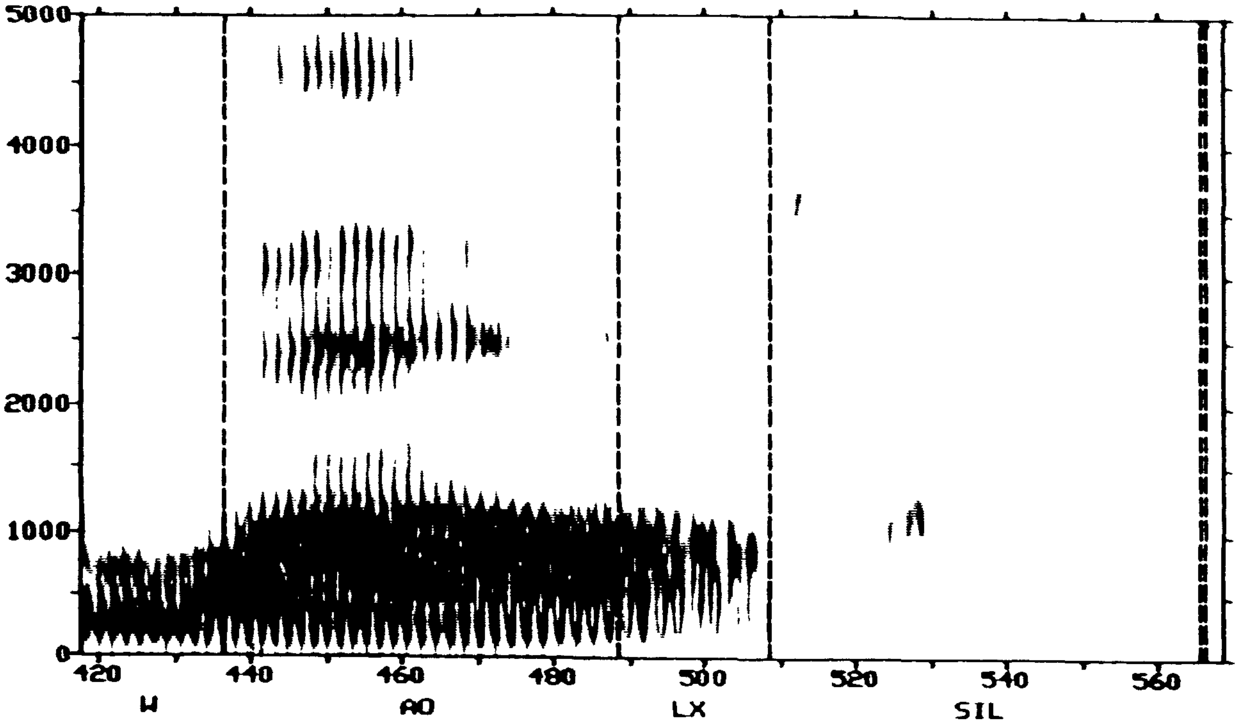


Fig. 2-14d. Speaker ML with induced MITalk segmentation and labelling.

Sentence 10: (Post no bills on this office) wall.



# FIGURES FOR

# CHAPTER 3

Fig. 3-1a. Synthetic speech with rule-based segmentation and labelling.

Sentence 1: The goose (was brought straight from the old market).

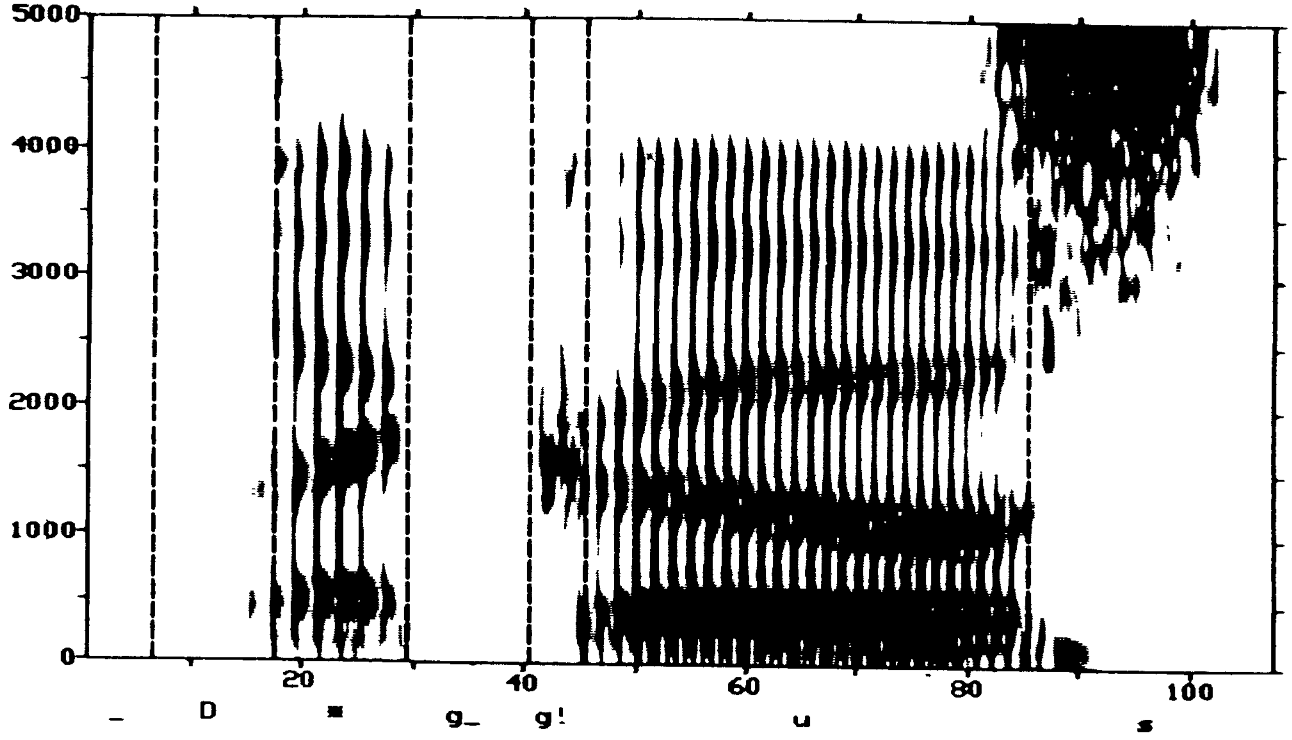


Fig. 3-1b. Synthetic speech with rule-based segmentation and labelling.

Sentence 1: (The goose) was brought (straight from the old market).

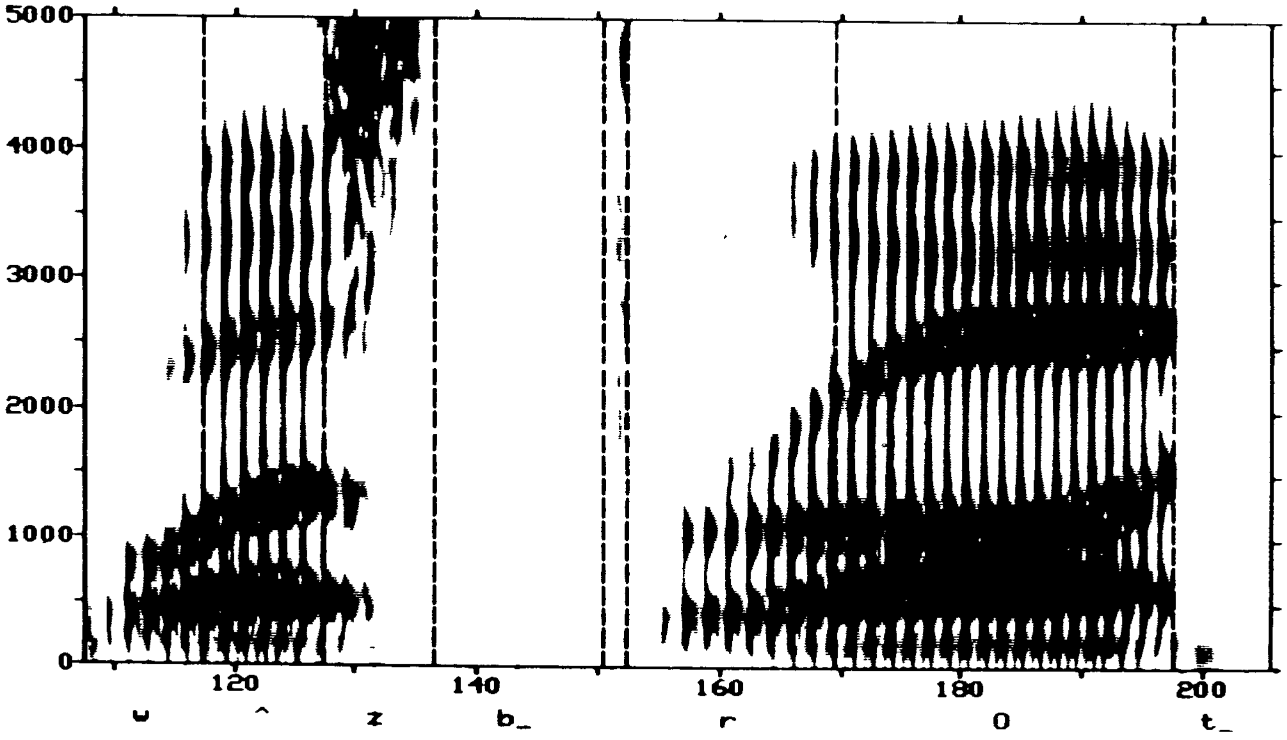


Fig. 3-1c. Synthetic speech with rule-based segmentation and labelling.

Sentence 1: (The goose was brought) straight f(rom the old market).

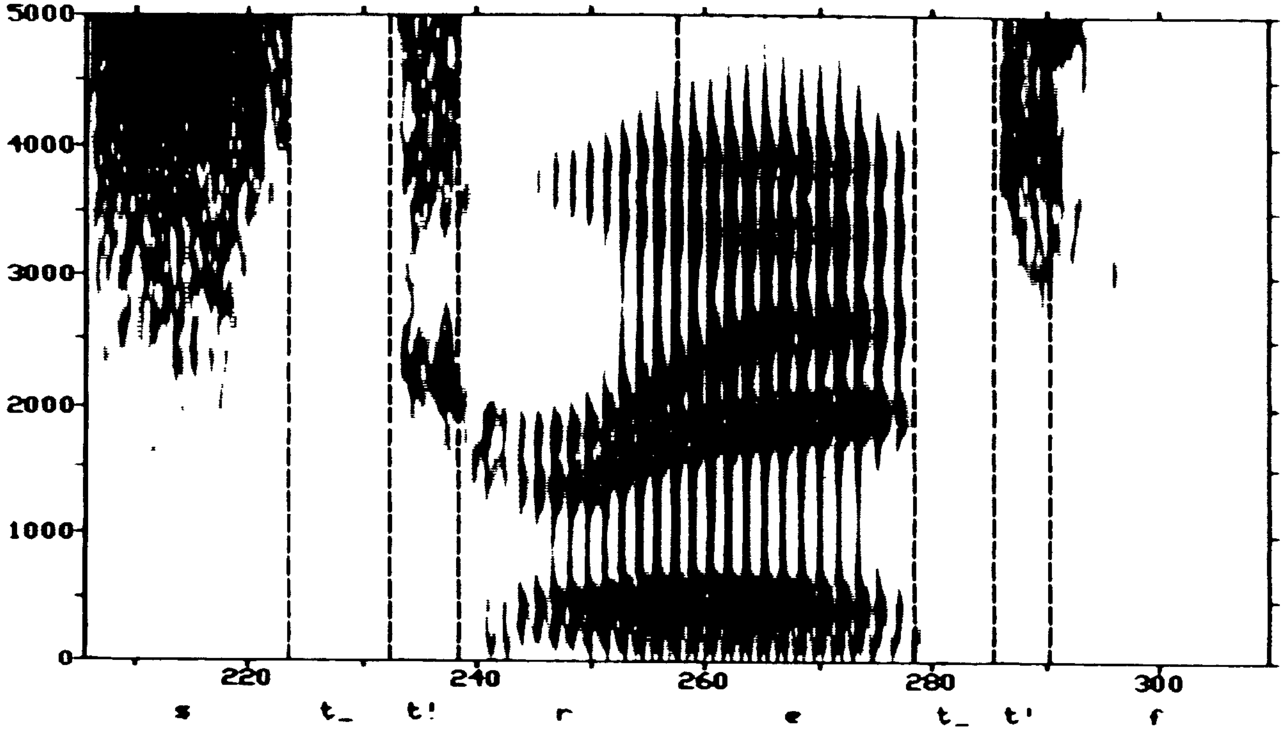


Fig. 3-1d. Synthetic speech with rule-based segmentation and labelling.

Sentence 1: (The goose was brought straight f)rom the ol(d market).

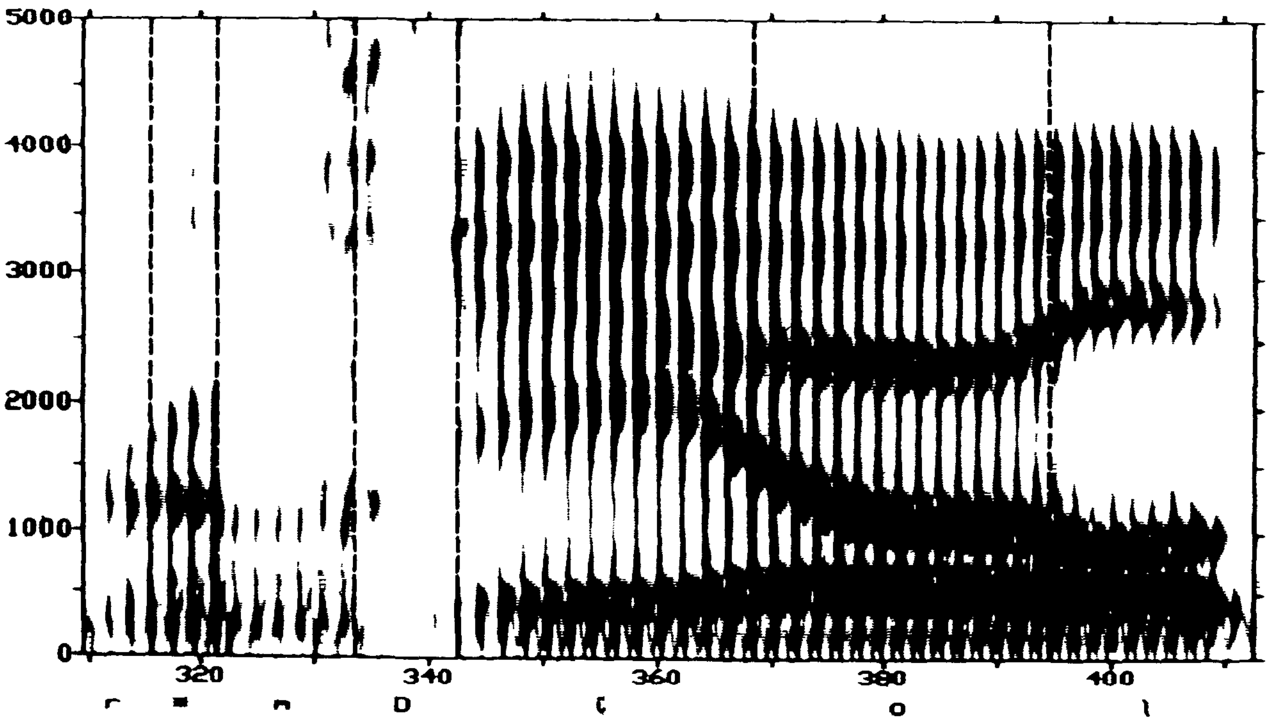


Fig. 3-1e. Synthetic speech with rule-based segmentation and labelling.

Sentence 1: (The goose was brought straight from the ol)d market.

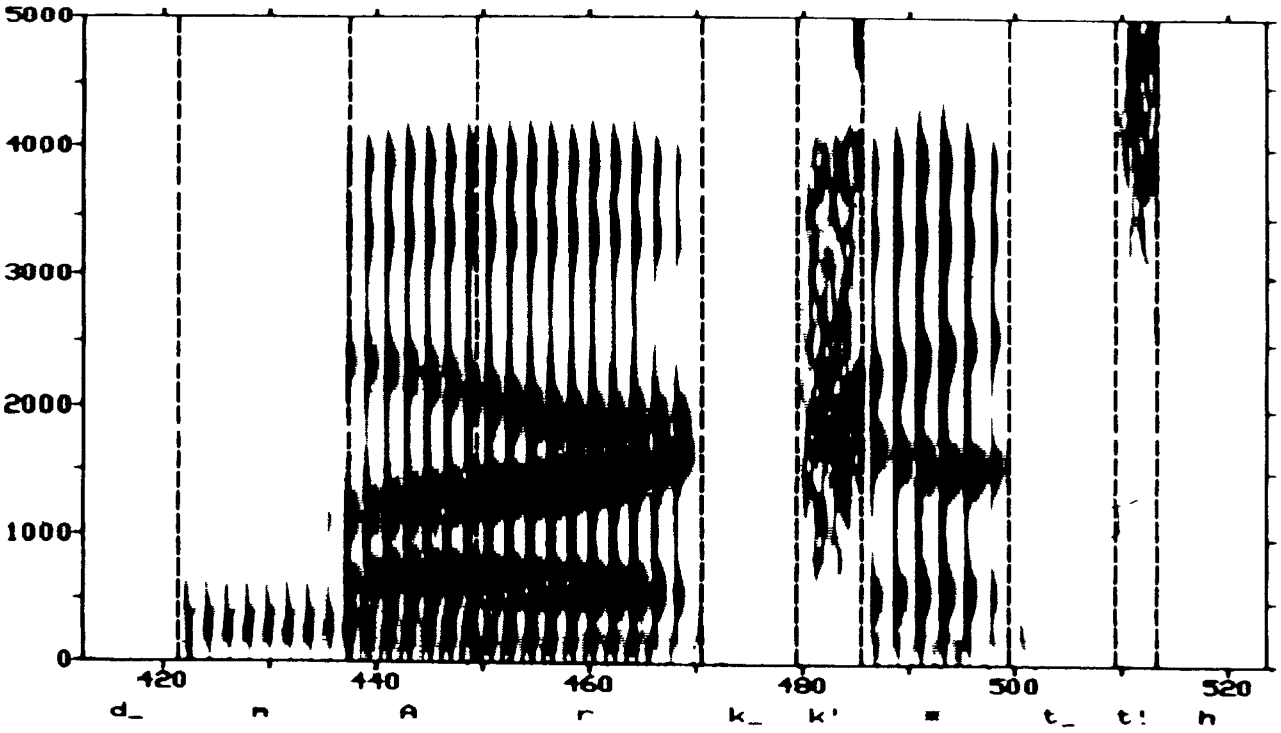


Fig. 3-2a. Synthetic speech with rule-based segmentation and labelling.

Sentence 2: The sink (is the thing in which we pile dishes).

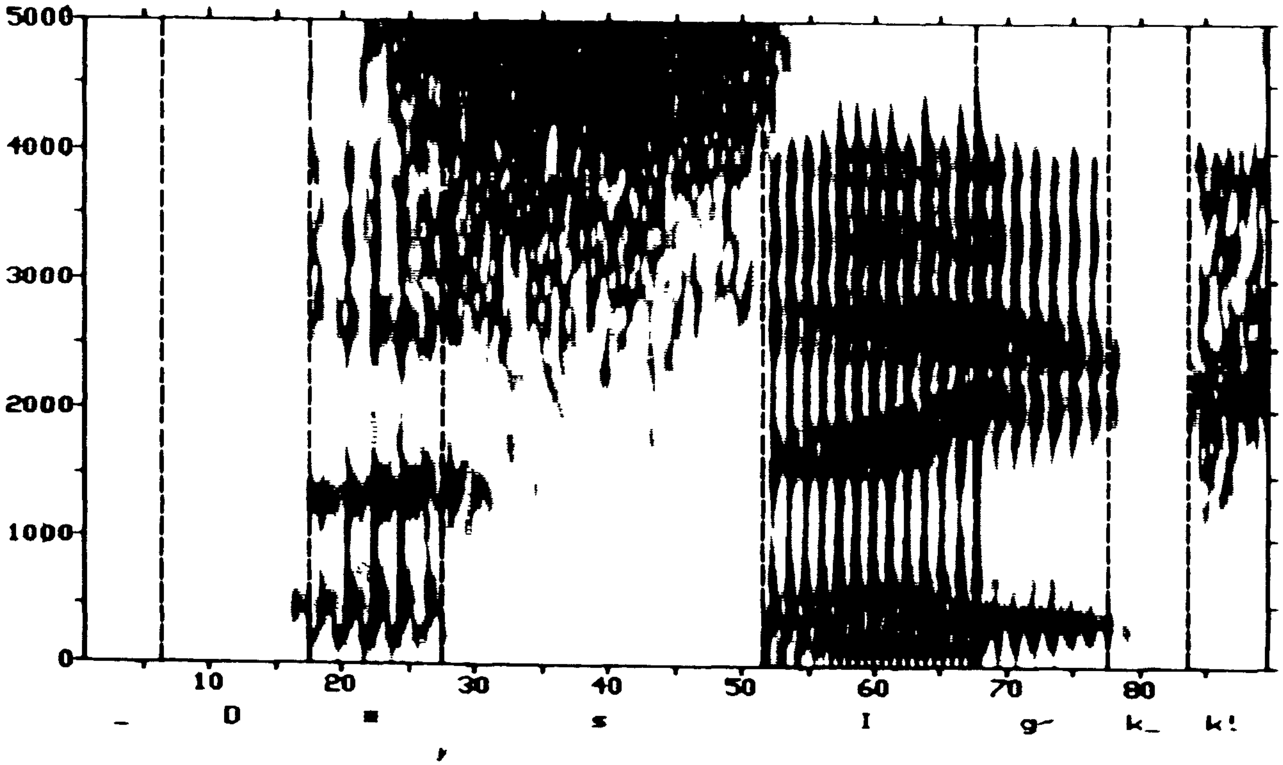


Fig. 3-2b. Synthetic speech with rule-based segmentation and labelling.

Sentence 2: (The sink) is the thing in (which we pile dishes).

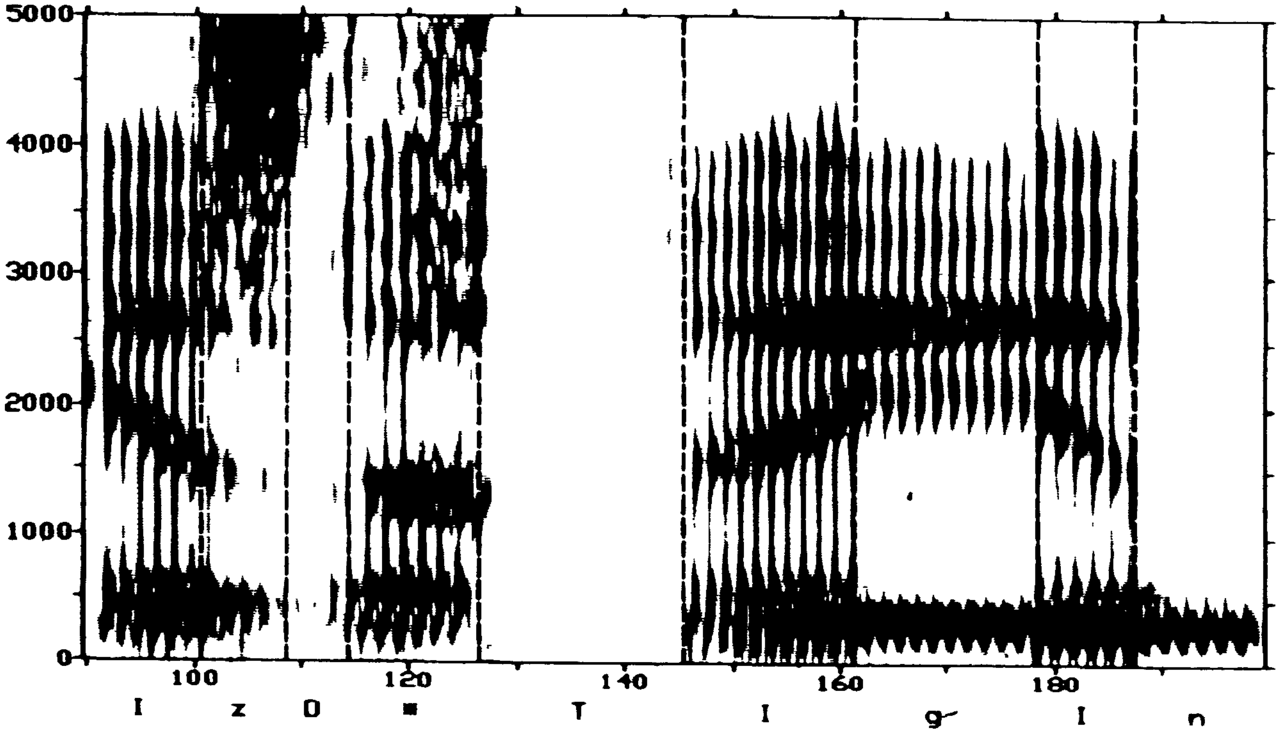


Fig. 3-2c. Synthetic speech with rule-based segmentation and labelling.

Sentence 2: (The sink is the thing in) which we p(ile dishes).

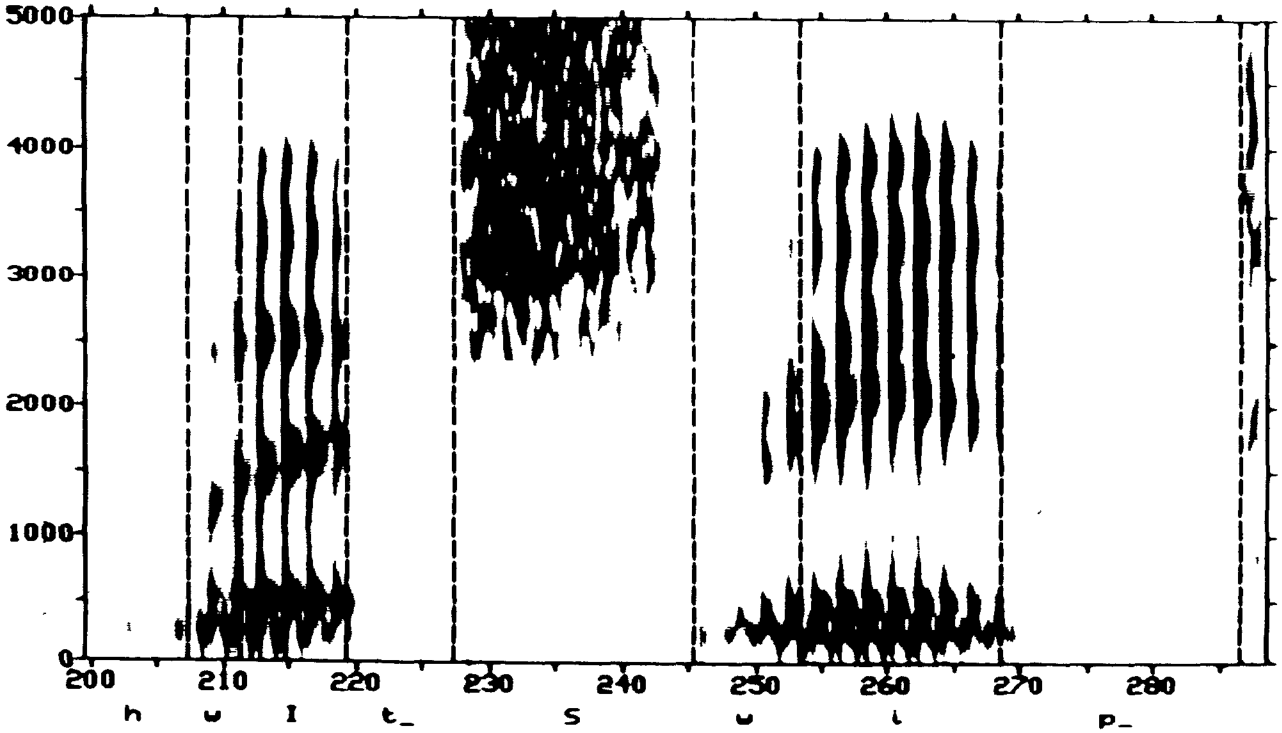


Fig. 3-2d. Synthetic speech with rule-based segmentation and labelling.

Sentence 2: (The sink is the thing in which we p)ile di(shes).

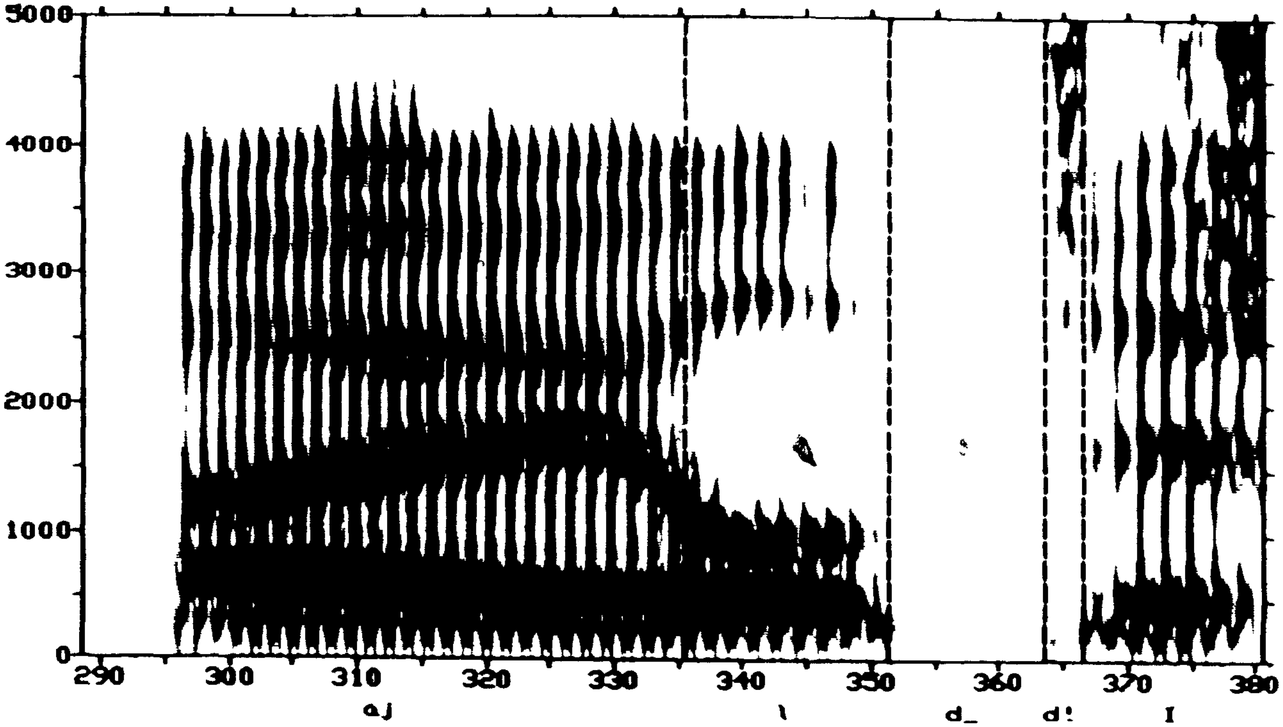


Fig. 3-2e. Synthetic speech with rule-based segmentation and labelling.

Sentence 2: (The sink is the thing in which we pile di)shes.

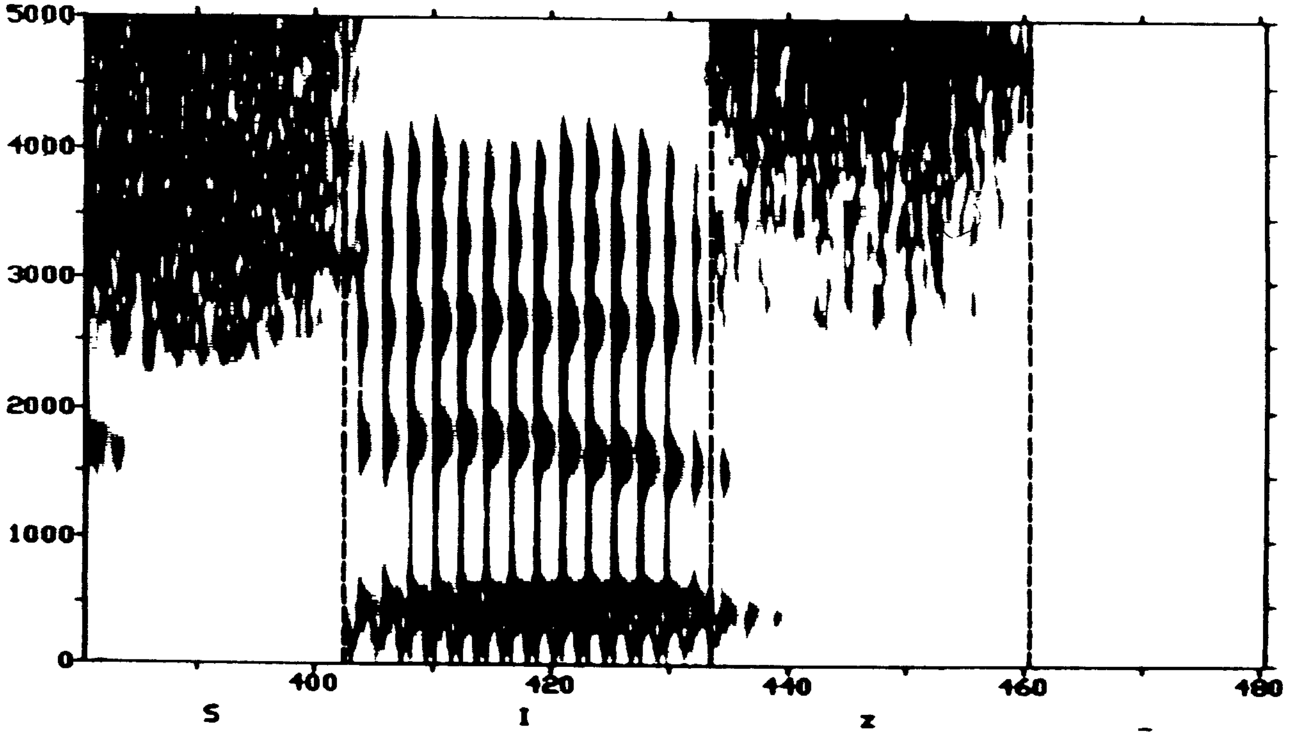


Fig. 3-3a. Synthetic speech with rule-based segmentation and labelling.

Sentence 3: A whiff of (it will cure the most stubborn cold).

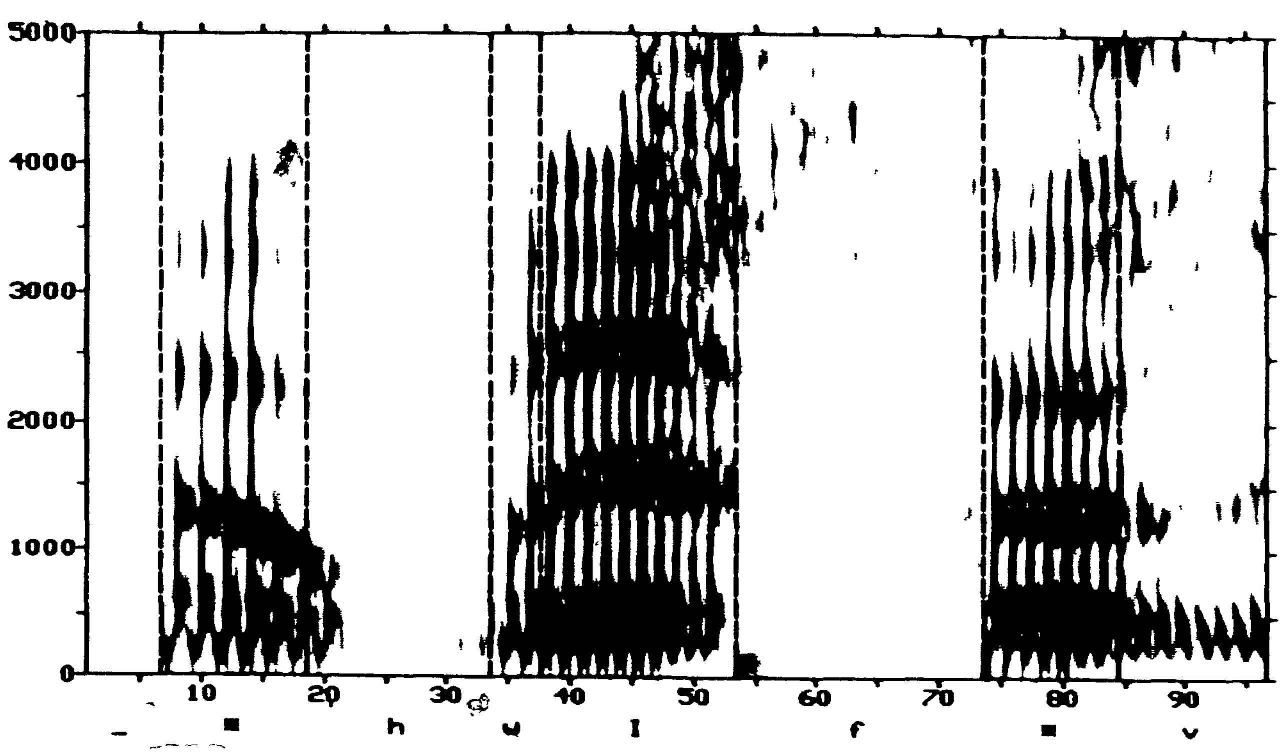


Fig. 3-3b. Synthetic speech with rule-based segmentation and labelling.

Sentence 3: (A whiff of) it will c(ure the most stubborn cold).

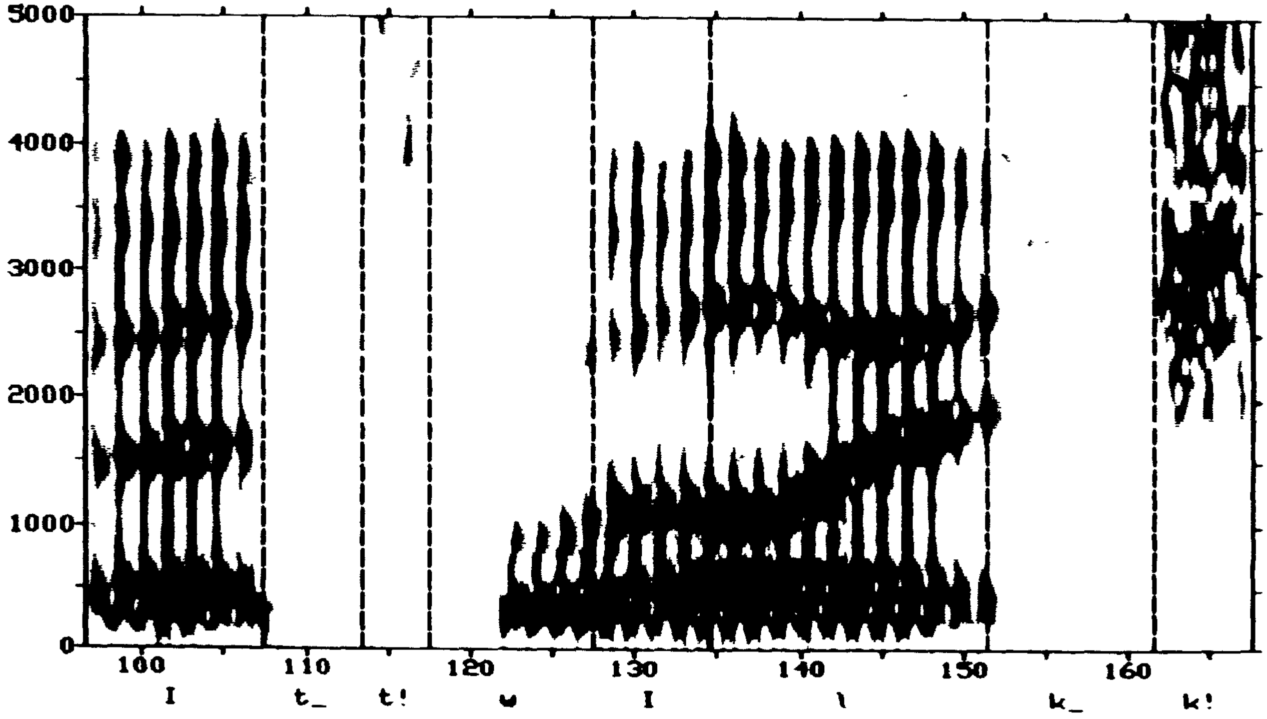


Fig. 3-3c. Synthetic speech with rule-based segmentation and labelling.

Sentence 3: (A whiff of it will c)ure the mo(st stubborn cold).

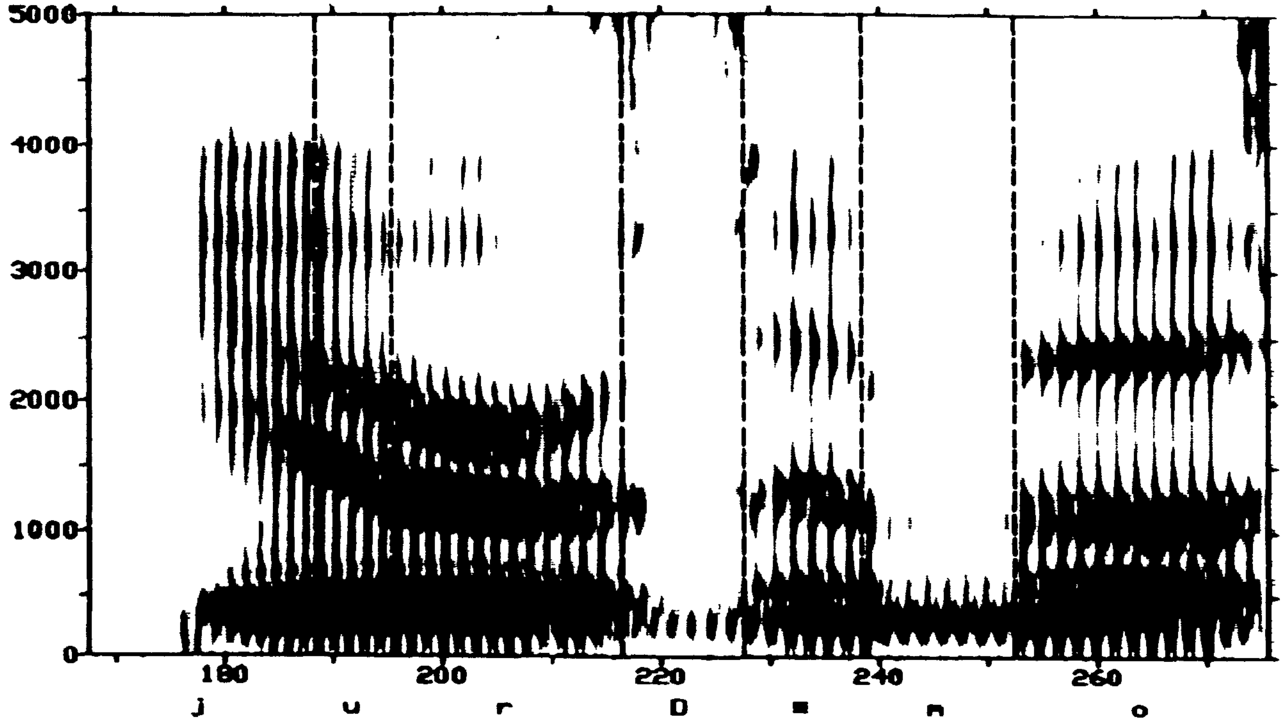


Fig. 3-3d. Synthetic speech with rule-based segmentation and labelling.

Sentence 3: (A whiff of it will cure the mo)st stubb(orn cold).

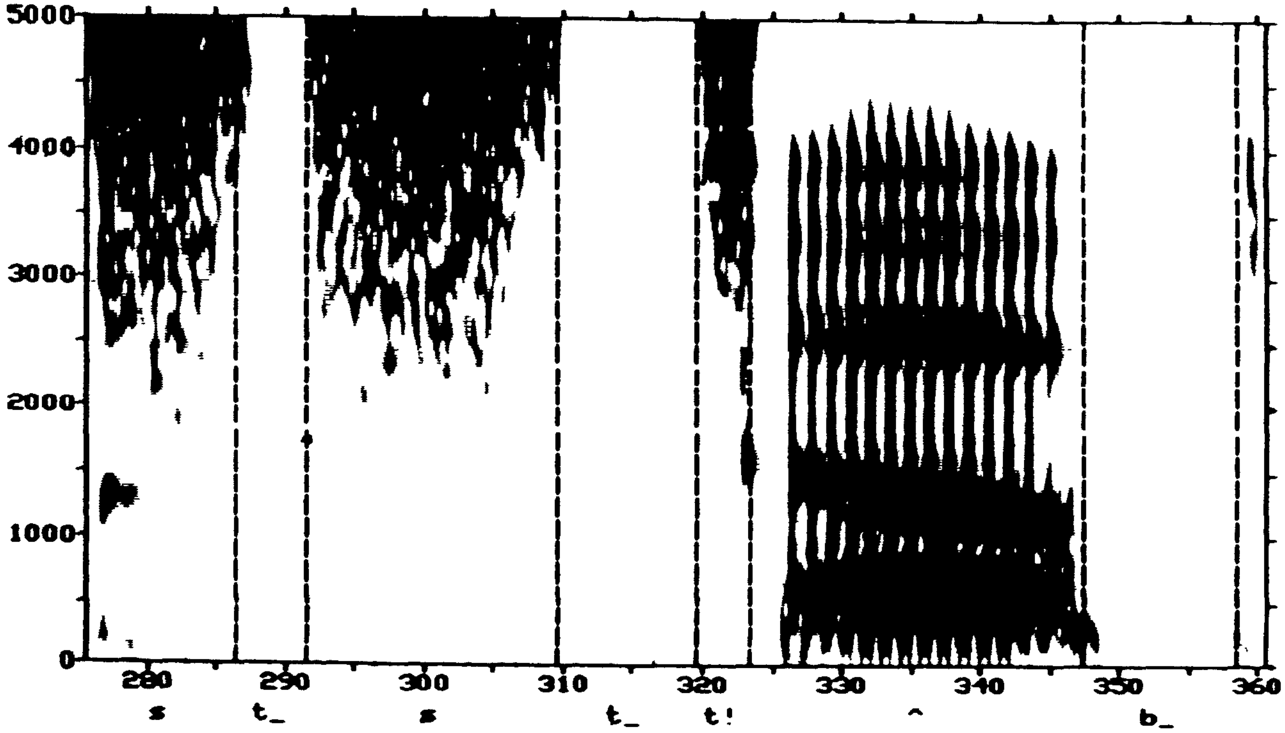


Fig. 3-3e. Synthetic speech with rule-based segmentation and labelling.

Sentence 3: (A whiff of it will cure the most stubb)orn cold.

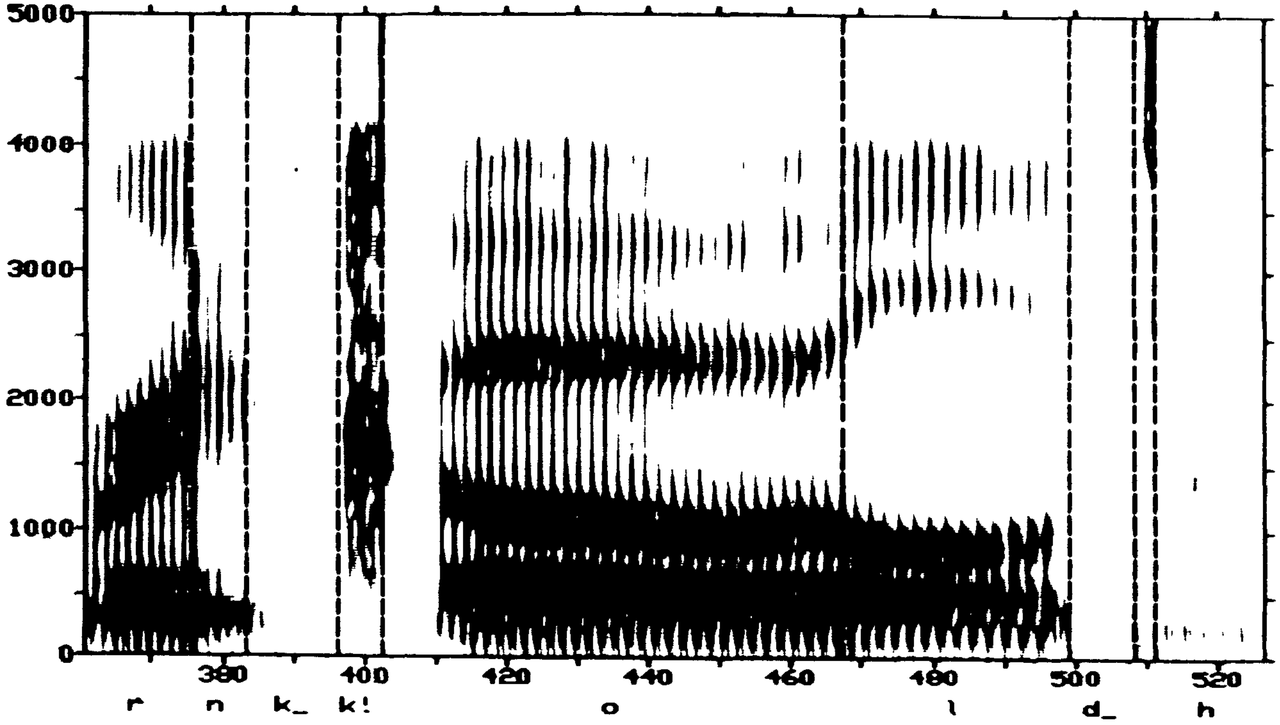


Fig. 3-4a. Synthetic speech with rule-based segmentation and labelling.

Sentence 4: The facts (don´t always show who is right).

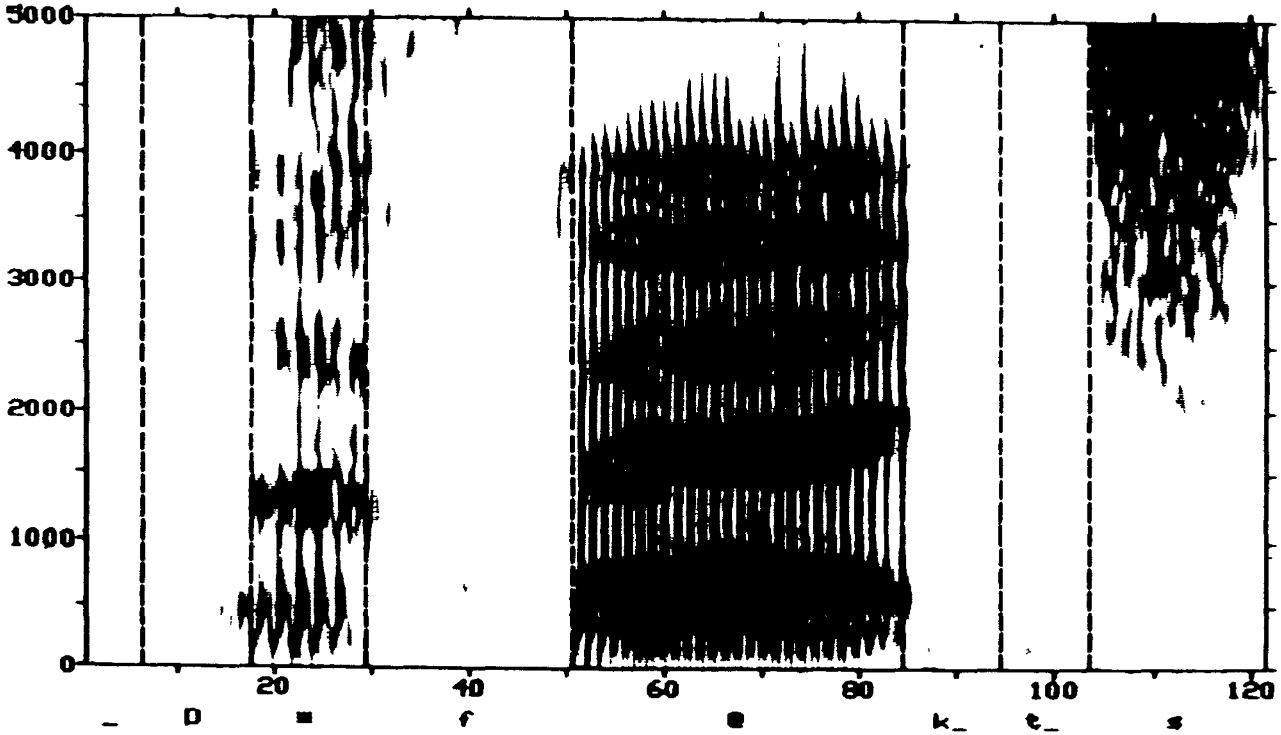


Fig. 3-4b. Synthetic speech with rule-based segmentation and labelling.

Sentence 4: (The facts) don´t (always show who is right).

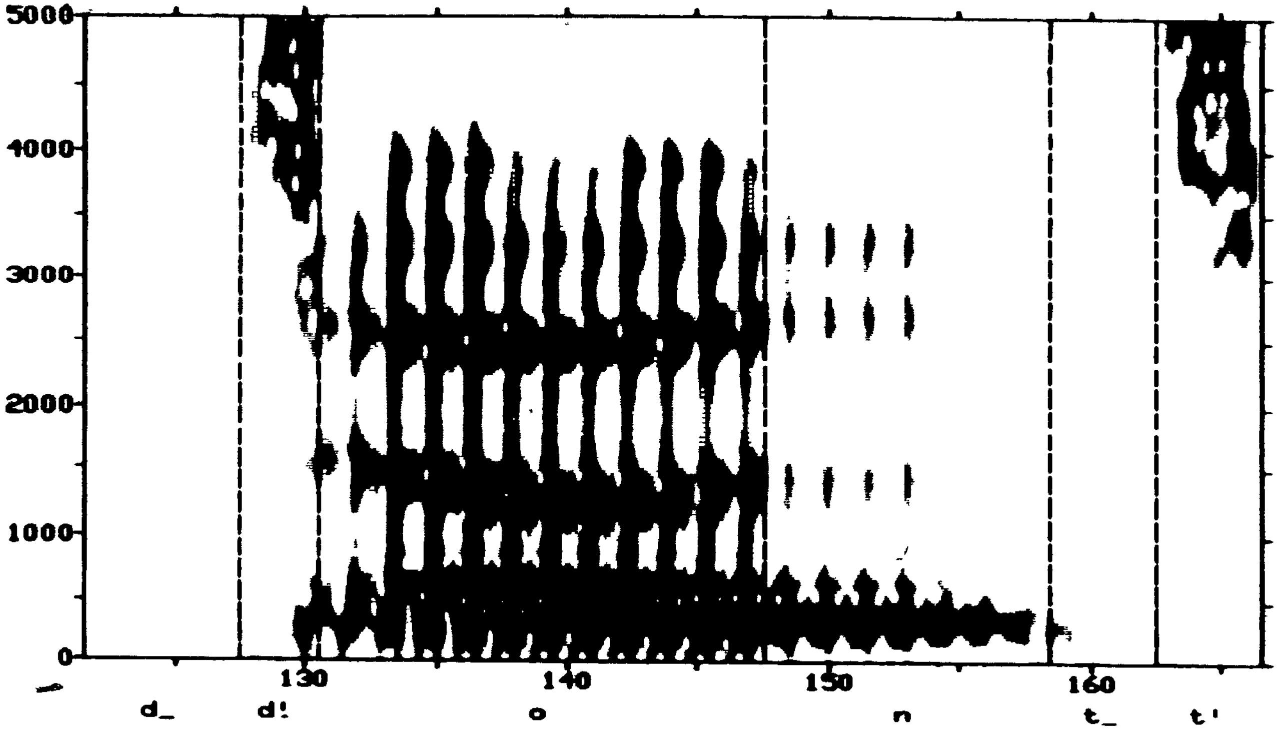


Fig. 3-4c. Synthetic speech with rule-based segmentation and labelling.

Sentence 4: (The facts don´t) always sh(ow who is right).

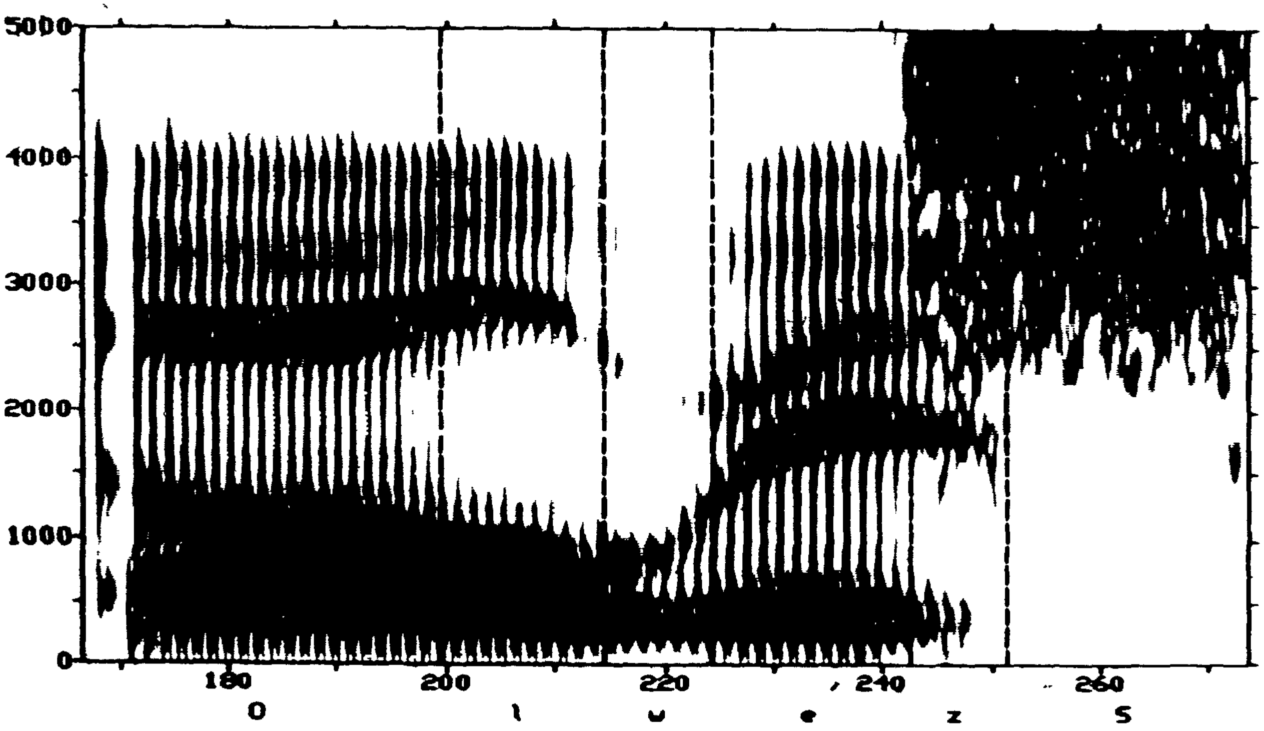


Fig. 3-4d. Synthetic speech with rule-based segmentation and labelling.

Sentence 4: (The facts don´t always sh)ow who is (right).

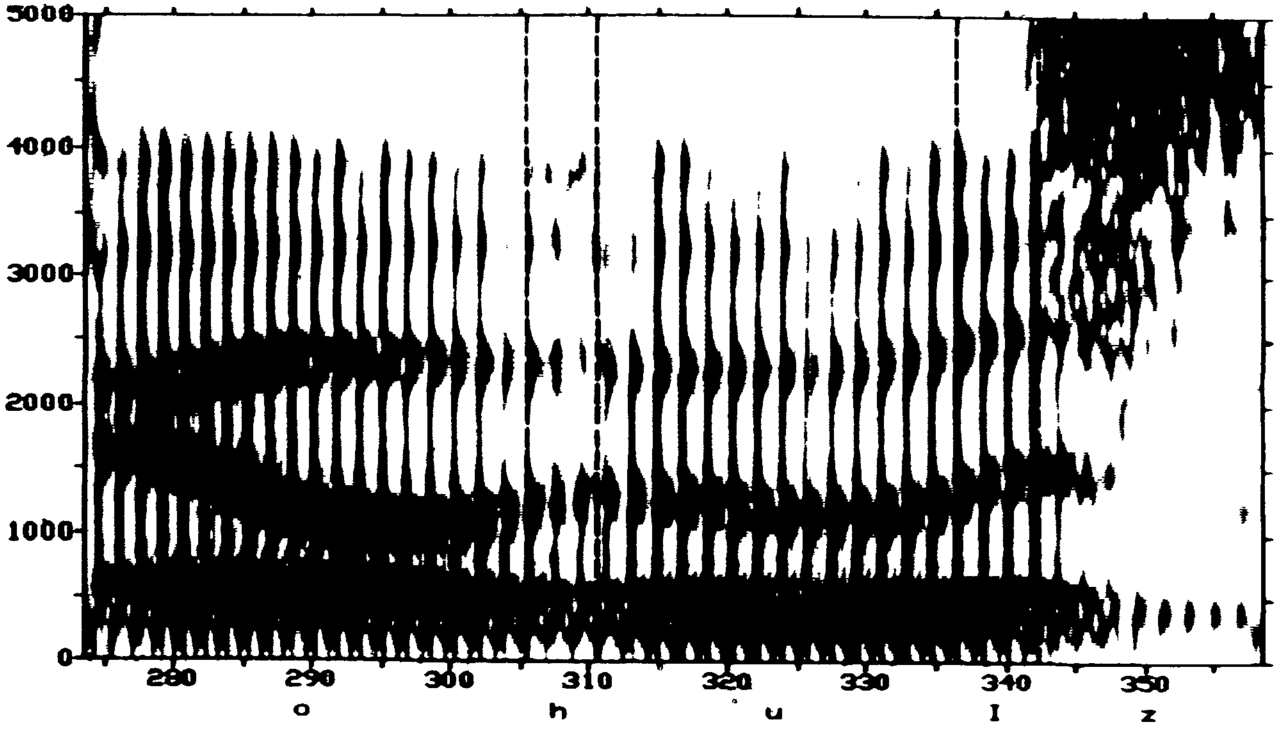


Fig. 3-4e. Synthetic speech with rule-based segmentation and labelling.

Sentence 4: (The facts don´t always show who is) rıght.

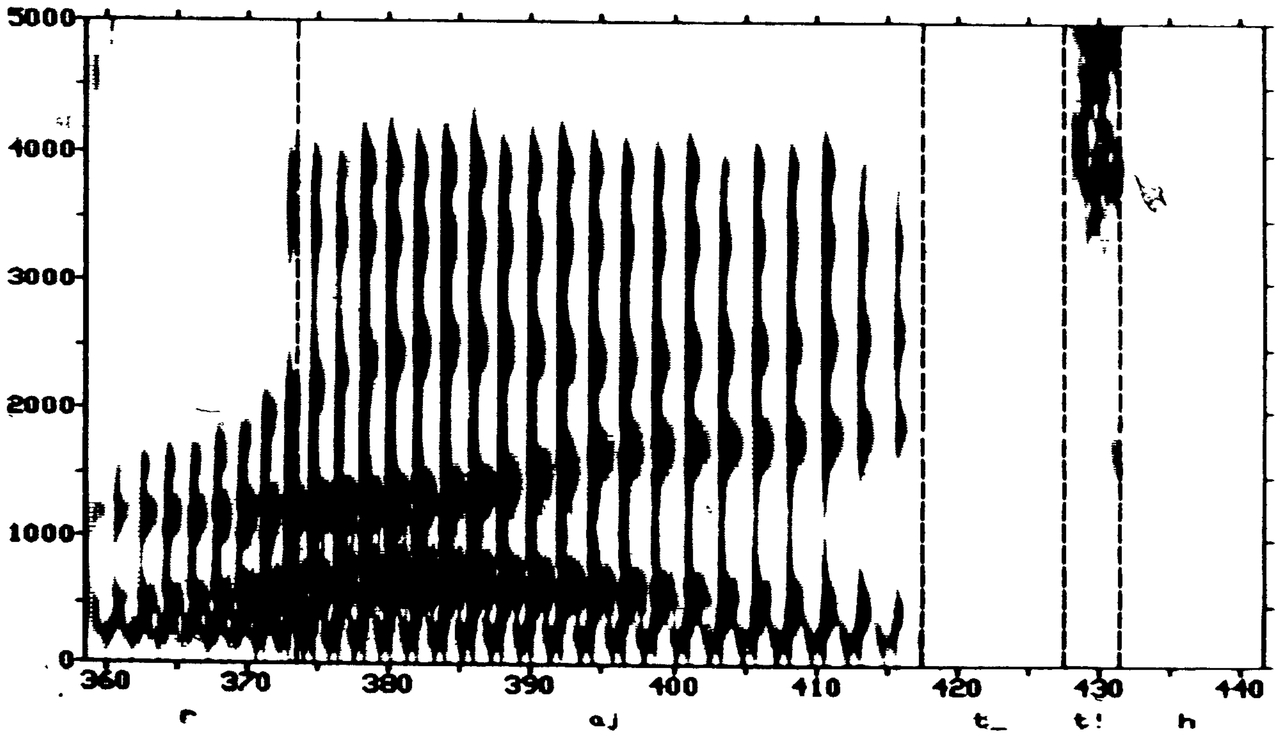


Fig. 3-5a. Synthetic speech with rule-based segmentation and labelling.

Sentence 5: She fla(ps her cape as she parades the street).

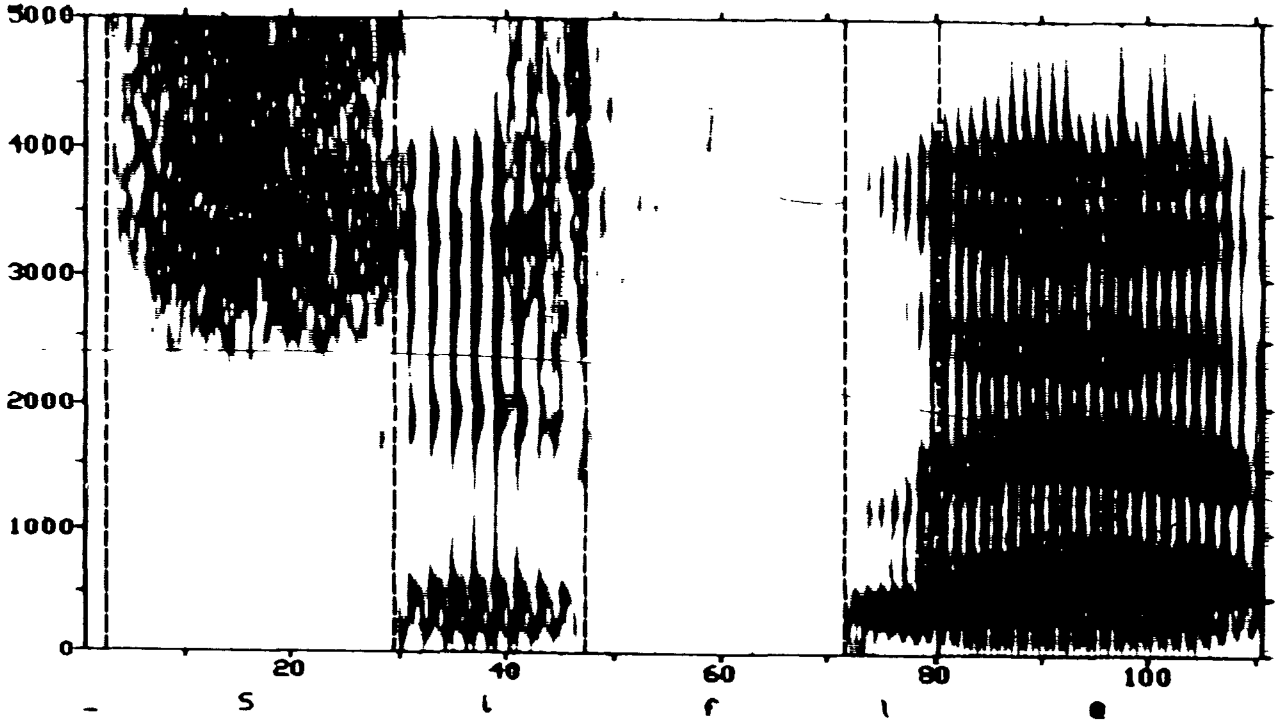


Fig. 3-5b. Synthetic speech with rule-based segmentation and labelling.

Sentence 5: (She fla)ps her cape (as she parades the street)

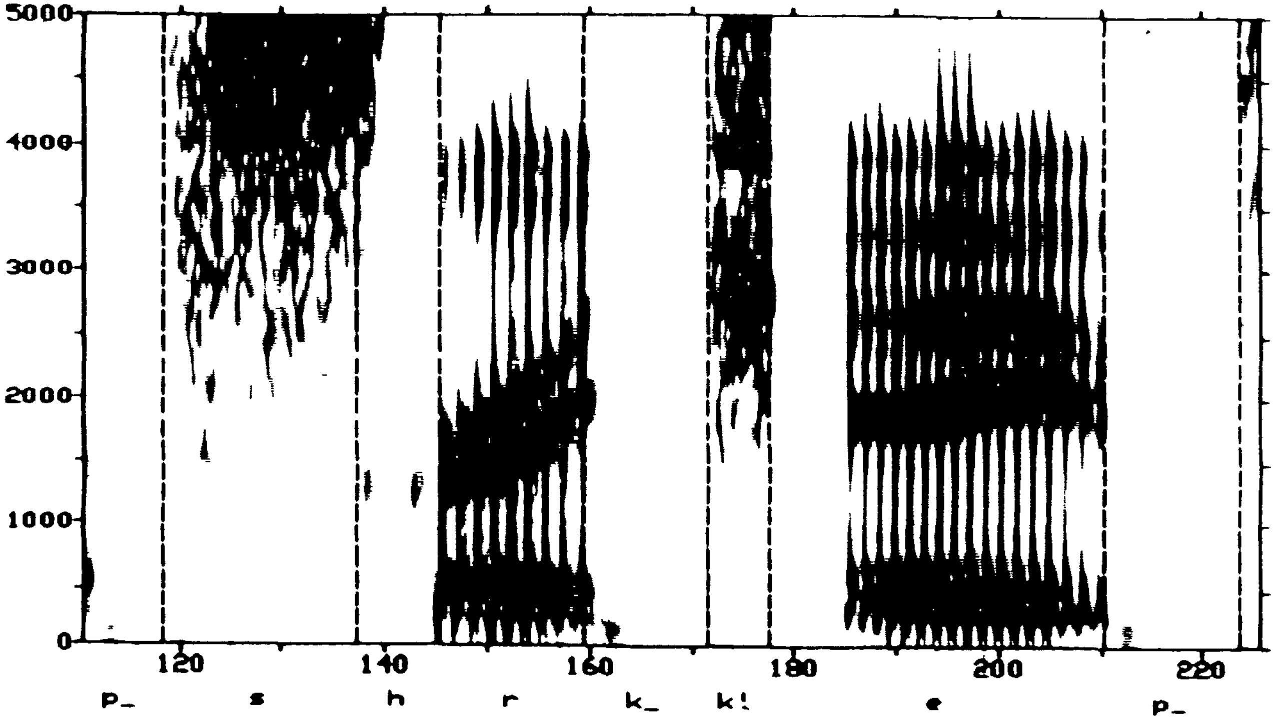


Fig. 3-5c. Synthetic speech with rule-based segmentation and labelling.

Sentence 5: (She flaps her cape) as she para(des the street).

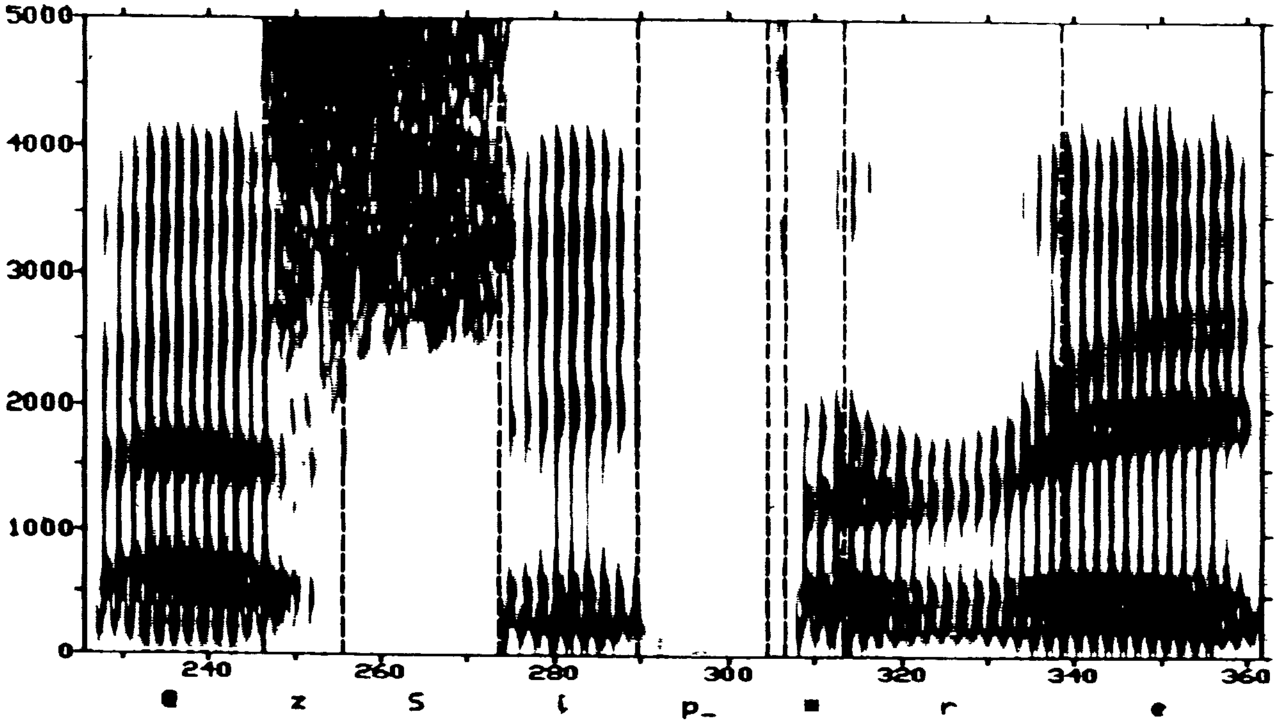


Fig. 3-5d. Synthetic speech with rule-based segmentation and labelling.

Sentence 5: (She flaps her cape as she para)des the street.

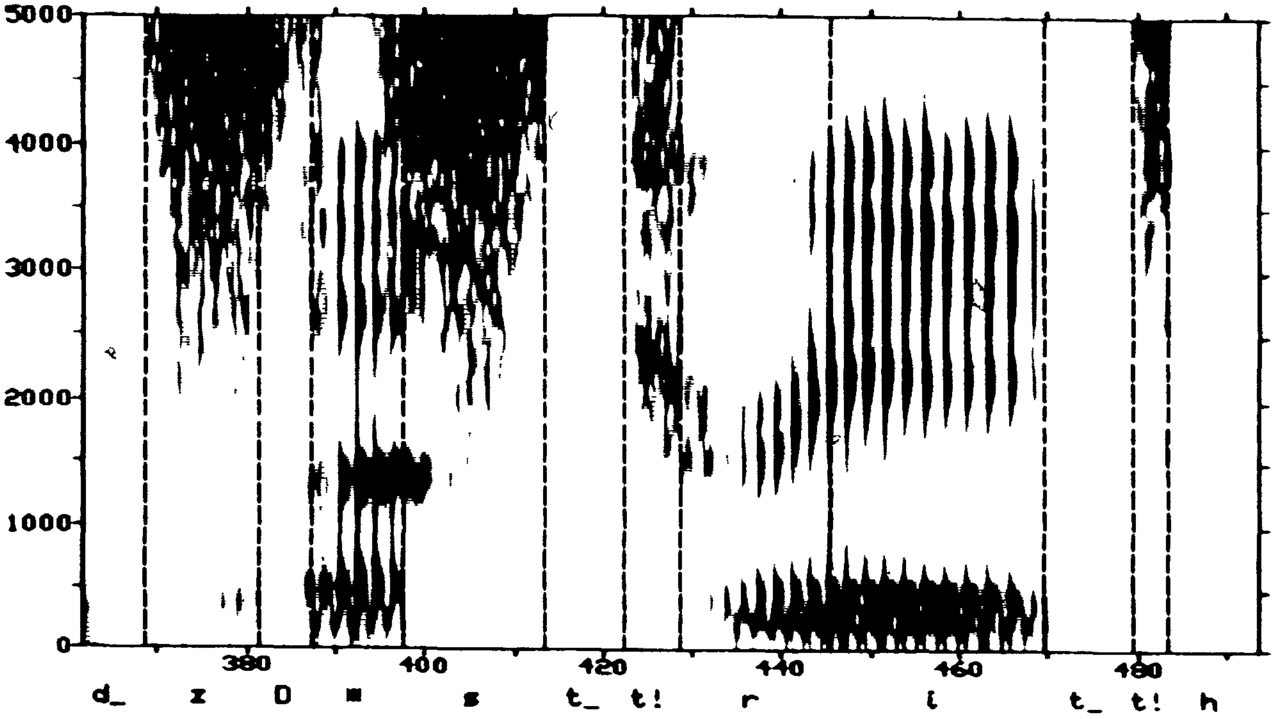


Fig. 3-6a. Synthetic speech with rule-based segmentation and labelling.

Sentence 6: The loss (of the cruiser was a blow to the fleet).

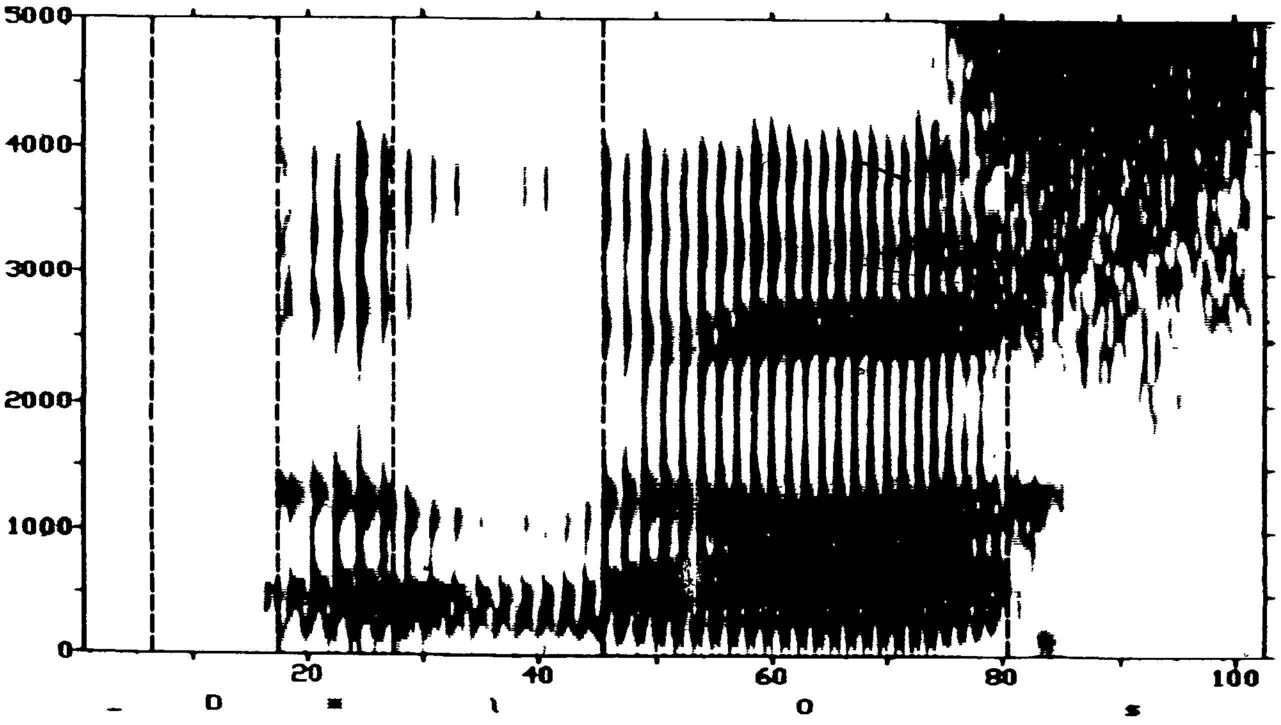


Fig. 3-6b. Synthetic speech with rule-based segmentation and labelling.

Sentence 6: (The loss) of the cruis(er was a blow to the fleet).

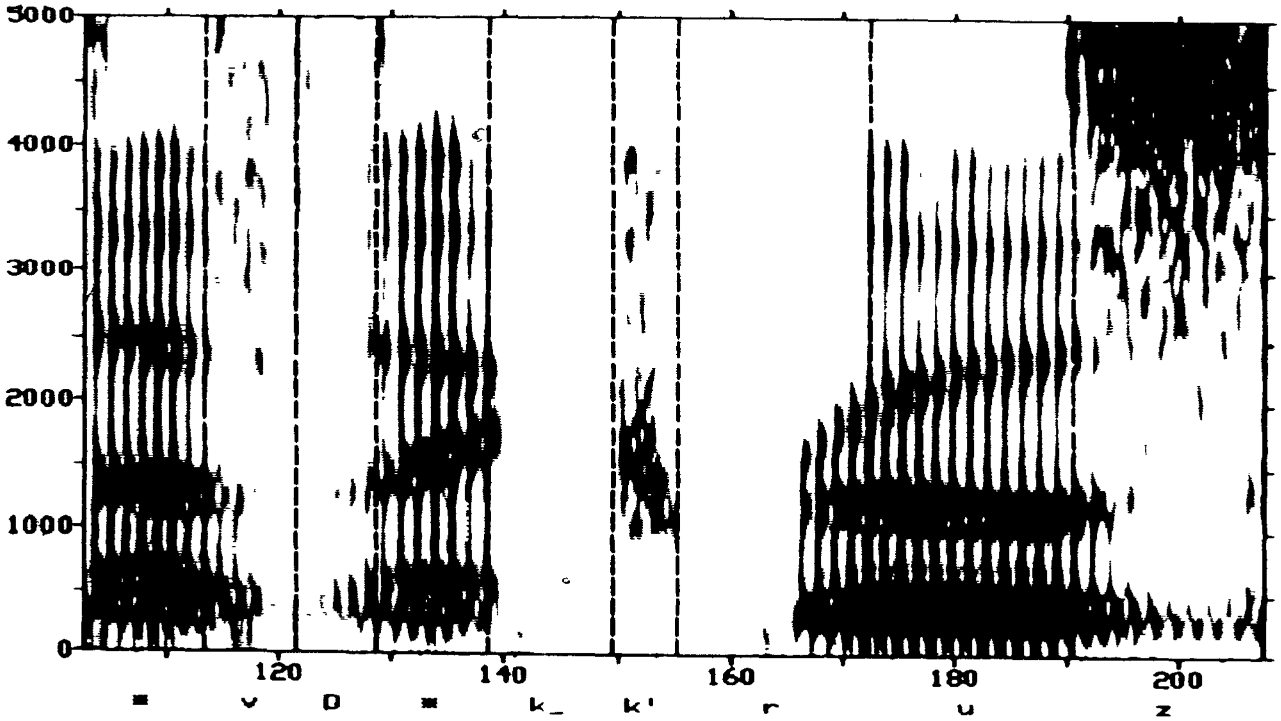


Fig. 3-6c. Synthetic speech with rule-based segmentation and labelling.

Sentence 6: (The loss of the cruis)er was (a blow to the fleet).

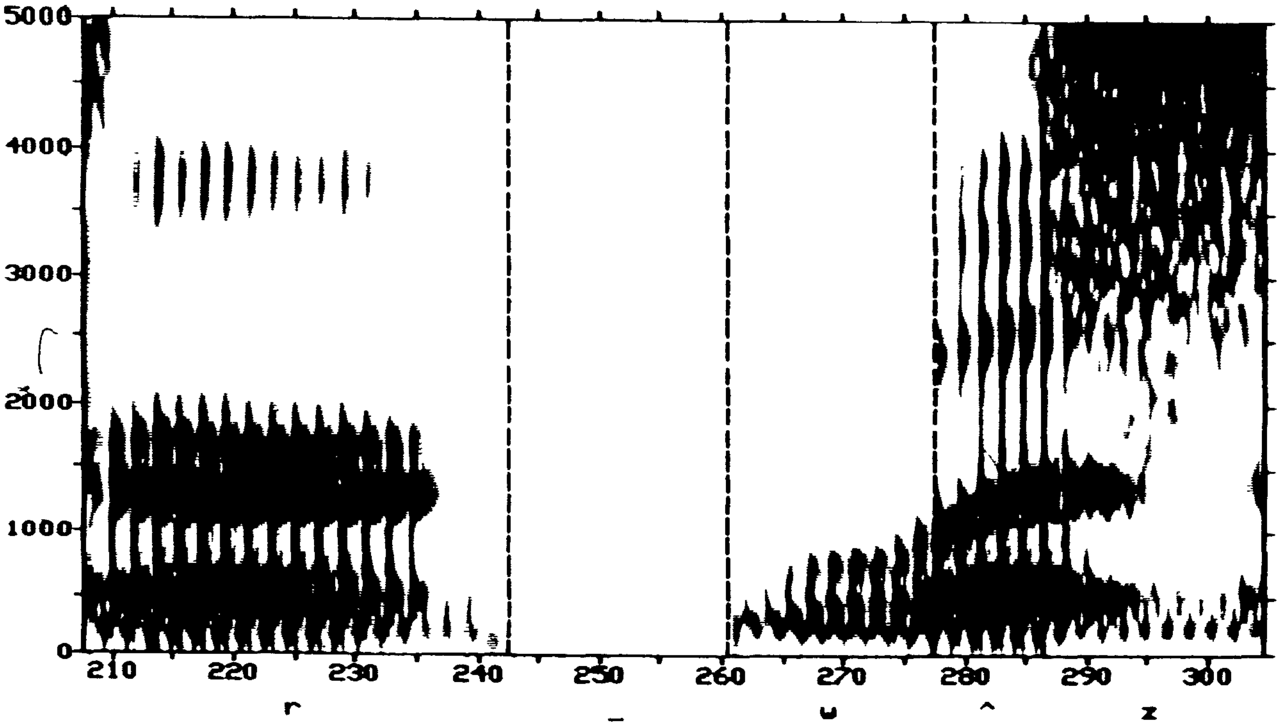


Fig. 3-6d. Synthetic speech with rule-based segmentation and labelling.

Sentence 6: (The loss of the cruiser was) a blow to (the fleet).

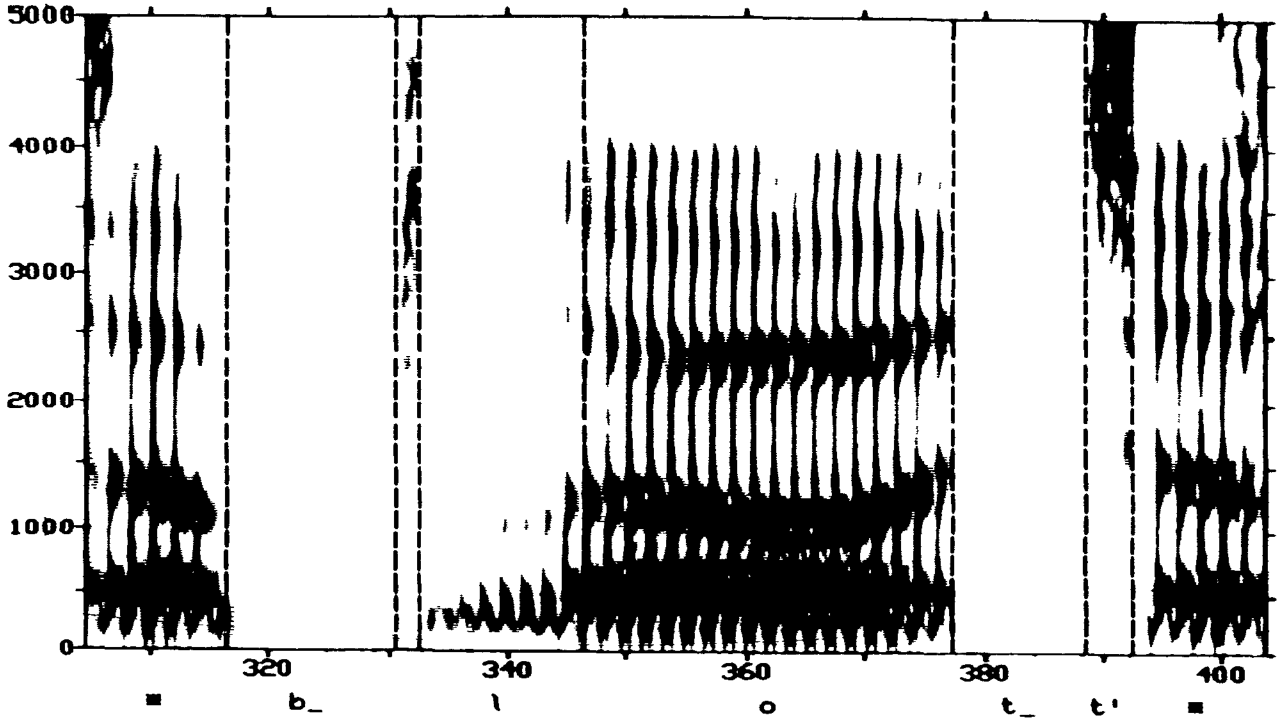


Fig. 3-6e. Synthetic speech with rule-based segmentation and labelling.

Sentence 6: (The loss of the cruiser was a blow to) the fleet.

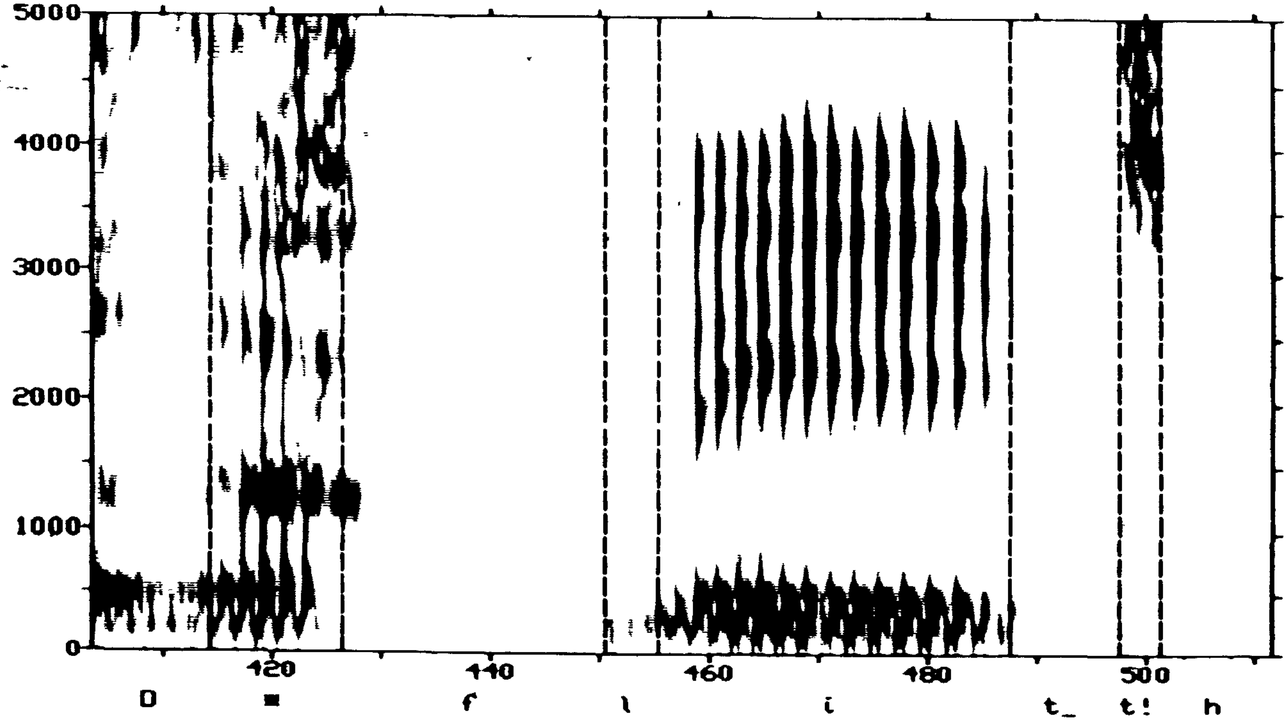


Fig. 3-7a. Synthetic speech with rule-based segmentation and labelling.

Sentence 7: Loop the b(raid to the left and then over).

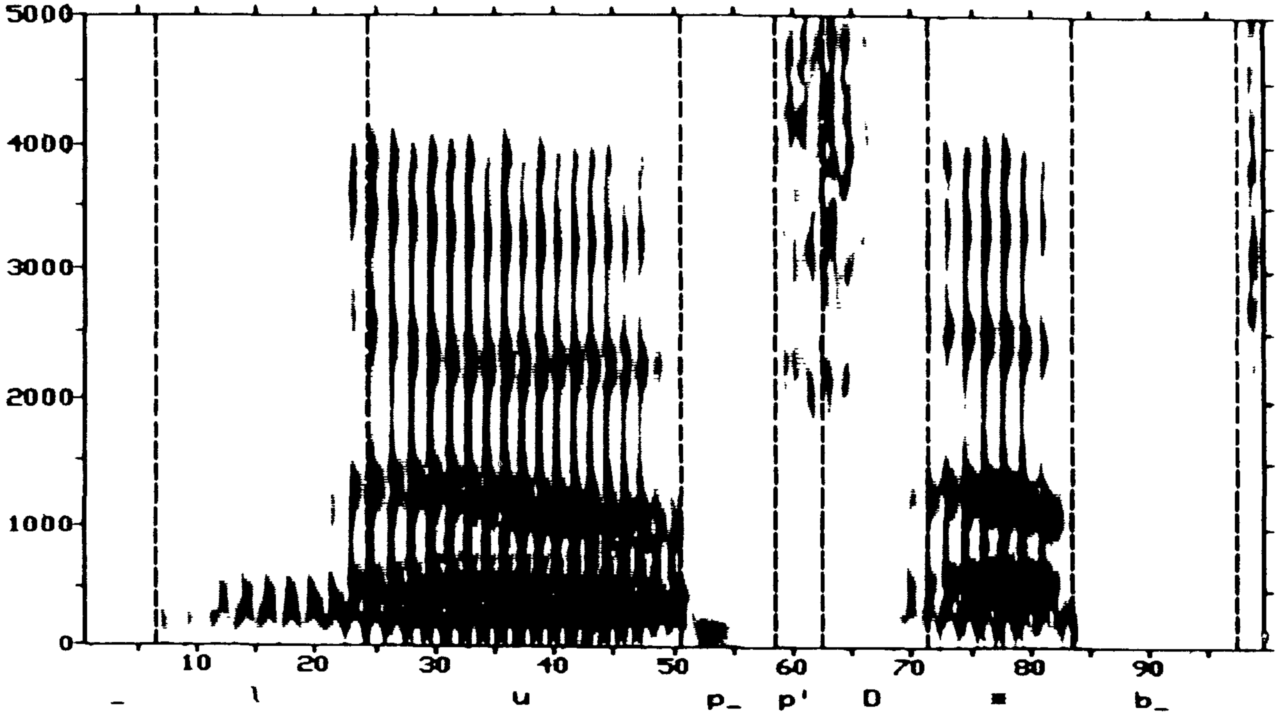


Fig. 3-7b. Synthetic speech with rule-based segmentation and labelling.

Sentence 7: (Loop the b)raid to th(e left and then over).

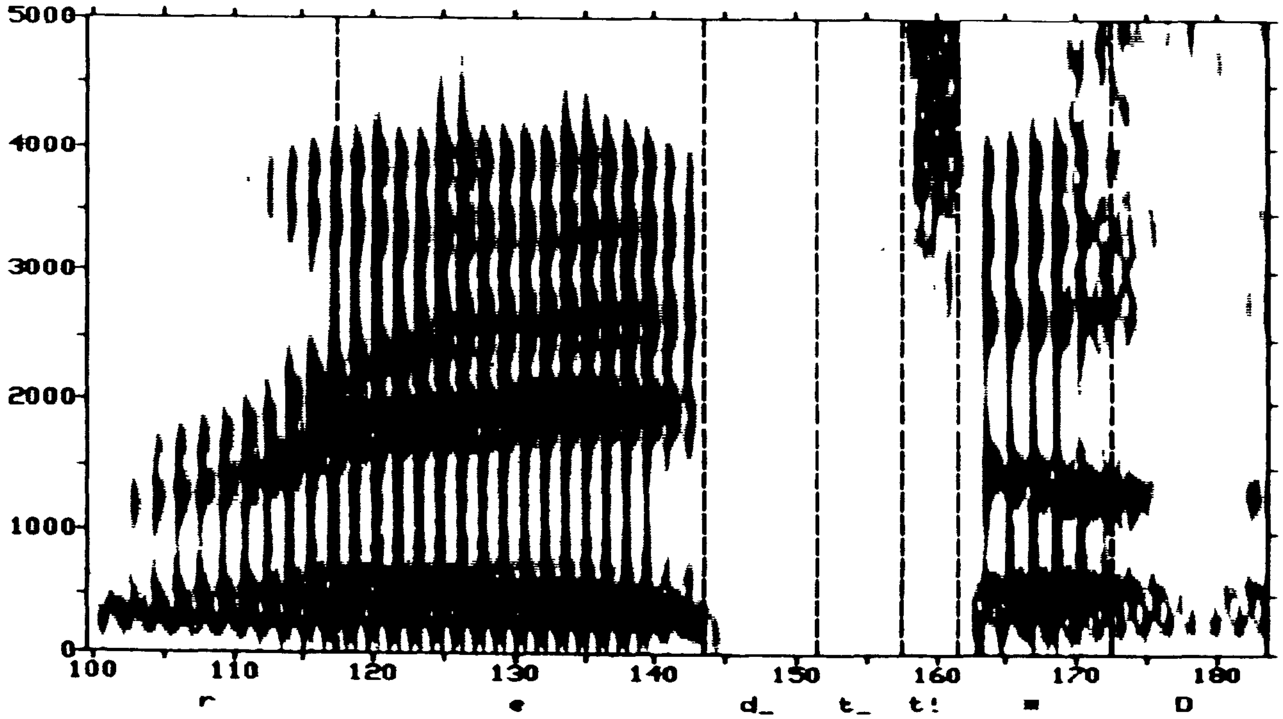


Fig. 3-7c. Synthetic speech with rule-based segmentation and labelling.

Sentence 7: (Loop the braid to th)e left (and then over).

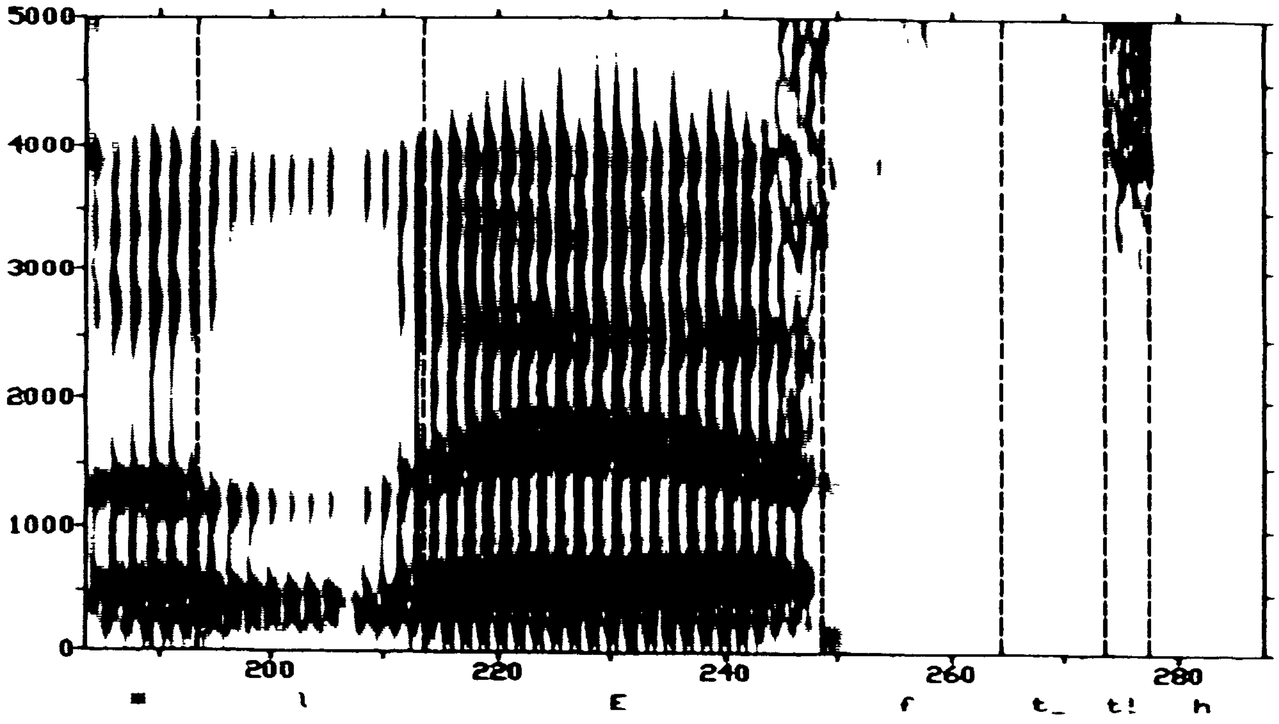


Fig. 3-7d. Synthetic speech with rule-based segmentation and labelling.

Sentence 7: (Loop the braid to the left) and then (over).

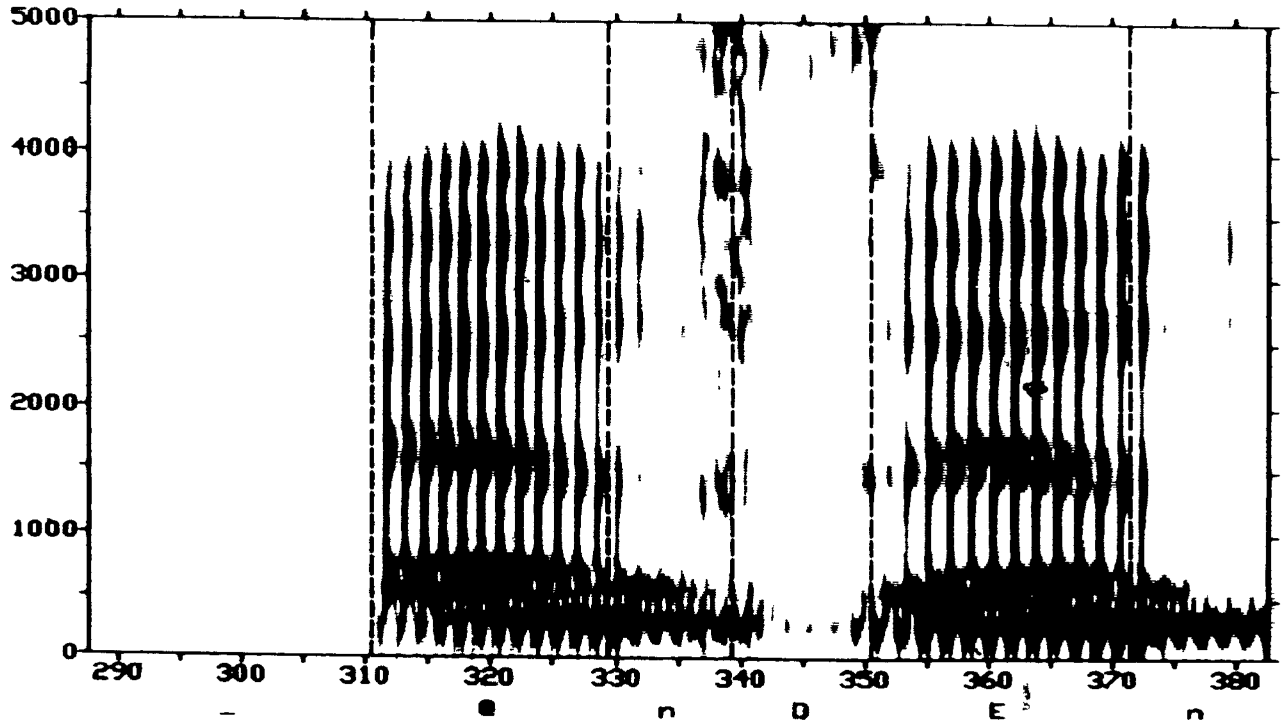


Fig. 3-7e. Synthetic speech with rule-based segmentation and labelling.

Sentence 7: (Loop the braid to the left and then) over.

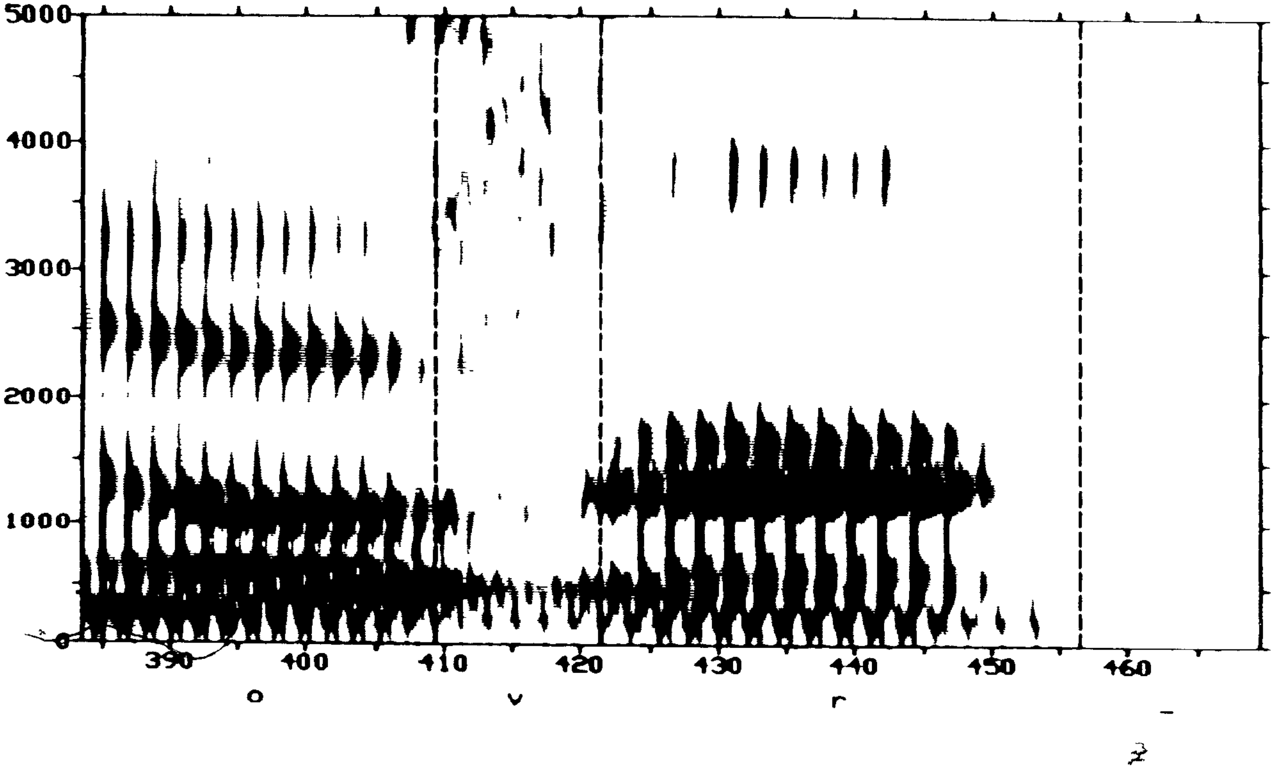


Fig. 3-8a. Synthetic speech with rule-based segmentation and labelling.

Sentence 8: Plead with (the lawyer to drop the lost cause).

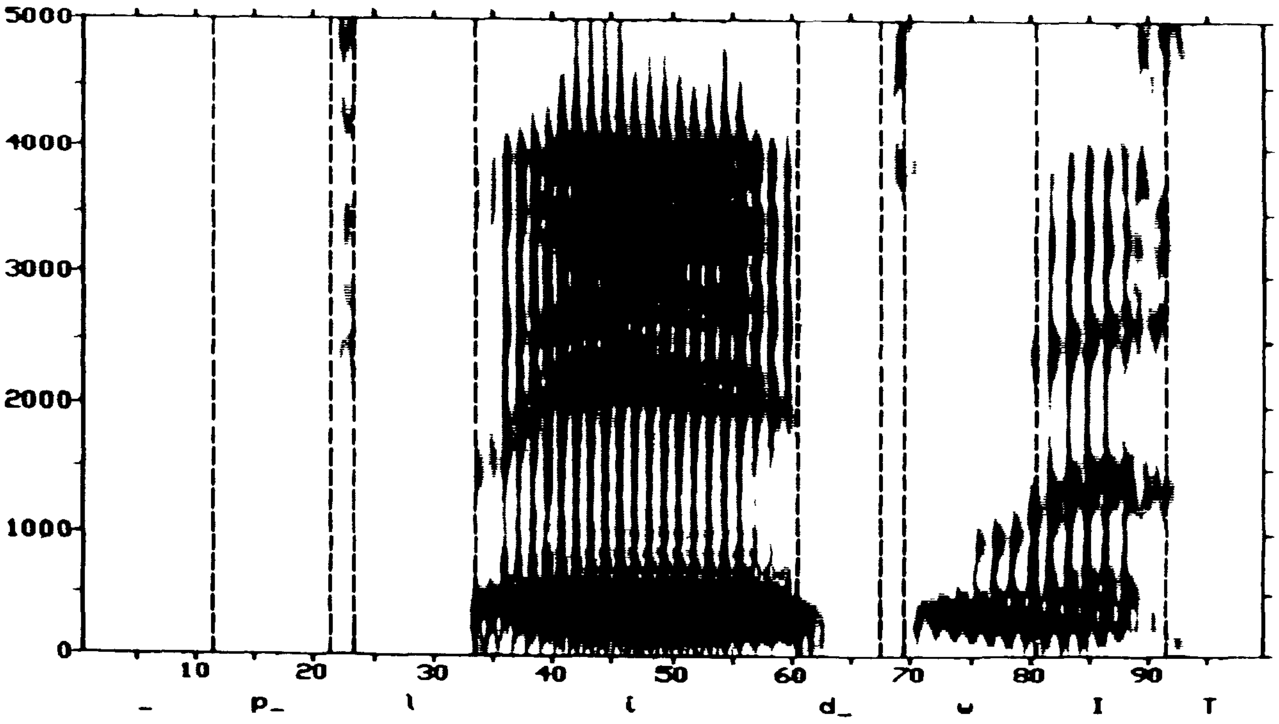


Fig. 3-8b. Synthetic speech with rule-based segmentation and labelling.

Sentence 8: (Plead with) the lawyer (to drop the lost cause).

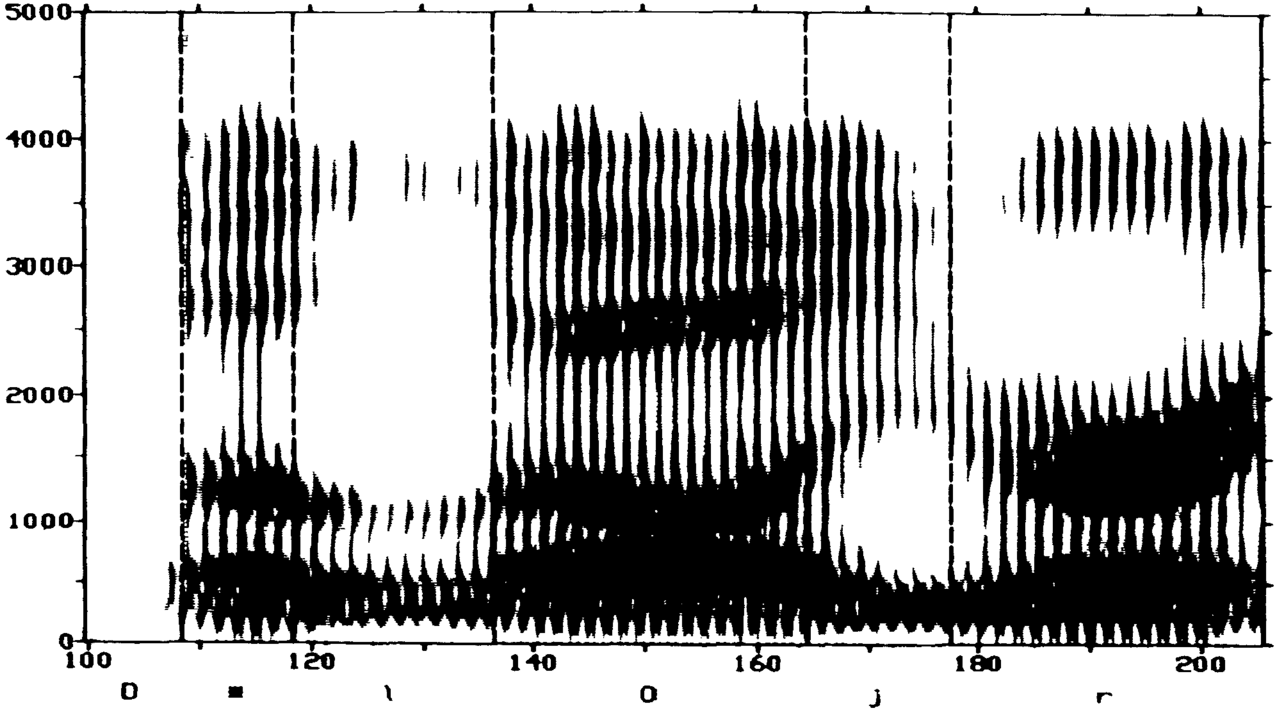


Fig. 3-8c. Synthetic speech with rule-based segmentation and labelling.

Sentence 8: (Plead with the lawyer) to drop (the lost cause).

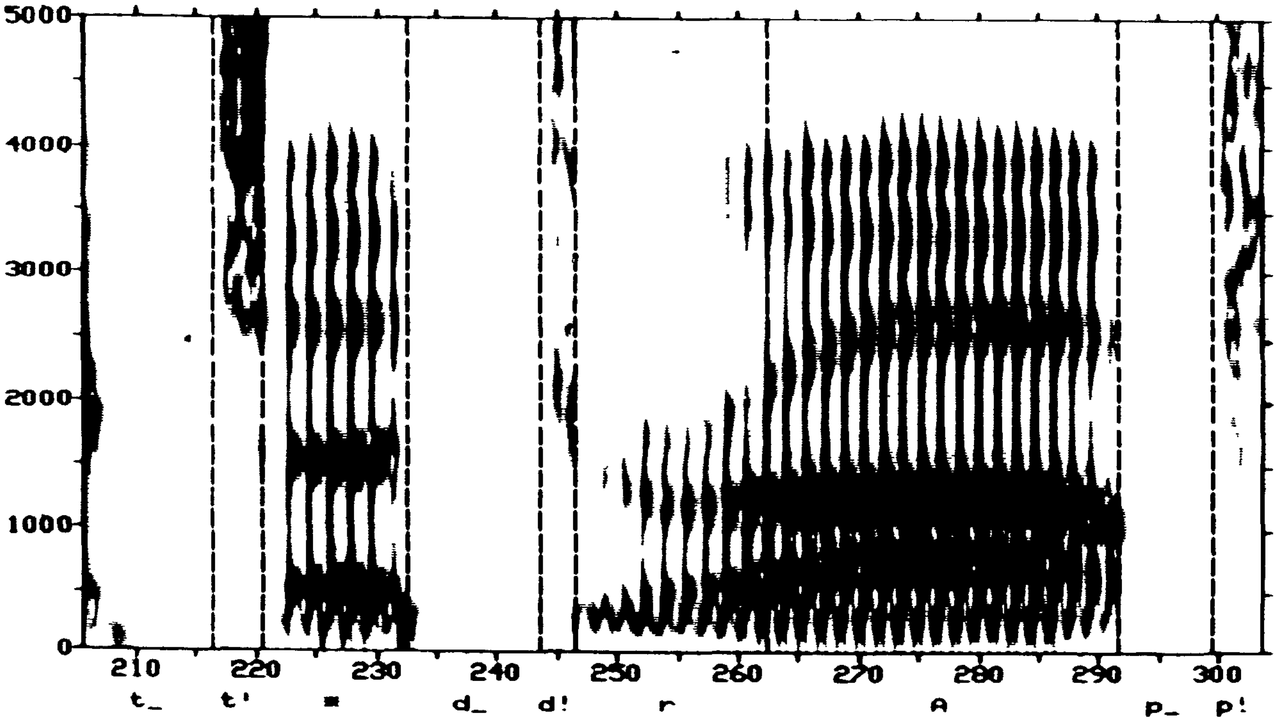


Fig. 3-8d. Synthetic speech with rule-based segmentation and labelling.

Sentence 8: (Plead with the lawyer to drop) the lost (cause).

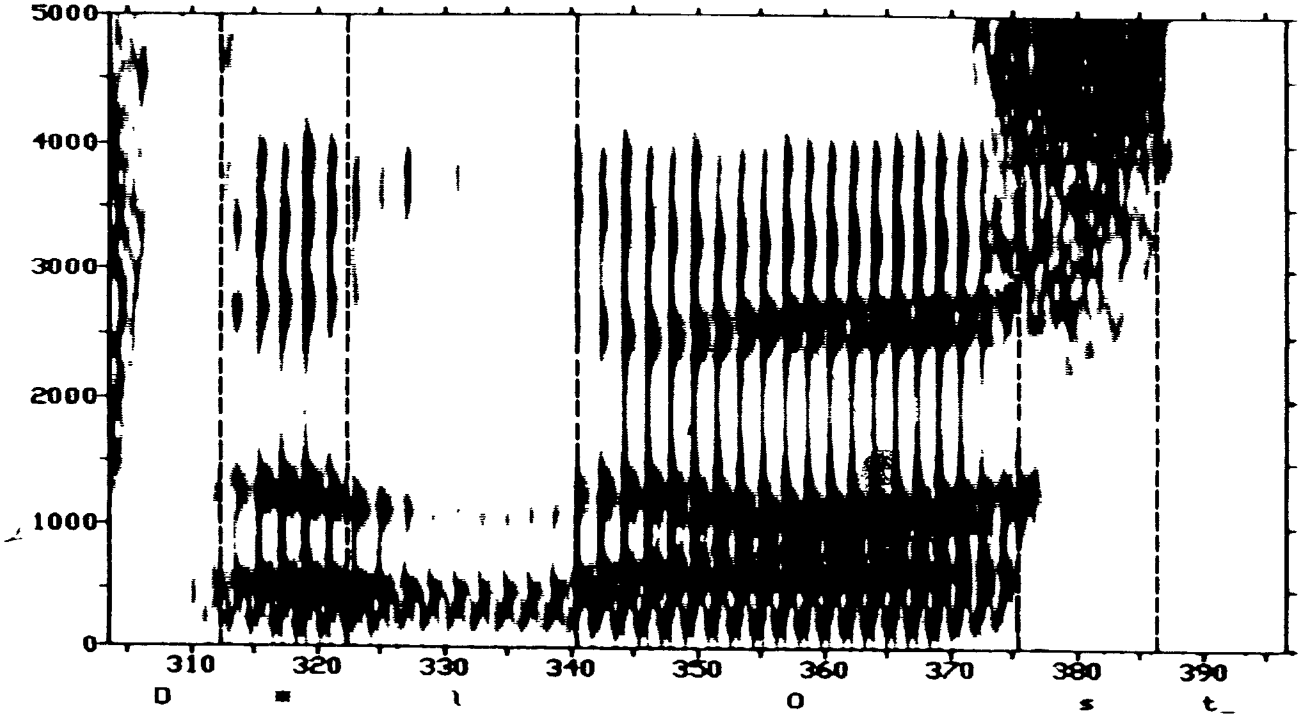


Fig. 3-8e. Synthetic speech with rule-based segmentation and labelling.

Sentence 8: (Plead with the lawyer to drop the lo)st cause.

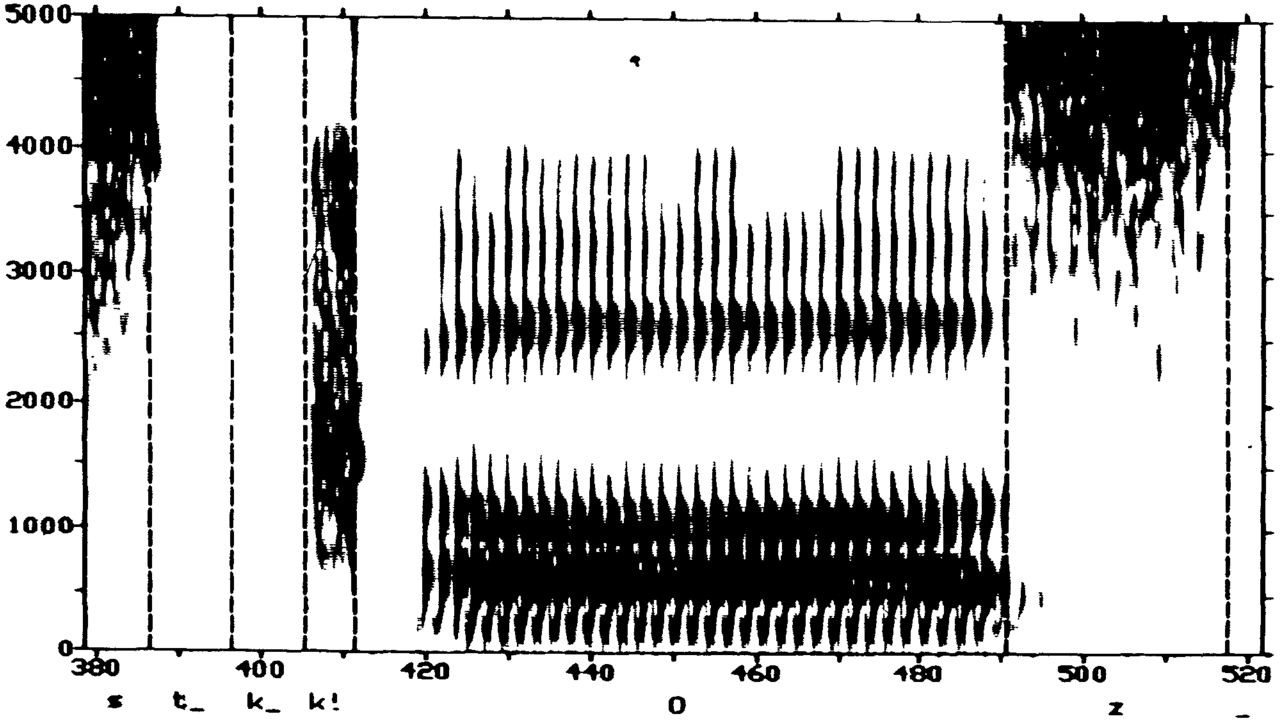


Fig. 3-9a. Synthetic speech with rule-based segmentation and labelling.

Sentence 9: Calves th(rive on tender spring grass).

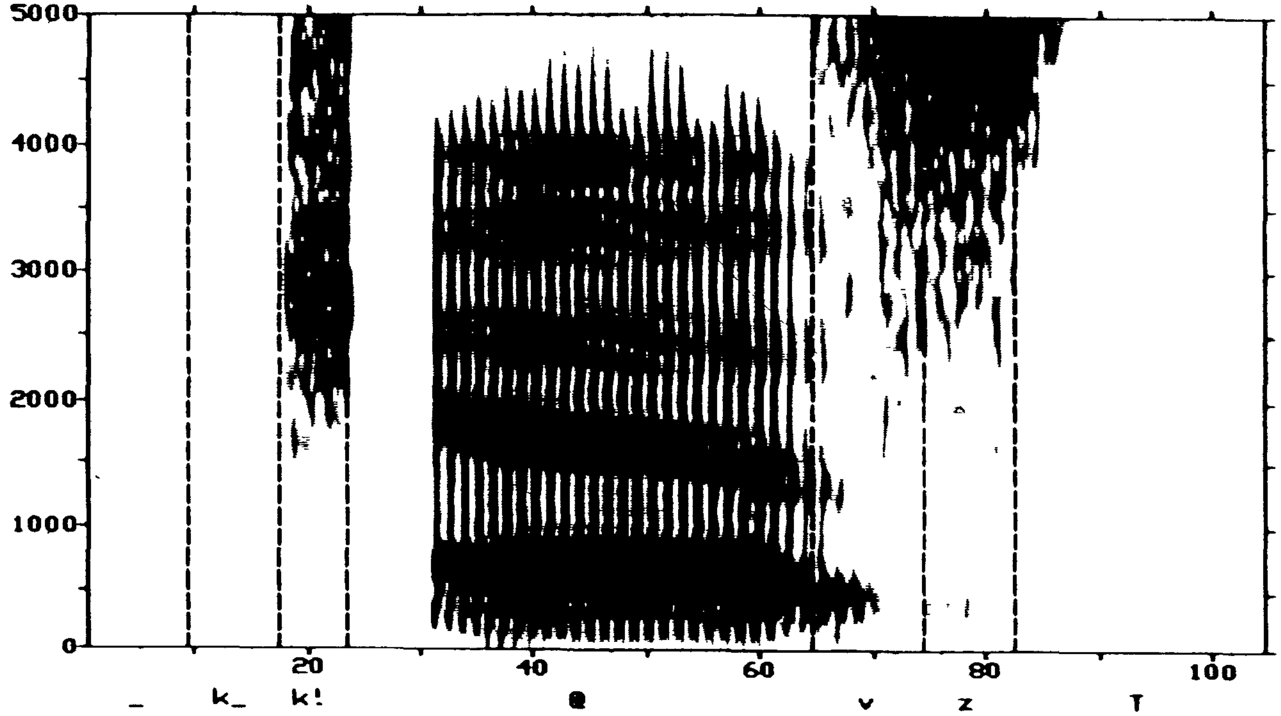


Fig. 3-9b.  Synthetic speech with rule-based segmentation and labelling.

Sentence 9:  (Calves th)rive on (tender spring grass).

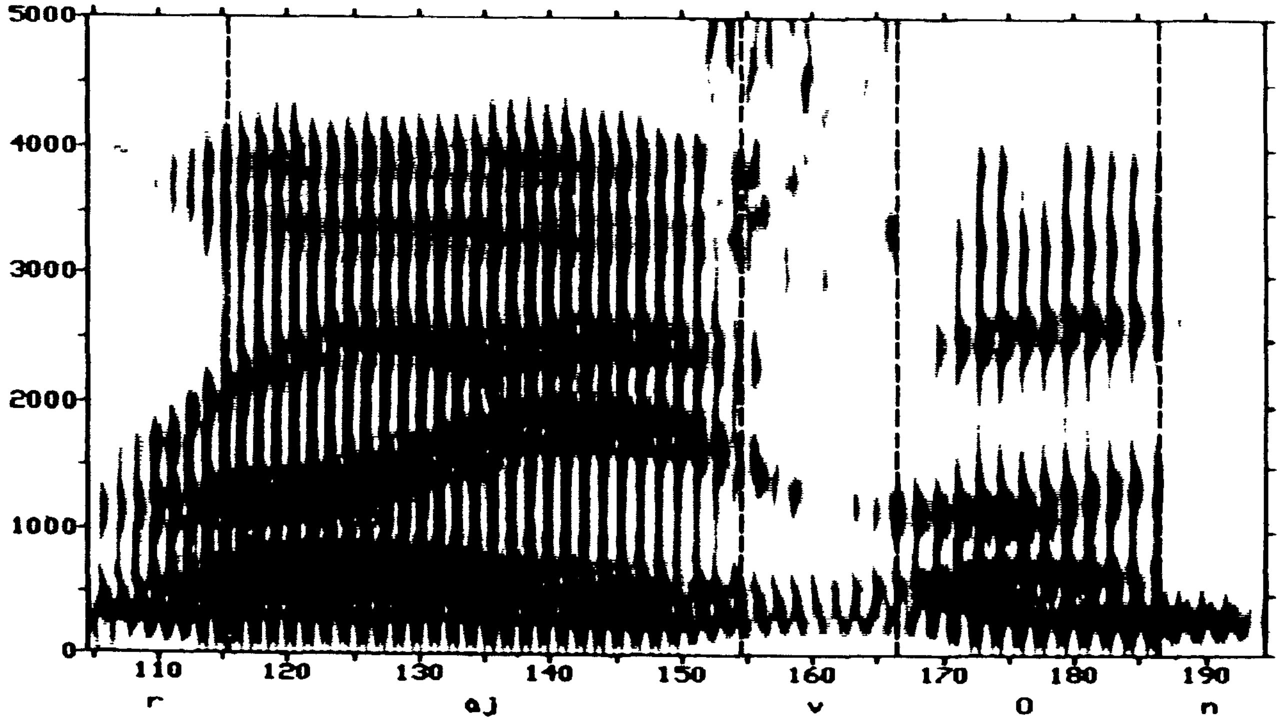


Fig. 3-9c. Synthetic speech with rule-based segmentation and labelling.

Sentence 9: (Calves thrive on) tender s(pring grass).

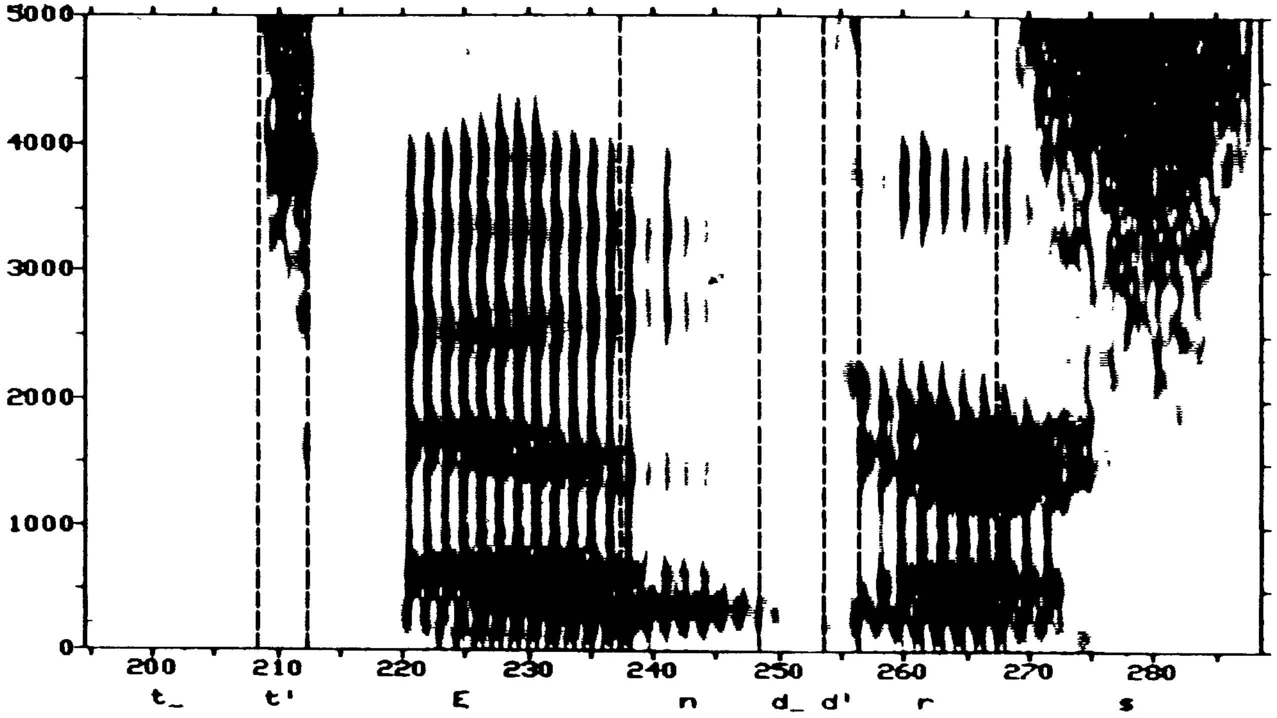


Fig. 3-9d. Synthetic speech with rule-based segmentation and labelling.

Sentence 9: (Calves thrive on tender s)pring (grass).

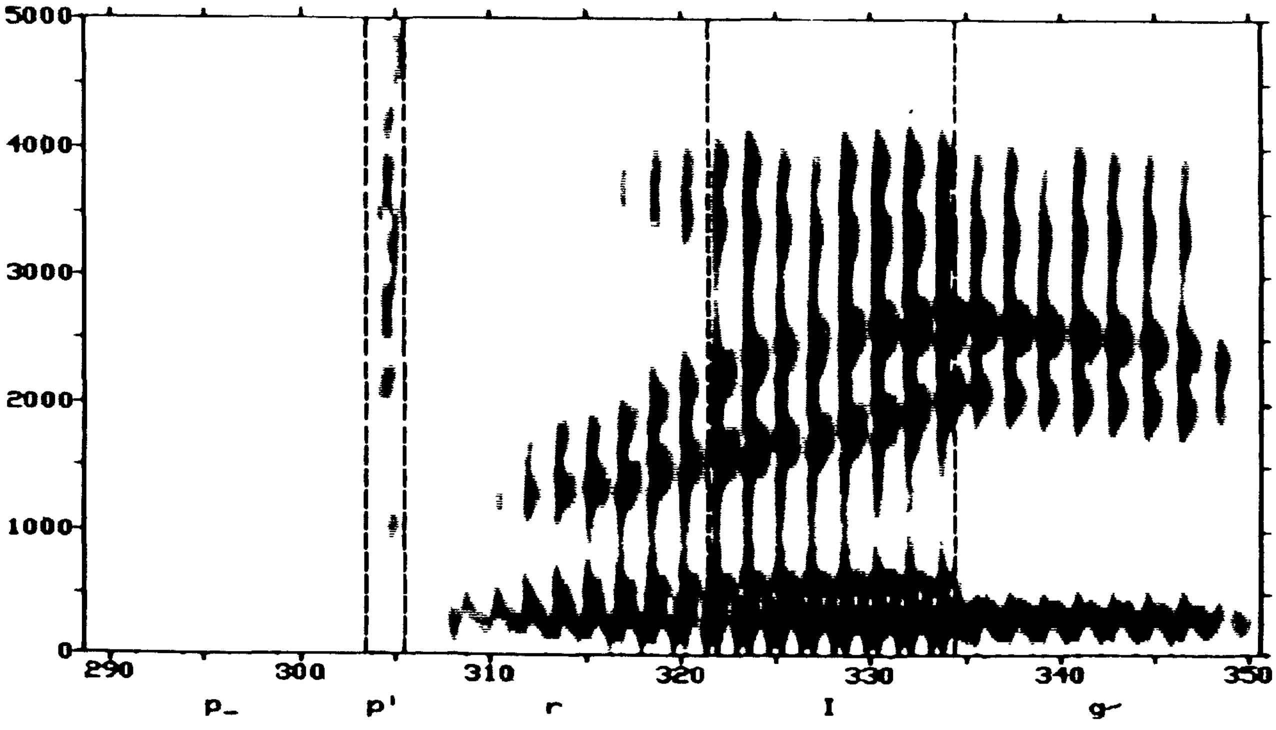


Fig. 3-9e. Synthetic speech with rule-based segmentation and labelling.

Sentence 9: (Calves thrive on tender spring) grass.

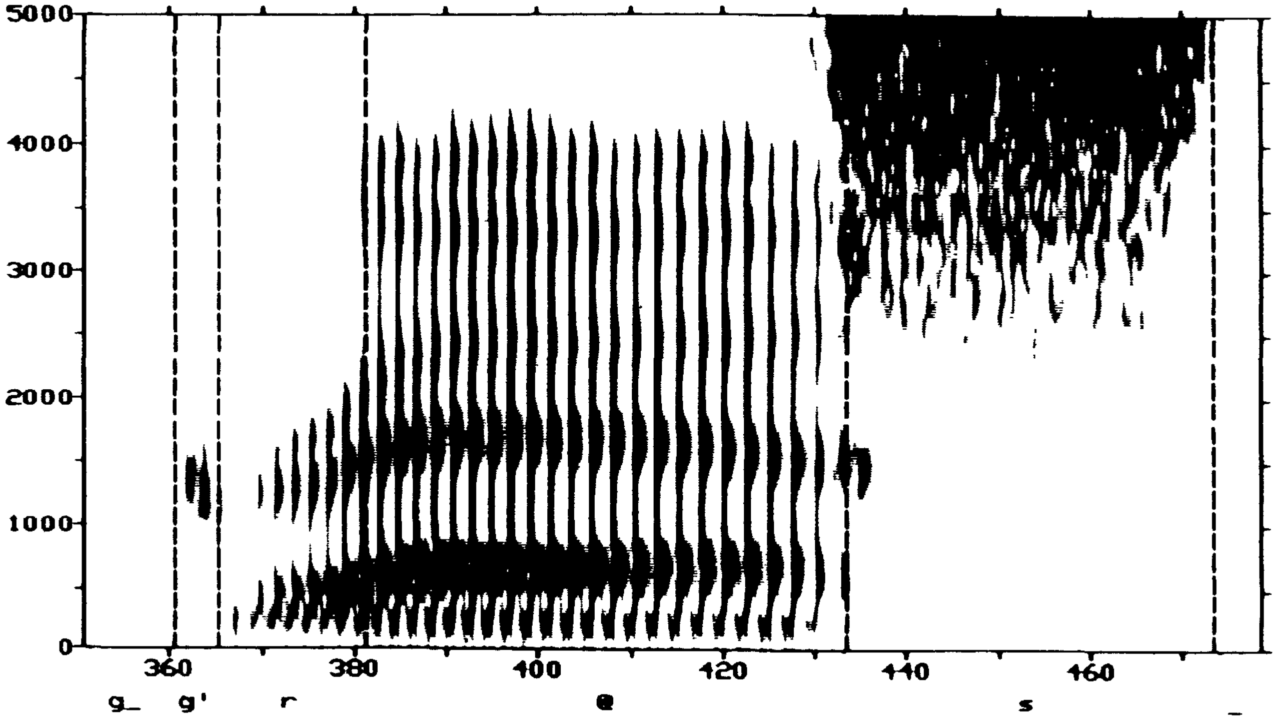


Fig. 3-10a. Synthetic speech with rule-based segmentation and labelling.

Sentence 10: Post no (bills on this office wall).

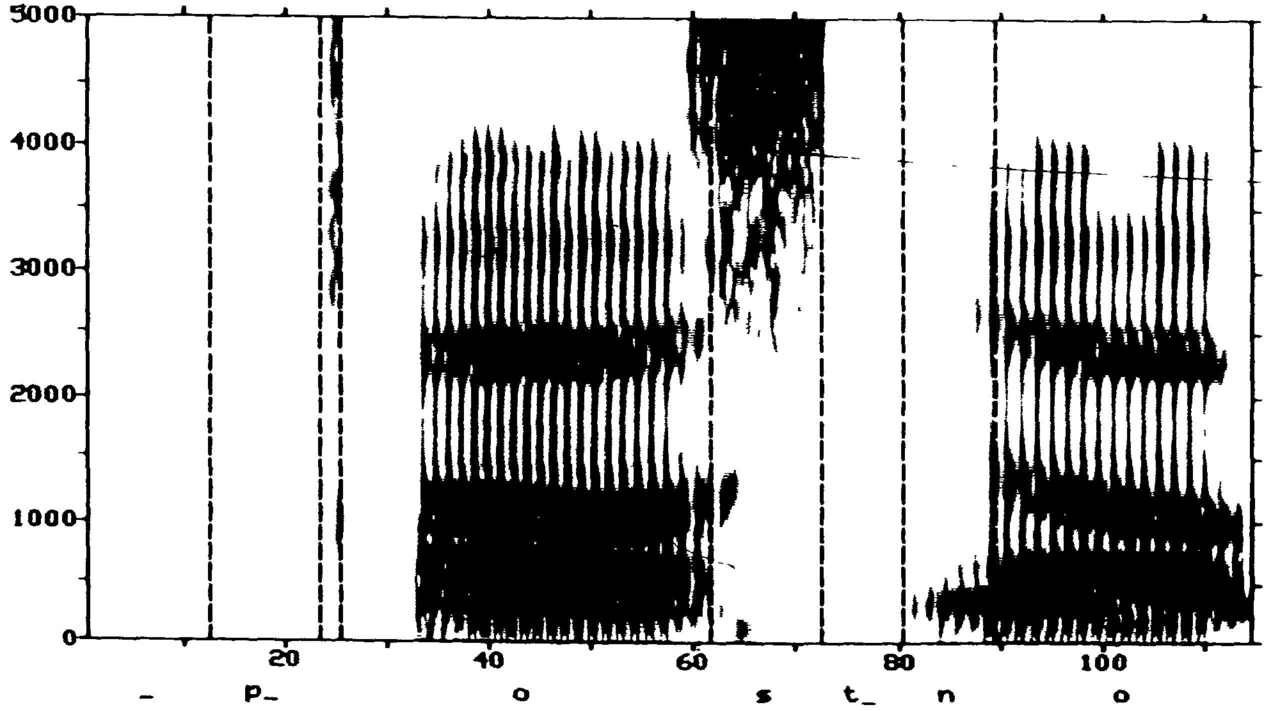


Fig. 3-10b. Synthetic speech with rule-based segmentation and labelling.

Sentence 10: (Post no) bills on (this office wall).

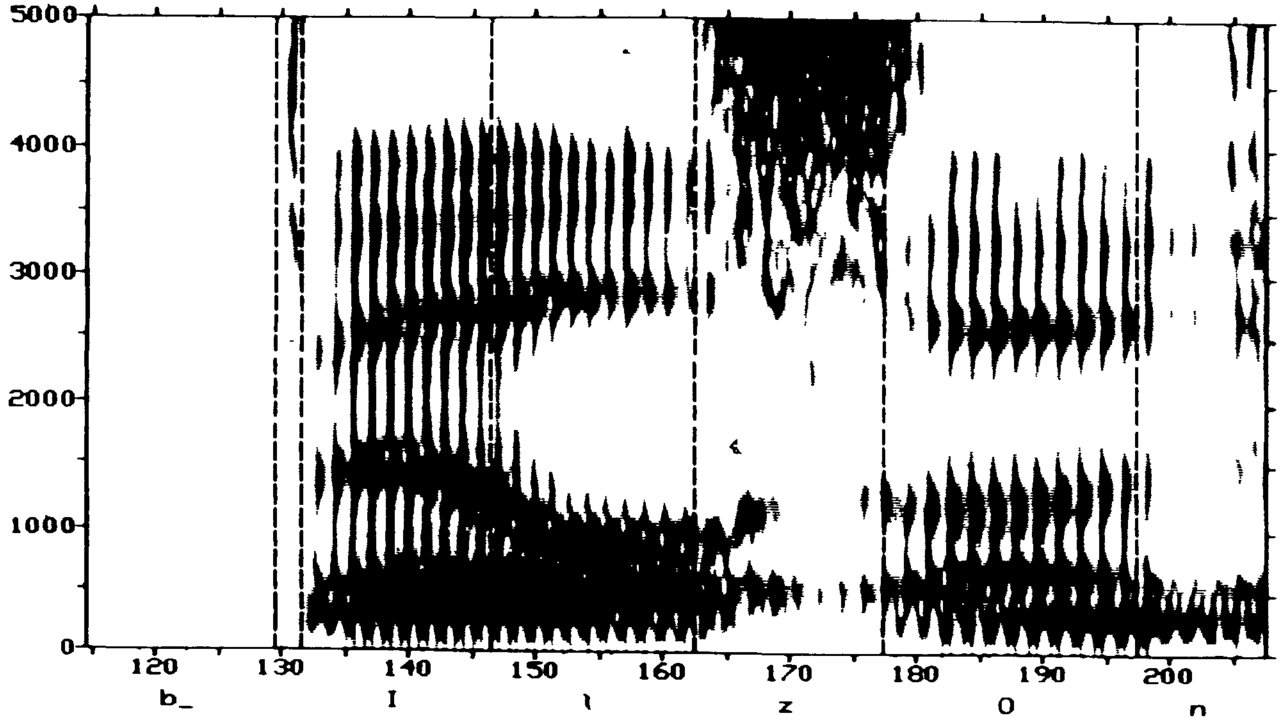


Fig. 3-10c. Synthetic speech with rule-based segmentation and labelling.

Sentence 10: (Post no bills on) this office (wall).

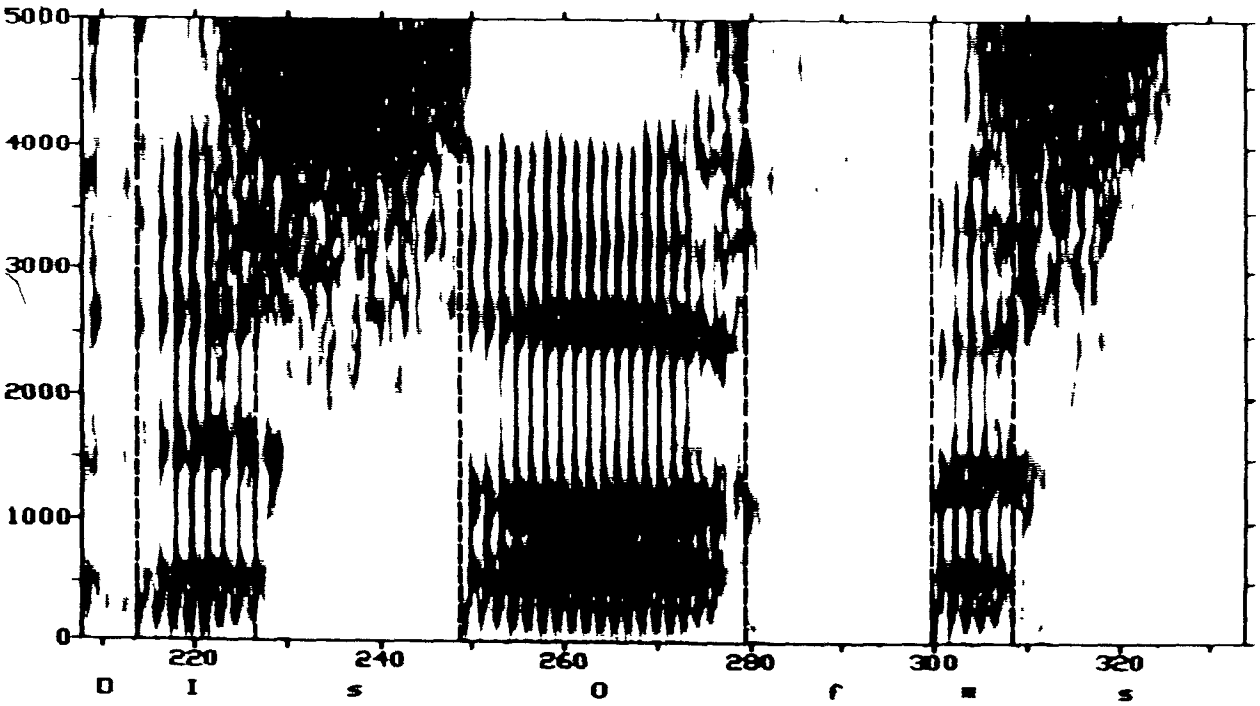
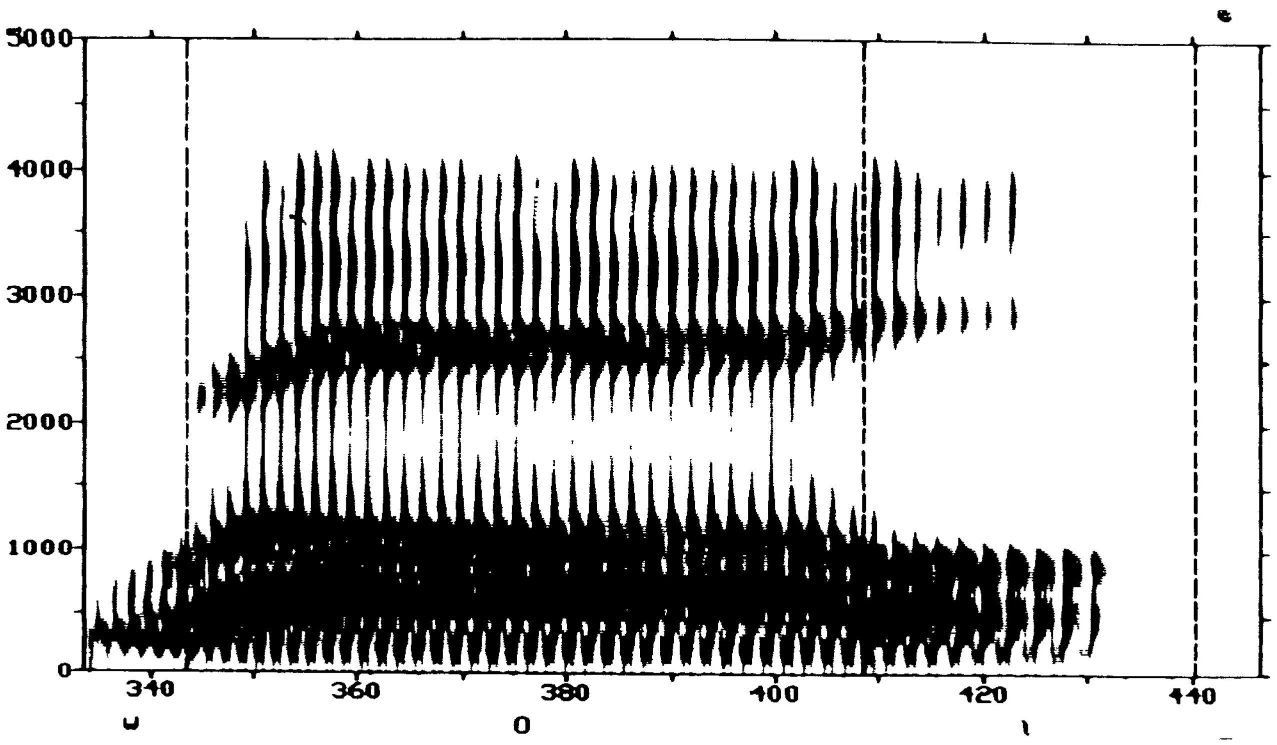


Fig. 3-10d. Synthetic speech with rule-based segmentation and labelling.

Sentence 10: (Post no bills on this office) wall.



# FIGURES FOR

# CHAPTER 4

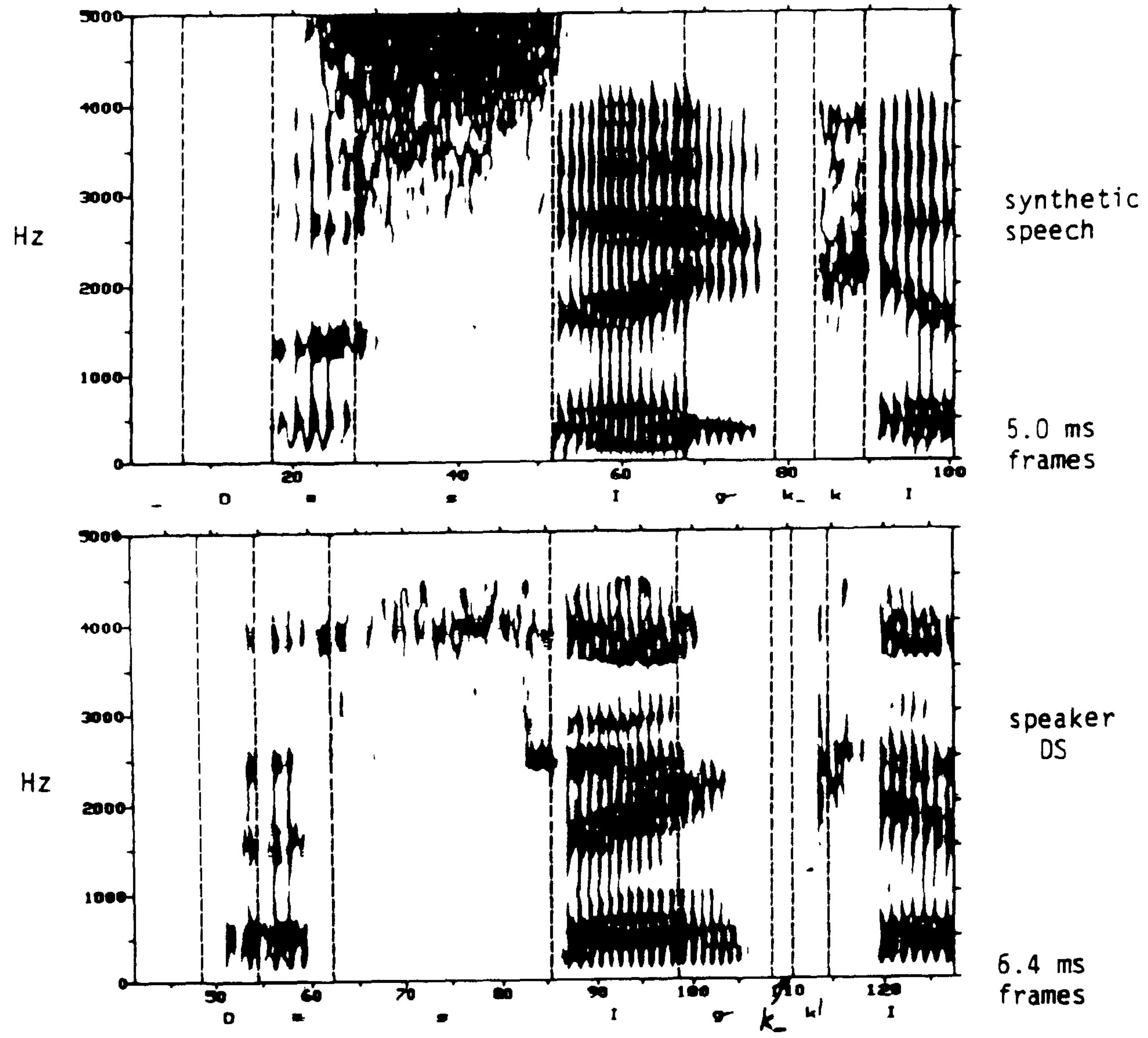


Fig. 4-1a. Aligned synthetic and natural speech (speaker DS).

Sent. 2: The sink i(s the thing in which we pile dishes).

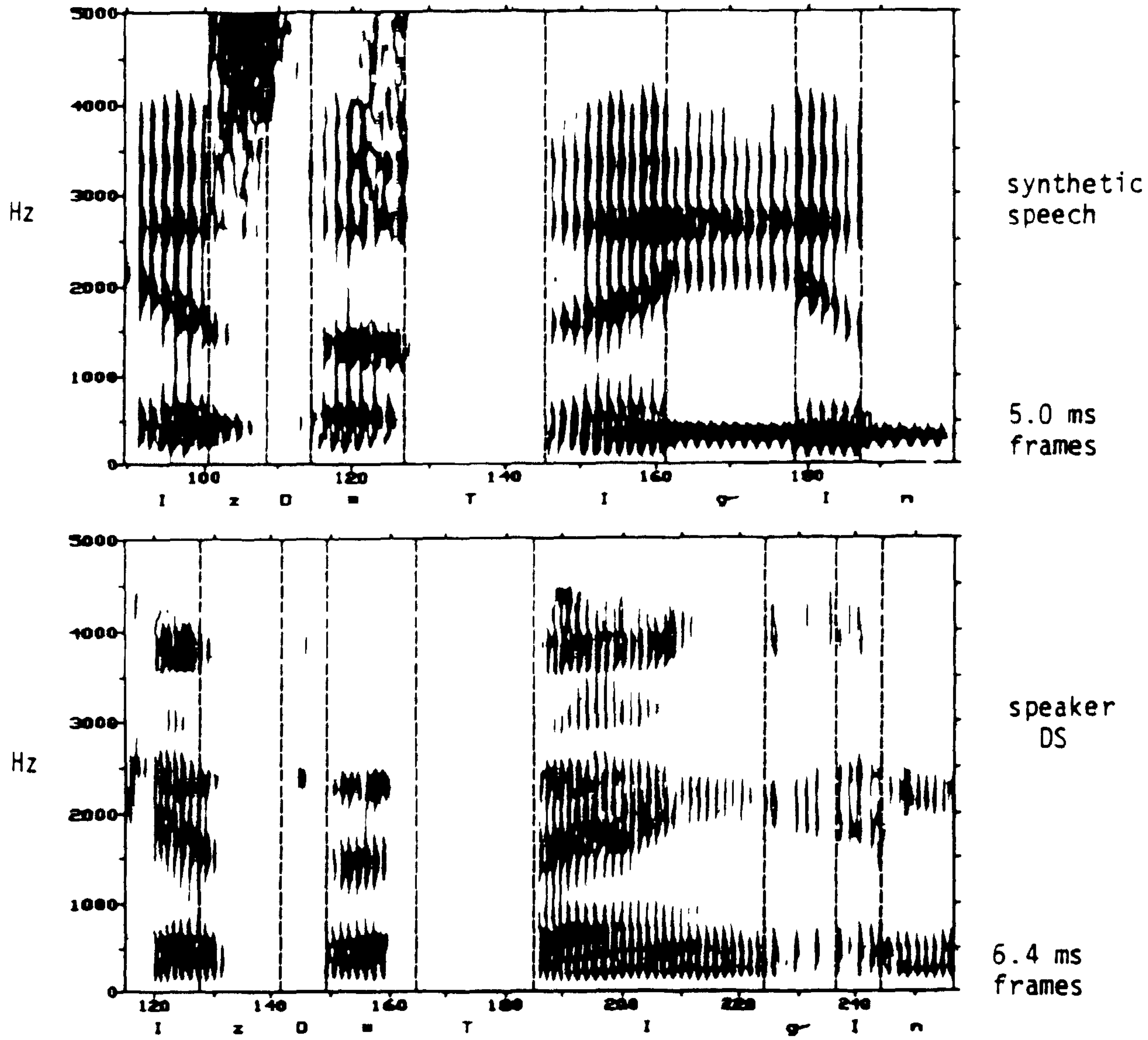


Fig. 4-1b. Aligned synthetic and natural speech (speaker DS).

Sent. 2: (The sink) is the thing in (which we pile dishes).

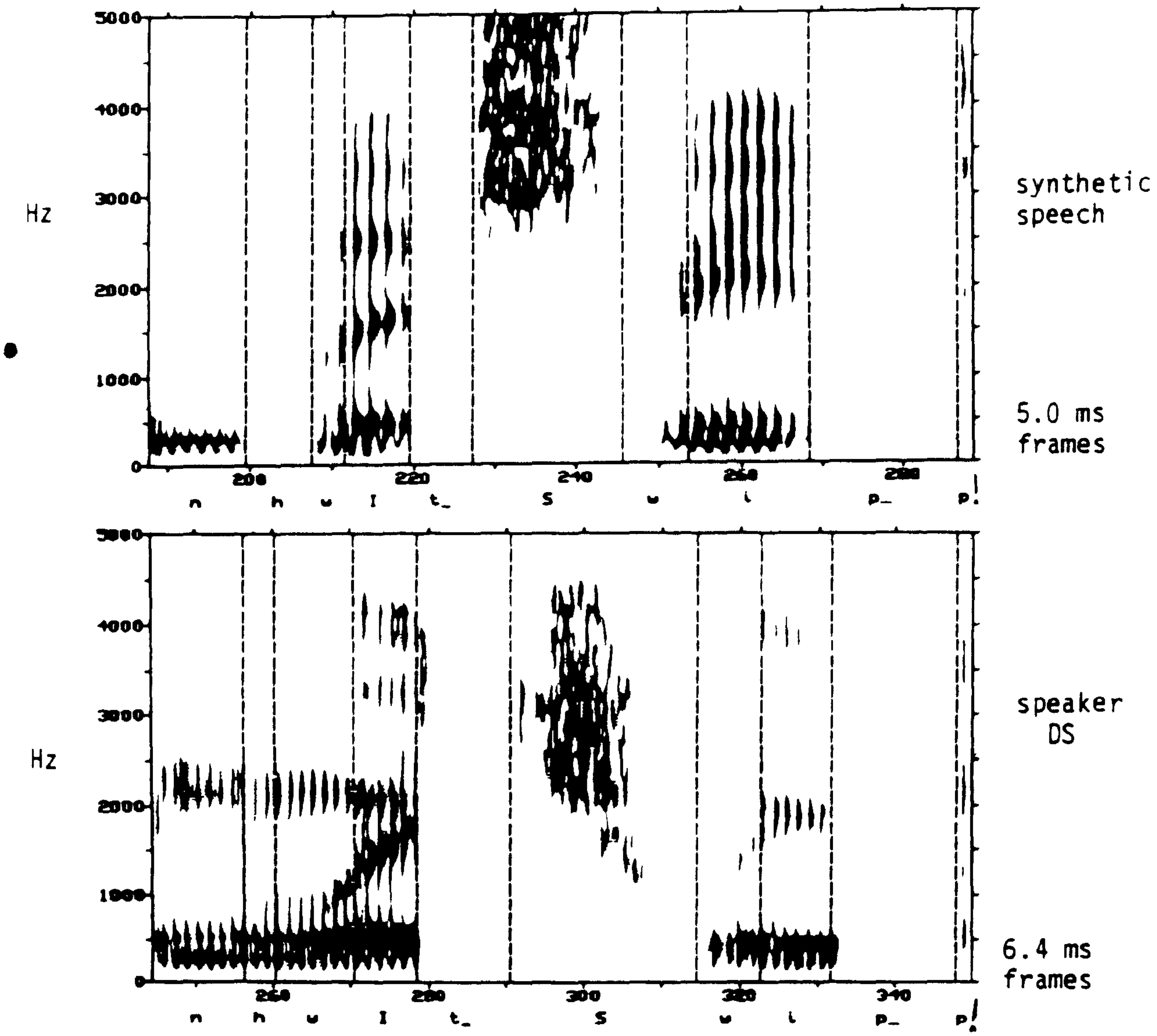


Fig. 4-1c. Aligned synthetic and natural speech (speaker DS).

Sent. 2: (The sink is the thing )in which we p(ile dishes).

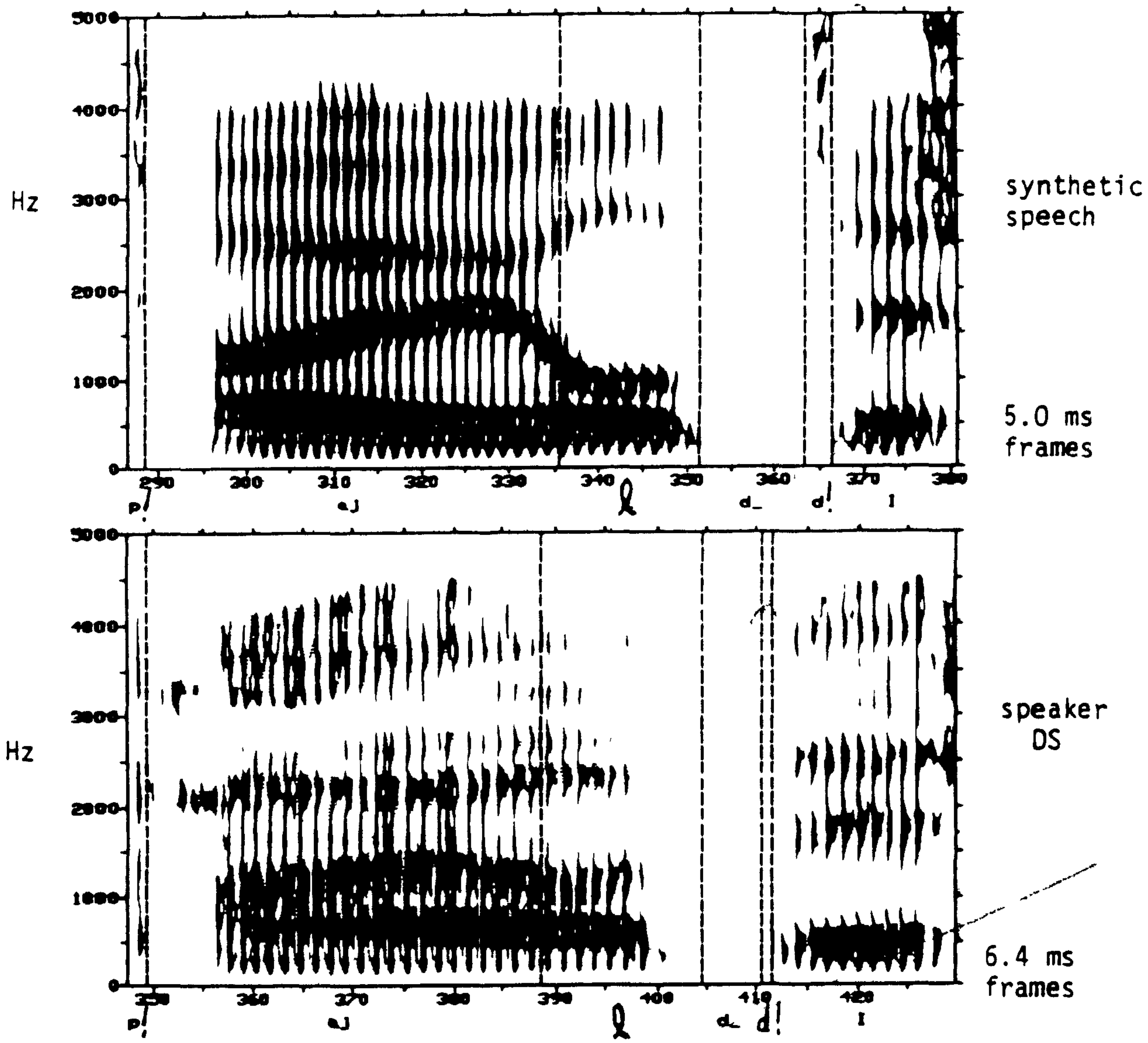


Fig. 4-1d. Alıgned synthetic and natural speech (speaker DS).

Sent. 2: (The sink is the thing in which we p)ile di(shes).

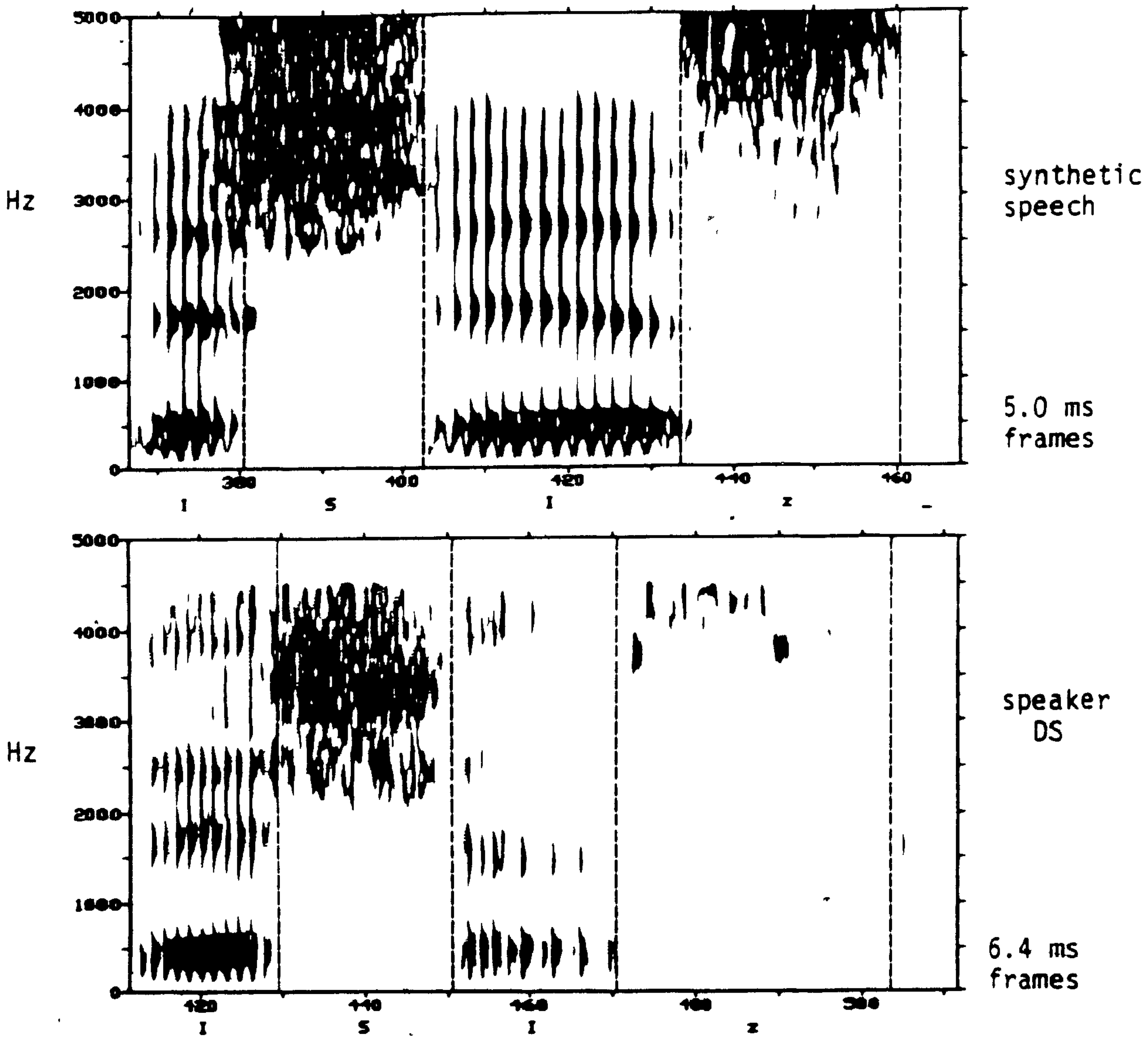
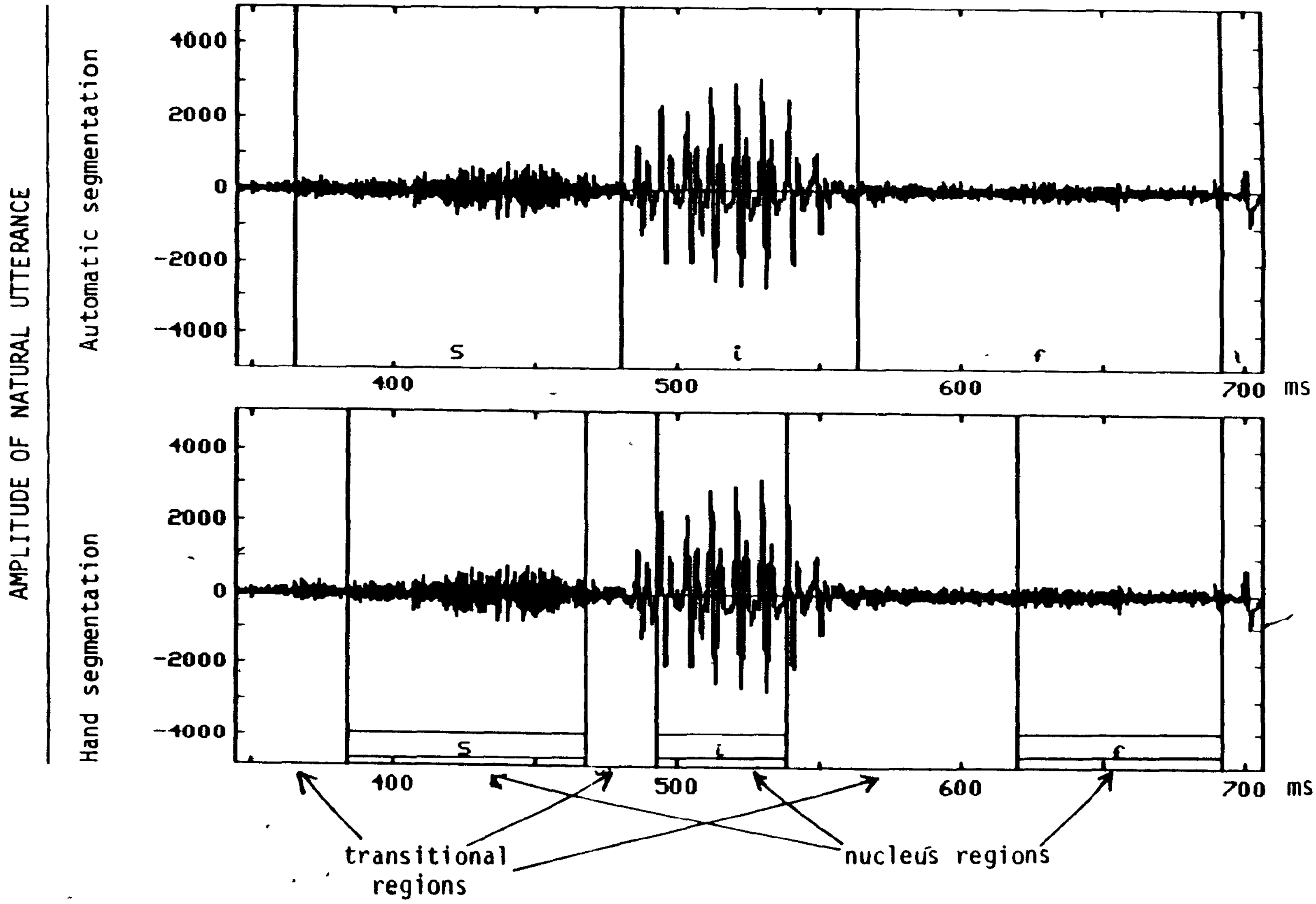


Fig. 4-1e. Aligned synthetic and natural speech (speaker DS).

Sent. 2: (The sink is the thing in which we pile d)ishes.

Fig. 4-2. Aligned waveforms of synthetic and natural speech.

Sent. 5: She fl(aps her cape as she parades the street).



FIGURES FOR

CHAPTER 5

Fig. 5-1. Speaker DS with segmentation induced from rule-based analysis of synthetic model. Sent. 8: (Plead with the) lawyer (to drop the lost cause).

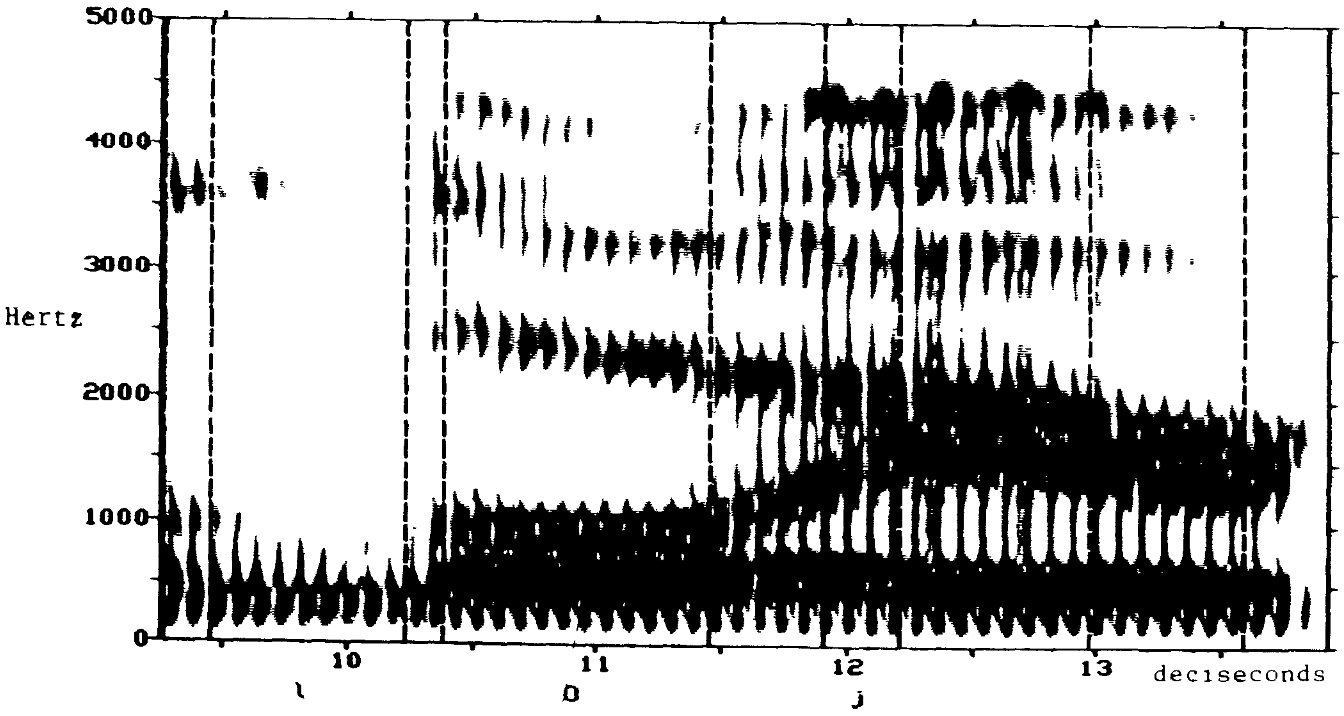
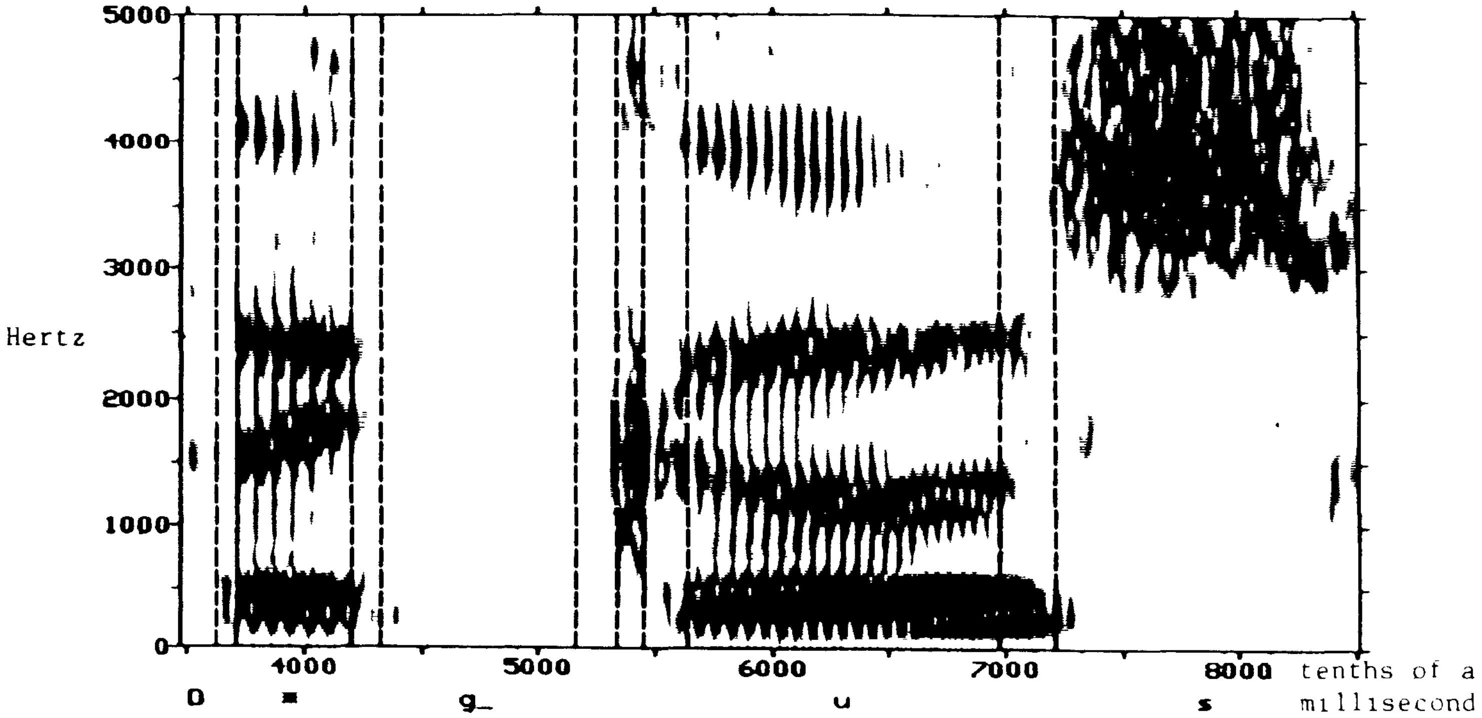


Fig. 5-2. Speaker ML with segmentation induced from rule-based analysis of synthetic model. Sent. 1: The goose (was brought straight from the old market).


Hertz
tenths of a millisecond
D
■
g_
u
s