TR-IT-0017

Description of fundamental frequency in read speech in the ATR 200 sentence database

Jaqueline Vaissiere

1993.8

This report describes the fundamental frequency characteristics as found in 1000 sentences read by five speakers from the ATR phonetically-balanced English database. It provides basic principles for the automatic segmentation of speech for prosodic analysis, and shows how meaningful chunking of sense groups is reflected in the prosodic structuring of the utterances, as indicated by their fundamental frequency contours. The work described in this report was performed during a six-week stay at ATR ITL and can be used in conjunction with duration and amplitude measures for auto-segmentation of speech by prosody.

ATR Interpreting Telecommunications Research Labs.
ATR 音声翻訳通信研究所

DESCRIPTION OF FUNDAMENTAL FREQUENCY IN READ SPEECH IN THE ATR 200 SENTENCE ENGLISH DATA BASE

J.Vaissiere

SUMMARY OF THE REPORT

The purpose is this document is to represent the information contained in the fundamental frequency (Fo) fluctuations along read sentences and to uncover inter-speakers commonalties and inter-speaker differences. The method used here if the listening and the visual observation of the large majority of 1000 sentences using Xwaves facilities (200 sentences by 5 speakers). The chosen method aims to put into light a series of inter-speakers commonalties and inter-speaker differences.

The prior knowledge which eases the interpretation of the Fo curves in any languages is summarized in the first part: the physiological aspects of Fo control, with the interplay of segmental and suprasegmental influences on the articulators, the perceptual relevancy of Fo movements, the language-dependent phonological point of view, the multiparametric character of prosody, the necessity of a superimposition model, and the factoring out of the segmental influence.

The second part exposes the basic principles underlying our description: the chunking of the continuum into successive hat pattern/sense-groups, each corresponding to an pitch excursion from the baseline, and the cues for the division of such units into smaller units or their groupings into larger units: pauses, leveling of the baseline and resetting, extension of the Fo range, continuation rise or extra low Fo values, etc.

The third part illustrates the derivation of the prosodic structuring of the sentence directly from Fo, and the fourth part exemplifies some of the observed speaker differences in Fo range, and realizations of pattern.

The next step could be

- to verify and quantify the observations automatically on the present data base (or on another),
- to complete the information derived from Fo contours by information from duration (Nick Campbell) and derive a more complete parsing of the sentences,
- to compare systematically the "prosodic tree" in English sentences in similar sentences translated in Japanese and English, and to establish the interplay between all the prosodic parameters in the three languages, which happen to be considered as very different from a topological point of view.

PART ONE: IN SEARCH OF AN IDEAL MODEL

An ideal model of Fo description probably does not exist, at least as far as I know. An ideal model should meet several requirements, most of them are exemplified in the following. Many models exist, and many of them may be judged as the only best, depending on what one want to do with the model. Specialists in speech recognition have to cope with inter- and intra-speakers differences and want to hear only about completely automatic description processes, while pure phonologists often do not even feel the need to look at grumpy acoustic data, but would like to have a deep understanding of what is going in the speaker head. We have to search for the best compromise.

1.1. PHYSIOLOGICAL ASPECTS OF FO CONTROLS

It is common sense to repeat that speech is a communication code, spoken by humans, which are all equipped with the same production apparatus and that this apparatus (the lung, the glottis, the tongue, the velum, etc.) is not designed primarily for speech (but for basic vital functions such as breathing, eating, etc.). An Fo model in a supposed-to-be physiological has to take into account the specificity of the Fo production mechanism in human. Each "active" command input of the models should correspond to a particular muscular activity.. Fundamental frequency vibrations (Fo) exactly correspond to the rate of frequency of the vocal folds. There is only three things that humans can do with the vibrations of their vocal folds and therefore with Fo:

- increasing the rate of vibrations, i.e. *raising* Fo which is equivalent to going from a lower (*Low*) to an higher value (*High*); a Fo rise can be represented by a movement (Rise), or by a change between two targets (L-H) or L-M (Low Mid), M-H, H to extra High, etc.,
- decreasing the rate of vibration, i.e. *lowering* Fo, which is equivalent to going from an higher Fo value to a lower one ((High to Low)
- or *maintaining* the same Fo.

Describing the Fo contours in one language is to **explain** the succession of Fo movements or targets in the pitch contour, their location and their relative amplitude.

Let us recall briefly the mechanism of Fo control in humans, because it is essential to have a minimum understanding about the differences between the Fo rises and falls which are due to *voluntarily* prosodic control by the speaker, from the ones which are (passive) *consequences* of the way speech is produced or of other movements.

A) Fo control at the level of the individual syllable.

There are two basic inputs for controlling the rate of vibrations of the vocal folds: the *laryngeal muscles* activities which control the state of the vocal folds (stiffness and mass of the folds, spreading of the two folds) and the *pressure drop* across the glottis[1]. As a matter of fact (unfortunate for modeling), the larynx mechanism is very complicated, and there is yet no full understanding yet.

1) Laryngeal muscles activity

There are two types of laryngeal muscles: the extrinsic muscles and the intrinsic muscles.

The larynx position influences the states of the vocal folds. Up and down of the larynx results in raising and lowering Fo, respectively. The extrinsic muscles suspend the larynx, and therefore control its height, and grossly speaking, the higher the larynx, the more stretched the vocal folds, and therefore the higher the pitch (like in untrained singers). The role of the extrinsic muscles in speech is not (yet) clearly understood, because they may control the larynx not only in a up and down movement, but in more complex movements which affect the vocal folds length .

The vocal folds states are primarily controlled by the intrinsic muscles such as the cricothyroid (CT), the vocalis (VOC) and the cricoarythenoid (LCA), at least in the chest register.

a) Raising pitch

Raising is known to be accomplished by contraction of CT assisted by VOC and LCA. When the action of one muscle may reach its limit, another muscle may be used. Only the CT activity is uniquely related to

[1]The pressure drop or transglottal pressure is the difference between the pressure above the vocal folds or Subglottal Pressure (Ps) and the pressure above the glottis or Supraglottal Pressure. A positive transglottal pressure is necessary for vocal folds vibration (Bernoulli effect).

Fo change, in particular rising movements. The CT contraction rotates the cricoid cartilage, resulting in a lengthening and stiffening of the vocal folds and thus in Fo rising. The CT seems to exhibit a peak of various amplitude for every syllable located above the lowest speaker range (or baseline). Such pattern suggests that there is an individual command for **each** syllable for which Fo is higher than the baseline, and not a single command encompassing several syllables. When two or more syllables are spoken in the higher Fo range, it should be understood then it is not the same continuous input encompassing the series of syllables (a unique tension of the CT), followed by a relaxation of the CT activity, but rather a series of activities of the CT muscle, which is, let say, synchronized with syllables.

Figure I.1 (extracted from Maeda's thesis) and Figure I.2 (from Samada and Hirose , 1971) illustrate the correlation between CT activity and Fo raising for both English and Japanese. The correspondence between the CT activity and the Fo curve (no delay), the non linear relations between the amplitude of the CT activity and the amplitude of the Fo rise and the successive CT activity for each syllable above the baseline can be observed in both figures.

Figure I.1: CT activity in English sentence
Fo contour in English and the corresponding EMG activities of the laryngeal muscles, CT, VOC, LCA, SH and ST for one sentence uttered by KN " Almost all farmers raise yellow sheep" (extracted from Maeda's thesis.

Figure I.2: CT activity, Japanese sentences
Fo contour in Japanese and the corresponding EMG activities of the cricothyroid (top) in two Japanese phrases, with different accent patterns (from Simada and Hirose, 1971).

b) Lowering pitch

The mechanism for Fo local *lowering* is less clear than for Fo rising. There may be passive and/or active lowering. Relaxation of the CT causes lowering (passive lowering). Fujisaki speculates only passive lowering in Japanese, by relaxation of the CT activity, but such an assumption may not be in accordance with the data obtained for different Japanese speakers. (see one hypothesis by Hirai and Honda, 1993). When the amplitude of lowering is larger than the amplitude of the preceding rise or when there is an extra low Fo value, for example at the end of certain sentences, or inside sentences (as a boundary marker), active lowering has to be assumed. From inspection of Fo curves produced by different French speakers, I have argue that there may relatively large speakers differences

in ways of chunking continuous speech: some speakers chunk the flow of speech merely by rapid rises (P4 speakers), or by a succession a relatively similar rises and falls (P3 speakers), or by extra peak on particular syllables (P2 speakers) (Vaissiere, 1975). Unexpectively the same speaker using different ways of chunking have been observed on the present data base (as seen later).

As shown in figures 1 and 2, the SH and the ST curves also exhibit peaks and seems to be related to the Fo lowering. Such activity is increased in final Fo lowering (see both figures), and in syllables lower than the "normal" baseline of the speaker.

A model should assume a passive lowering by relaxation of CT activity and the possibility of an active lowering by the complex external laryngeal mechanism.

B) Fo control along long stretches: the declination tendency

We have seen the mechanism for local rises and falls, encompassing one or two syllables, at most. These locals movements are superimposed with larger movements, which encompass long stretches or speech, up to a whole sentence.

A general slow decline of the Fo values along (declarative) long sentences is generally observed. Such a decline has been observed in the data of many different languages, and should therefore correspond to a "natural" tendency, a passive consequence of the way humans speak, Figure I.3 illustrates the lowest Fo values along the first sentences for one of the speaker (sab). As it can be seen in that picture, the tendency to fall can be observed in all sentences, but it is not possible to give a fixed rate of decline.

In contrast with professional speakers, that speaker has the tendency to translate its Fo bottom toward lower value, from one sentence to the next. The change may be related to the volume of air she intakes between each utterance. Larger intakes should lead to higher Fo value (increased larynx height and increases subglottal pressure). The measurement of the tension of elastic rubber band around the lung of the speakers could be set up to estimate the correspondence between the volume of air intake and such the amplitude of decline.

Figure I.3 Declination tendency for one speaker (sab)
Lines joining the local minima Fo values in the first sentences in the list

Such a tendency can be however controlled, and the slope may of declination may vary along the sentence, or even be suppressed. There are arguments among researchers concerning the presence or absence of the Fo declination. But it is a fact that the Fo declination often occurs, at least in declarative sentences, which corresponds at least partially to a **passive** lowering and any Fo description model, therefore, should include the description variable that account for the declination. Local downstepping[2] and such a general tendency should be well separated as two separate inputs, although if it is not always easy to separate the two from the inspection of single sentence.

The speakers tend also to start and to stop voicing with the same Fo values, visible only when the first phoneme is voiced, and this natural tendency conflicts with the tendency of a linear decline. Lower Fo values should be expected at the end of longer sentences, which is not the case. For example, in the 20 first sentences by speaker gsw, Fo values tend to start around 130-140 Hz (maximum 147 Hz), and to stop between 95 Hz-110 Hz (statistics could be systematically run for the 1000 sentences).

The Fo values tend to decline more rapidly in the sentence beginning (at least when a linear scale is used), and then the declination rate is reduced (leveling tendency) or there may be even a reset at a major boundary in the sentence . All the Fo contours in this report illustrate this tendency.

The *declination tendency* probably does not correspond to a single mechanism in all cases, but to both a passive and an active control. In the following, we shall describe some of the possible physiological correlates of Fo declination, which may explain passive lowering.

1) Subglottal air pressure P(s)

The *subglottal pressure* falls along some sentences might account at least partially for the declination tendency. However, Ps is not clearly falling along all sentences, and the speakers tend (probably by making an effort) to maintain Ps constant. As far as I know, there is no extensive data (on a large data base) to establish the correspondence between Ps and the declination tendency (such an experiment does not represent technical problems, and will be useful).

---

[2]Down stepping is marked by ! in the TOBI representation

2) Larynx height

At rest position, the larynx is lower than during phonation.

The higher the larynx the higher the Fo frequency.

The larynx is known to *rise and then fall* during short sentences (for Japanese: Kakita and Hiji, 1974; English: Vanderslice, 1967). The same large larynx movement can be clearly seen in long sentences. The initial larynx rising movement tend to end after phonation starts, in the sentences, and may be responsible for the sentence initial rising tendency (ceteris paribus, the sentence initial syllables tend to have a rising Fo contour, whatever their status) and for the higher Fo values at sentence initial portion. The higher Fo values during the first syllables of the sentences (generally during the first lexical word), and the Fo maximum during the first two or four syllables) seems to be a rather language-independent tendency.

Declining larynx as seen in some sentences height may contribute the declination tendency, but the larynx does not decline in sentences in a consistent manner.

Figure I.4 illustrates the movements of the mandible (JAW), the hyoid bone (HYD) and the thyroid cartilage (TYD), and the corresponding Fo contours (FO) and amplitude envelope (AE) for two English sentences spoken by KN: " My dog like the enormous gorilla" and " Ken raises the light yellow sheep." (extracted from Maeda's thesis). As can be seen, the larynx starts to rise before the uttering of the sentence. to such a rise probably corresponds **the so-called speech ready gesture**, which corresponds to a hypothesized tensening of many muscles (all?) involved in speech production, (such as the velum at the onset of English sentences; see Vaissiere, 1989, among others), and which has its **maximum at the first stressed syllable** and which falls toward the end of the sentence, generally with the last stressed syllable in English, or before for some speakers (as seen by a lowered velum). In the first sentence, shown in Figure I.4, the movement of the larynx seems to be correlated with the declination tendency, and even with the two successive "hat patterns"[3], but there is no declining larynx height in the second sentence, while Fo is declining. The main Fo dip in the sentences may (on the top) or may not (on the bottom) correspond to a dip in the larynx height.

Figure I. 4: Vertical movements of the larynx in English sentences

[3] The notion of "hat pattern" will be seen in detail later

Movements of the mandible (marked JAW), the hyoid bone (HYD) and the thyroid cartilage (TYD), and the corresponding Fo contours (FO) and amplitude envelope (AE) for two English sentences spoken by KN: " My dog like the enormous gorilla" and " Ken raises the light yellow sheep." (extracted from Maeda's thesis).

3) Shortening of the ventricles

As seen before, the temporal course of Ps and of the larynx height are sufficient to account for the general declination tendency.

A *gradual shortening of the ventricles* from beginning toward the end of sentences, which may be explained by a tracheal pull due a decreasing lung volume, has been postulated (Maeda, 1975). As a general principle, the shorter the ventricles, the lower the fundamental frequency. The rate of change in Fo with respect to vocal-fold length may vary from 7Hz/mm to 20 Hz /mm depending in the individual subjects (Hollien, Brown and Hollien 1971). Such a gradual shortening has been attested in all English sentences studied by Maeda, but there is in fact only very few sentences.

Figure I.5 represents the fluctuations of the laryngeal ventricle length and the corresponding Fo contours for two sentences, read by KN: " I like the cat in the tree in the park." and " almost all farmers raise yellow sheep." The magnitude of the baseline is equal to 32 Hz. The straight line superimposed on the data representing the frame-to-frame variation in the ventricle length is determined by a least square fitting algorithm. The concomitant decline in Fo frequencies and the shortening in ventricles length can be observed on that figure.

Figure I.5: Laryngeal ventricle length and declination tendency.
Fluctuations of the laryngeal ventricle length and the corresponding Fo contours for two sentences, read by KN: " I like the cat in the tree in the park." and " almost all farmers raise yellow sheep." The magnitude of the baseline is equal to 32 Hz. The straight line superimposed on the data representing the frame-to-frame variation in the ventricle length is determined by a least square fitting algorithm (extracted from Maeda's thesis).

C) Controlled and uncontrolled Fo fluctuations

The suprasegmental fluctuations may correspond to *local* fluctuations involving one phoneme or one or two syllables (local rise or jump and fall between two syllables) or one vowel (gliding tone). They may correspond to more *global* fluctuations (such as the baseline, and the resetting of the baseline, involving long stretches of speech). Part of the fluctuations, local and global, is the uncontrolled consequence of other movements (such as the *microfluctuations*, as consequences of the

production of stop consonants and part of *declining tendency* due to declining air volume in the lung). Small fluctuations may be under the controlled of the speaker (such as a gliding tone on a vowel in a tone language, or the expression of the continuation rise[4] in a number of languages, such as German, French, Spanish, etc.). Part of the declination tendency (the slope, the location of the resetting and/or the leveling) is also under the full control of the speaker. The factoring out of uncontrolled Fo fluctuations cause a real problem in modeling the part of Fo fluctuations due to suprasegmental control, because they are intimately interlaced. It is very important to differentiate between a fluctuation (a local rise or fall on a vowel, or a larger movement, stretching along the whole sentence or large part of it) that correspond to a precise command (let say an effort), to fluctuations that are uncontrolled consequences of changes in subglottal pressure or states of the vocal folds and in tongue position..

Both laryngeal muscles activity and Ps variations may be actively controlled by the speaker for prosodic purpose: increasing CT activity leads directly to an increase in Fo value and increased Ps leads also to an increase in Fo value, for example ad may be used to assist CT activity, such as in focusing). Vocal folds length and tension, and Ps may vary under influences other than prosody: higher cricothyroid activity for unvoiced consonants increases Fo values at the release of the consonants, and declining Ps due to a declining volume of air in the lungs along the sentence leads to declining Fo values. What may things rather complicate to interpret is that uncontrolled fluctuations may be heard by the listeners, and contribute to the prosodic patterning of the sentence. For example, word initial unstressed syllables beginning by a unvoiced stop (such as the syllable "com" in the word "computer" which is stressed on the second syllable) may be heard as more stressed that unvoiced counterparts .

The local Fo contours on a vowel, surrounding by two unvoiced stops may be falling because of the *cointrinsic*[5] *segmental Fo fluctuations*,, and the same vowel in an equivalent phrase, surrounded by two sonorant consonants may have a *flat* Fo contour. The falling/flat paradigmatic

---

[4]Notated H% in the TOBI system.

[5] The difference in Fo height of the vowel in [pi] and [pa] is said to be due to **intrinsic** difference between the two vowels [i] and [a], while the difference in Fo height of the vowel in [pa] and [ba] is said to be due the **cointrinsic** influence of the consonants [p] and [b].

contrast in that case should be interpreted as the "natural" consequence of particular laryngeal and pressure conditions for the production of the surrounding consonants: it is not the results of a choice. The cointrinsic effect are particularly important for English, and therefore will be detailed.

D) Fo microfluctuations due to segments...

1) Perturbations due to the presence of voiced and unvoiced stops and fricatives

The vocal folds vibrates during production of the voiced sounds (the vowels, and most of the consonants) and stop to vibrate during the production of the unvoiced sounds. The fact that the vocal cords are vibrating or not (voicing contrast), or the timing between the release of the consonants and the onset of vibrations is used as distinctive feature in most of the languages; the rate of vibrations during the consonants is not used as a distinctive feature.

The languages differ widely in the use of the larynx for differentiating consonants. Japanese and French use mainly a voicing contrast: there are vibrations of the vocal folds (and therefore detection of fundamental frequency) during the voiced stops, as opposed to the lack of vibration during the unvoiced stops. In English, the presence versus absence of vocal folds vibrations is not the primary characteristics that differentiates between [b], [d], and [g], on one side and [p], [t] and [k] on the other side. There is often, as illustrated late, no Fo detection (no voice bar) corresponding to the production of [b], [d], and [g]. The difference is merely a difference of tension and there is in English more segmental effect due to consonants than in French and Japanese. In that respect, English is more difficult to model.

For what concerns microfluctuations of the Fo contours, the consonant can be divided into four different types, as illustrated in figure I.6: *unvoiced stops,*, voiced stops and fricatives and sonorant consonants

Figure I.6: typical Fo contours for the four consonant types

Typical Fo contours corresponding to the four types of consonants: the unvoiced stops, the voiced stops, the voiced fricatives and the sonorants.

a) Intrinsic perturbations

a) Unvoiced stops

The production of unvoiced consonants involves a maneuver for stopping voicing. That is done generally by tensening the vocal folds and widening the glottis. Such maneuver has a consequence on the Fo contours of the surrounding vowels.

The tensening and widening of the vocal folds are mainly responsible for the delay (Voice Onset Time or VOT) in starting vibration after the release of the unvoiced stops consonant and the higher Fo values at the vowel onset after the consonants in English.

Figure I.7 illustrates the voiced/voiceless contrast in an initial consonant cluster of a stressed syllable upon the Fo contour and the EMG activities in CT, VOC and SH. The continuous curves correspond to words with voiceless stops, and the dashed curves represent the words with voiced stops. As it can be seen in that figure, CT level of activity and timing is different for voiced and unvoiced consonant. shows an earlier and increased activity of the CT during the unvoiced stops as compared to the voiced stops. The difference in Fo at the onset of the vowel between /p/ and /b/ is as high as 20 Hz. The vocal folds may tense and lengthen for the Fo rise already during the consonant, since there is no incompatibility between the prosodic gesture and the gesture for the consonant.

Figure I.7: CT activity during voiced and unvoiced stops
Cricothyroid (CT), the Sternohoid (SH) and the vocalis(VOC) activity during voiced (dotted line) and unvoiced consonants (plain line), for one American English speaker (KN).

b) Voiced stops

The presence of voiced stops perturbs also the Fo contours.

As well known, Fo is dependent on transglottal pressure: sudden drop in transglottal pressure due to the oral constriction leads to a sudden drop in Fo contour. A supraglottal cavity sudden complete occlusion (such as for the stops) or partial occlusion (such as for the fricatives) leads to a sudden increase in supraglottic pressure. Ps being fairly constant along the sentence, the transglottal pressure decreases, leading to a drop in Fo, and Fo voicing become more and more difficult to maintain with time. If the occlusion is very long, the vibrations of the vocal folds may even stop. The place of occlusion has a small effect on maintaining voicing: the smaller the cavity between the closure and the glottis, the larger the effect.

It is less difficult to maintain voicing during the labial /b/ than during the velar /g/. Statistics could be done on the present data base to verify such a tendency.

There are two typical maneuvers for maintaining voicing: increasing the volume of the cavity between the glottis and the constriction and changing the state of the vocal folds.

- In the case of the labial /b/ (closure at the lips), the supraglottal volume is large, and it is possible to extend the mouth volume (cheeks), to maintain transglottal flow positive. In the case of velars, for which the constriction is more black in the mouth, it is more difficult to extend the (laryngeal and pharyngeal) cavity. This may explain while voiced stop /g/ tend to be lacking in a number of languages where the voiceless counterpart /k/ exists. In a number of languages also, voiced stop /g/ may often be devoiced, or even **nasalized**: the opening of the velopharyngeal port allow to maintain a transglottal air flow (as it seems to be the case in Dutch and Japanese: n instead of [g])). Maintaining voicing during voiced stops can also be done by **lowering the larynx** (Ladefoged, 1968). Another possibility is to have a contrast not based on the presence of voicing, but merely on VOT (voice onset time).

- Stevens suggested also a **slackening of the vocal folds** for maintaining voicing. This slackening, it exists will contribute to lower Fo values even more.

Figure I.8 schematizes the language tendencies and Figure I.9 schematizes the effect of the place of articulation. A systematic verification of the influence of the place of articulation on the English and the Japanese data bases available at ATR could be used to analyze statistically language-dependent tendency for each particular speaker (degree of voicing in the stops). Figure I.10 illustrates the different maneuvers for maintaining voicing: slackening of the vocal folds themselves, lowering of the larynx, enlargement of supraglottal cavity, eventually the opening of the velo-pharyngeal port (there may be other maneuvers as well).

Figure: 8 VOICED-UNVOICED CONTRAST AND TENSE-LAX CONTRAST

Figure: 9 VOICING: The effect of the place of articulation

Figure: 10 Maneuvers for maintaining voicing in voiced stops

In languages like English, the voiced consonants tend to be devoiced. Figure I.11 illustrates the Fo contours during the sentence initial portion "Jane adores...) for three speakers. as it can be seen in this figure, the first speaker voices the two consonants /d / and /d/, and the two other speaker devoices /d /. In the case of the second speaker, [d] is partially devoiced and it is almost completely devoiced for the third speaker. The typical micromelodic perturbations for the production of an voiced stop can be seen for the first speaker. Fo is typically flat or rising after a voiced stop (while it is typically falling in the case of an unvoiced stop).

Figure I.11 Inter-speaker difference: Speaker differences in stop contrast
Fundamental frequency contours in the sentence beginning "Jane adored", illustrating differences in voicing.

Since maintaining vocal folds vibrating requires special maneuvers from the speakers to compensate for the decreasing transglottal pressure, inter-speaker variations are expected. Figure I.12 illustrates such inter-speaker variations. It displays the Fo contours for the same utterance spoken by two Japanese professional announcers. The first speaker has a typically larger drop for the [b] consonant, and such observations can be extended to all voiced stops.

Figure I.12 : Interspeaker variations: Fo drop in voiced stops.
Fo contours during the same utterance spoken by 2 Japanese professional announcer, The arrows point to the micromelodic fluctuations due to the presence of an voiced stop.

c) Fricatives

The fluctuations corresponding to the voiced fricatives are similar, but the minimum Fo value correspond to the point where the obstruction is maximum that is about the middle of the consonant (see figure I.13), while it happens just before the release for the voiced stops (note that the Fo tracker fails to detect Fo when the signal is too weak, but the vocal folds are still clearly vibrating).

Figure I.13: Fo drop in voiced stops and fricative.
Fo contours with intervening voiced stops and voiced fricative.

b) Cointrinsic influence

The higher Fo values at the vowel onset after a voiceless consonant are well explained by a progressive assimilation: the tensening of the vocal folds for ceasing vibrations leads to higher Fo values, particularly after aspirated consonant. Such higher values are audible and give rise to phonological changes. For examples, some of the high and low tones differences in tones languages historically stem from a difference between voiceless and voiced consonants respectively (see Hombert): Vowels preceded by a voiceless stops became high tone vowels, and the vowels preceded by voiced stops become low tones vowels.

The higher Fo values due to the presence of a preceding voiceless consonant explains the earlier peak in the vowel intervening after a rise, ceteris paribus. We may predict that the vowel in the initial syllable in the word "computer" (lexical stress on the second syllable), which is supposed to be Low, will be higher than expected because /k/ is unvoiced (which tends to heighten Fo) and its initial position.

3) Vowel intrinsic Fo values

a) Influence of vowel articulation on Fo values

Since long, it is known that ceteris paribus, close vowels (like /i/ and /u/) have an higher intrinsic Fo value than open vowels (like /a/)[6] . The difference can be as much as 15 Hz and varies according the speaker and the position of the vowels within the word (more different in the stressed syllable than in the unstressed ones) and the position of the word in the sentence (more difference at the sentence beginning than in the sentence end) (Ref.: House and Stevens on JASA, Intrinsic and cointrinsic, Phonetica, Di Cristo and Hirst on Phonetica, Silverman on Phonetica). One of the hypothesis is that such effect is due to the heightening of the tongue for closed vowels (tongue pull hypothesis)[2] .

The intrinsic and the cointrinsic influences combine, so that we can expect the following order:

ga<ka; gi<ki; ga<<ki, etc.

where < means lower Fo, and << much more lower.

b) Influence of Fo contour on vowel articulation

---

[6]The close vowels tend also to be shorter in duration than the open vowels, and the most common hypothesis is that it takes time to "open the mouth"

[2] There are other hypotheses as well.

In turn, it seems that the Fo contours may affect the position of the tongue during the production of the vowels, through biological interactions between the tongue and the larynx. In vowel [i], the articulation in higher Fo mainly produces F2 shift to a higher frequency. In low vowel [a], F1 shift, although less marked. The CT shows a lower coefficient for low vowels than for high vowels: jaw opening gesture (vowel [a]) may affect Fo control (Honda, 1991).

The intrinsic values may also influence the pitch accent shape. A rising pitch accent on a low vowel tends to stay level and then rise, when the same pitch accent superimposed on a high may only rise. The TOBI notation postulate a "scooped accent" a low tone target on the accented syllable which is immediately followed by relatively sharp rise to a peak in the upper part of the speaker's pitch range. A lower Fo may be more realistic attested when the vowel is a back vowel.

2) Word Initial tensening effect on vowels and consonants

The effect of initial position can be summarized as follows. First, the **unvoiced stops** tend to become aspirated. Second, the **voiced stops** tends to become unvoiced. Third, there is an increased in microprosodic fluctuations in voiced stops and fricatives and even the **sonorants** may have the typical fall-rise pattern of the obstruents . Fourth, the word initial **vowels** tend to start with a glottal stop.

Factoring out the segmental effect due to the identity of the phonemes itself and to its position in the word is not easy. There seem to be an extra tensening of all muscles involved in speech production at word initial phoneme, a vowel, or a consonant. Such a tensening can be viewed in the higher velum position at word initial position, whenever the first consonant is (a nasal consonant or a non nasal consonant). It seems to extend to all articulators, including naturally the vocal folds.

The tense consonants become even more tense, and aspirated in English. The lax and unvoiced consonants become more tense, and this tension renders voicing even more difficult: they tend to be more devoiced in word initial position. The vowels tend to start by a glottal stop, which is typically a vocal folds tensening (in German and English). The tendency may vary, depending on the speaker. Figure I.14 illustrates the Fo contours in the beginning of the sentence "I always seems to follow my instinct ..." for three speakers. The first speaker inserts a glottal stop at the

beginning of the word "always" and "instinct", the second at the beginning of "instinct", and the third one inserts no glottal stop. The unvoiced stops may be voiced in non initial and non prestressed position, but the tendency is largely speaker dependent. Figure I.15 illustrates the Fo contours corresponding to the beginning of the sentence "It is difficult..." by three speakers. The word initial [d] is devoiced in all cases. One of the speakers (on the middle) clearly realizes [d] and [k] as devoiced. The unvoiced phonemes [d] and [k] tend to be realized as voiced by the other speakers. When devoiced, the stops tend to have no strong release.

The effect of position in word can be also seen on the signal: for a given speaker, the word initial sonorant tend to have a more increasing amplitude than the word final sonorant. Figure I.16 illustrates Fo contours and signal envelope for two utterances (extracted form the 200 sentences): "forty love", and the "smell", where /l/ is word initial and word final respectively (light and dark allophones of /l/). For the speaker on the left, initial /l/ has a clearly rising envelope, which contrasts with the rather leveled envelope for final /l/. In the same line of reasoning, final /l/ has a clearly falling amplitude on word final /l/ and rather level for word initial. For both speaker, the amplitude of vibrations of the vocal folds are more reduced at the very beginning of the word initial l l . This effect is not particular to the consonant /l/, but it seems (to me at least) independent of the language spoken. Remark that the effect for the voiced stops is a reduction of the amplitude of voicing **at the beginning** of the stops, when length, place of articulation has an effect on the **maintenance** of voicing.

Figure I.14: Micromelodic effects and word initial tensening

Figure I.15: Voicing of unvoiced stops in non initial, non stressed position
The sequence "It's difficult" as spoken by three spoken. As it can be seen on the top, the consonant /f/ and /k/ have been fully voiced.

Figure I.16: signal envelope in word initial and final /l/.

Summary

The above remarks are some of the necessary ingredients to formulate a model with physiological motivations. There are relatively few places on the world where physiological data are taken and they are the most useful to validate the models. The necessity of constructing models on physiological grounds has been advocated by Ohman (1967) and

Fujisaki (1971). Ohman 's model has three inputs: the sentence, the word and the segmental input. Fujisaki's model has two inputs: the phrase (resetting of the baseline) and the word accent (local rises and falls). It may be the case that in Japanese, the identity of the phonemes (intrinsic and cointrinsic influences) intervene less in the general Fo contour, or that none of the small subset of speakers selected in Fujisaki's studies for study has strong micromelody influences. Four input is probably the minimim necessary to resynthesize unconstrained sentences of different sizes in general. It is not clear, except for the segmental input, whether or not and how far each input (sentence, phrase, word) correspond to a separate physiological input. In lack of a definitive answer, we may hypothesize the following scheme:

- the **sentence level** is mainly related to the declining volume of air in the lung and, declining subglottal air pressure, and the initial raising and final lowering of the larynx for phonation. The sentence level is related to the **very slow declining** (often exponential) tendency from the first phrase to the last in the sentence, to the sentence initial rising tendency, to the sentence initial and final Fo values, and to the presence of the maximum Fo value toward the sentence beginning.

- The **intermediate phrase level** related to the **resetting of the baseline** (in speaker lower Fo range) or of the top line (higher Fo range of the speaker), due to an elevation of the larynx by extrinsic muscles.

- the **local rises and falls** in the sentence to cricothyroid activity (tensening for raising pitch, relaxation for lowering) and other muscles for rising and lowering,

- and the **segmental input** corresponding to the **tongue pull hypothesis** for the vowels, to the change in **transglottal pressure** for the obstruents and to the contribution of the **laryngeal muscles** for stopping or maintaining vocal folds vibrations.

Unfortunately, there are not enough physiological data (they are difficult to obtain, and sometimes painful for the subjects) and not yet a enough clear understanding to control such models in a proper way. But the goal well formulated by Ohman for Swedish and Fujisaki for Japanese to a build model which reflects the Fo control mechanisms of humans should be kept in mind for the descriptions of other languages as well.

1.2. PERCEPTUAL RELEVANCY

Speech is produced by human to communicate with other humans. Not only the production mechanism, but also the perceptual mechanisms of the listeners have to be taken into account at some level in the model. While it is known that the human ear is very sensitive to fundamental frequency (and able to perceive change of 1% for a stationary vowel), the sensitivity varies with the position of the syllables in sentences (a difference of 10 Hz may not be perceived, even 20 Hz in some positions in sentences, as I have shown for French, Vaissiere, 19xx).

The notion of perceptually relevant contour comes from Holland, and the work of Collier, t'Hart and Cohen (the so-called Dutch school). Two originally complex contours may be represented at some stages by the same perceptually equivalent contour[1] . Some of the details may be therefore considered are perceptually irrelevant. This representation is important since only audible details are candidates for carrying pertinent information. The method for schematization consists in modifying original contours by steps, and to listen if each modifications is heard. I use such a method at ATR during my stay in 1992.

A number of researchers have follow this line of reasoning as a first priority, some ignore this concept, some find it difficult to apply, some think that only the production aspect of Fo is important.

In any case, there is a link between the production aspects and the perception aspects. On one side, we speak in order to be understood, so it is not economical to make efforts to create contrasts that cannot be decoded by the listeners, just because they a not audible. It is economical to use as much as possible the physiologically related characteristics, the natural tendency, as standard Fo, and to create contrasts from this "neutral" patterning. If a contrast is needed, it is also economical to make contrast in the part of speech where small production movements (less articulatory effort) are the easiest to decode perceptually (great perceptual effect) such as on the final syllable, before the pause, etc.). Languages general and specific segmental and suprasegmental aspects with the present form must be the results of a long evolutional, historical compromise between production

[1] The use of straight line (such as in the original models) or smoother lines to represent equivalents contours are not important considerations (the use of second-order and spline equations seems to fit particularly well to observed contours).

and perception constraints (see Martinet's least effort theory and Lindblom's theory on adaptation of the languages)).

Unfortunately, it is not easy to measure a quantity of perceptually equivalent contours.

First, there is a **technical problem**. A slight modification of real Fo contours and resynthesis of speech lead most of the time to a degradation of the speech. Such degradation may responsible of masking details. The results depend therefore on the resynthesizing method: PSOLA method, considered as better than LPC synthesis, is by no way perfect. The results depends also on the capabilities of the listeners and on the questions asked.

Second, there is also a well known interaction between duration, Fo and intensity perception. For example, the direction of a glide (rising or falling) on short vowels cannot be identified and the vowel is perceived equivalent to a vowel with steady Fo values. Vowels with higher Fo are perceived with higher intensity and vice versa.

Third, each movement is not perceived independently. This was clearly demonstrated in the famous experiment where Fo of the utterance "for Jane" was systematically changed (Garding and coworkers), and presented to Swedish and American speakers. The perception of the Fo movement on the second syllable depends on the height of the first syllable. A larger amplitude of the Fo rise is needed on the final syllable to be perceived as a rise (or a question), if Fo on the first syllable is higher. As a consequence, two contours differing on several syllables may be equivalent. This is important to understand intra- and inter-speaker variations. One of the implications is that the ear is sensitive to relative change of Fo values, not to high or low levels. Some authors prefer the Fo description using R and L notations (Rise and Lowering) rather than H and L (High and Low), due to this perceptual reason.

For all reasons, and probably other, it is difficult to measure the perceptually relevant contours in quantity. Nevertheless, psychoacoustic considerations and perceptually relevance are important aspects that cannot be ignored. It gives at least the feeling of what may be important to examine in the Fo contours and what should be left out of the descriptions. Because of that, it is important to have the facilities of changing Fo and duration in an easy way when examining Fo contours. Such a tool has been developed in certain labs.

1.3. PHONOLOGICAL MOTIVATIONS

No need to say, it is a necessity to draw a more or less direct relationship between the phonological representation of the sentences and Fo contours. Models without phonological constraints in their input (exact timing of the movements relative to the underlying segments) have no explanatory force.

What does a phonological representation adapted to prosody should look like is not entirely clear.

First, the **lexical input** (such as in a dictionary) delivers information about the word *category* (content/function, lexical/grammatical, dependent/independent), the word and morpheme *boundaries*, , on one side, lexical tunes (tone languages) , the type of accent (two in Swedish), the location of the stressed syllable (in stress languages like English, Italian, German), the position of the pitch accent or whether or not there is pitch accent (Japanese), etc. on the other side.

Second, a number of language dependent rules for **compound word formation at the phrase level**, etc. have been elaborated (such as the famous nuclear stress rule, and the compound stress rule for English, Chomsky and Halle, 1968; see also the work by Sagisaka and Sato, 1984, Pierrehumbert and Beckman, 1988; Kubozono, 1992 for Japanese), to convert the individual lexical patterns to the pattern in a single compound word pattern. Typically, a compound word in English carries the main stress in the first word and the second in the second word (initially stressed): "The 1'bath 2'plug". The nuclear stress rule predicts primary stress on the final word: "The 2'presbitarian 1'minister". In Japanese, the compound formation process predicts a unique pitch accent in the whole phrase, etc. Such rules leads the displacement or suppression of some of the characteristics of the underlying components and a hierarchy between the different prominence.

A simple **word and phrase** (nuclear stress and compound rules) input suffices to give a good account of the observed Fo in some (read) sentences, as it seems in Japanese and in short sentences for other languages. In French, the Fo contours superimposed to a phrase depends greatly on the position of the phrase in the sentence: the addition of a **sentence command** is absolutely necessary. The adding of a **sentence command** in the English sentences allows to take into account the fact

that in sentence initial position, the first stress tends to be reinforced, overwhelming the application of the phrase rules. In sentence initial position, the adjective in the sequence (adjective + noun) tend to be relatively more stressed because of its initial position, than in the rest of the sentence.

The examples that we will developed later shed light on the importance of the logical groupings of the word or semantics. The main function of prosody in read speech is demarcative, and grouping is a recursive process. The notion of **sense groups** (variants of prosodic word, phonological word, phonological phrase, bunsetsu, syntagma, hat pattern, etc. depending on the authors and the language under consideration) is the basic notion to explain prosody in a large number of languages. A sense group (and other equivalent word) is composed by one (at least) or more words semantically related in the sentences. The basic rises and falls on the stressed syllables in English read sentences are adequately explained by this basic grouping function. If two words are relate, the Fo in the stressed syllable of the first word will tend to rise, and the F0 in the stressed syllable of the last word will tends to fall (the process if recursive), forming an hat pattern[2] . This observation is not new, it was largely exemplified in Maeda's thesis.

If the combination of a local rise and fall indicates primary grouping into phrases (adjective(s) + noun, compound word), the successive reaching of the baseline indicates the succession of larger grouping. The presence of a pause, eventually preceded by a continuation rise indicates that all preceding phrases are to be grouped ("function integrative). The resetting of the baseline announces that a new grouping starts. The final pause and preceding very low Fo value indicate to the listener that the sentence is ended, etc. Lengthening phenomena have also a demarcative function (final lengthening) and a culminative function (stress, focus marking). The "prosodic" code does not seem very complicate, as expected, as it should be decoded by the listeners. The

---

[1] Notated by L* accents followed by a H- phrase accent and H% boundary tone: (L*) L* H- H%.iin the TOBI convention.

[2] The rise seems to equivalent to the TOBI notation L+H*, a rising peak accent on the accented syllable preceded immediately by relatively sharp rise form a valley in the lowest part of the speaker's speech range, and probably the fall is notated by H*+L (notated in TOBI by a single H*), may be followed by the break index values 3 (?), which is an intermediate phrase boundary notated by L-, so the notation for the complete succession becomes L+H* H* L- 3 (I am note sure to be correct).

prosody in read sentences reflects the structuring of the global meaning into components, and as far as I am concerned, despite apparent differences, it seems to deliver about the same information, at least for French and for English, in read speech. The most important input to the model, reflecting the demarcative function of prosody, should be exactly the same for French and English. The prosodic differences are superficial, only due to very different input at the lexical and segmental level.

It would be wrong to start with a hypothesis that classical phonological concepts are sufficient to describe Fo fluctuations and that the phonological representations is well adapted for other purposes than phonology. The use of exclusively H and L symbols to describe variations between speakers or the differences in phonetic realization due to the identity of the underlying phonemes is not sufficient. Pragmatic information, influence of rhythm, inter-and intra-speakers differences have also to be input of the model.

1.4. GENERALITY OF THE MODELS

A) Language-independent

All languages have the same function: to communicate. All human are equipped with the same production and perception apparatus, and have the same processing capabilities, I hope.

Each language has particular ways of structuring the sentence, forming compound word, indicating the word function (subject, or actant objet, indirect object, verb, etc.), grouping the words or phrasing (determinant and determinate), indicating the degree of dependency between the words (dependent adjective and nouns), topic marking or focusing a word. Each language uses a (limited) number of means. These means are **lexical**, **grammatical**, such as a particular **word order** (subject-verb-object, like in French and English), the declination of the words (Latin), the use of affixes or suffixes, or particles (like in Japanese, "*wa* " and "*ga* " for topic marking) and **prosodic**.

Some common tendencies can be seen in the linguistic use of prosody:

- initial tensening, Fo jump from one syllable to the next marking the **beginning**,
- final relaxation, lengthening, low Fo marking the **final relaxation**,
- and **higher Fo** and Fo rise generally correspond to the notion of not finite, question, doubt, hesitation, unfinished and **final tensening**.
- higher Fo values on the ed parts of the utterance, and reduced Fo range on the other parts.

When studying a particular use of prosody in one language, it is necessary to study all the means that the speaker is free to use.

B) Dialect-independent

It is important to distinguish well dialect characteristics. One model that fits well to one dialect, may be less adapted to another dialect.

C) Style independent

1) Read and spontaneous speech

The models should be able to describe not only read speech, but also spontaneous speech. The main difference is that read speech is not addressed to someone specific and the message is complete. In

spontaneous speech, Fo may carry more emotional information and the comprehension of the message may require pragmatic information. At least two more inputs are necessary: the **pragmatic** input and/or the **emotional** input (neutral feeling, the speakers don't believe what he says, has doubt about what he says, wants to emphasize one word in the sentence, etc.) More and more studies are concerned with spontaneous speech and models developed form read speech are proved not sufficient. But the description of read speech itself is by far not complete, and as stated in the introduction, there is not yet a uniformly accepted way of transcribing prosody. In particular there is a lack of work on the description of inter-speaker variations on speech, and the effect of hyper- versus hypo-articulated speech (see the work of Lindblom's group in Sweden, on the effect of hypo- and hyper-articulation).

2) Casual and non casual styles

When studying one language, it is important to identify the style of the sentences under study. In formal French, the word order is rather strict (subject-verb-object), and emphasizing a word requires to change the sentence. In less formal speech, it is possible to use up to a certain extend higher Fo values on the word to be emphasized. In very unformal style, the word order becomes almost completely free, and French becomes similar to Japanese for topic marking: the word under topics become the first word in the sentence.

example (casual speech)

Aujourd'hui, il fait chaud, hein?

Il fait **chaud**, aujourd'hui, hein?

Hein qu'il fait beau aujourd'hui?

D) Speaker-independent

No need to say, a model should be adaptable to different speakers. In standard French, I have divided the speakers into four categories dependent on their main tendency. Such a division may not be sufficient but eases the description. It will be used for describing the differences between the English speakers[7].

1.5. MULTIPARAMETRIC MODELS

[7] I have been really surprised by the similarities in inter-speaker French and English differences while conducting this study at ATR.

It is common to equate prosody with Fo, or at least to design it as the "most important" prosodic parameter. Fo contours are relatively easy to obtain, and most studies concern this parameter.

The study of **duration** of the phonemes requires much more time and requires a manual segmentation of the phonemes. Moreover duration is the most basic variable in not only prosody but also in segmental, since it determines the time, In fact, The availability of segmented and labeled large data base has changed this aspect (CF the work of Nick Campbell). The present data base have been segmented and hand-labeled and the correspondence between Fo patterning and durational patterning could be systematically exploited.

Reduction is also another important parameter. In Russian , reduction pattern seems to be even the most important parameters to detect the position of stress, at least according to Gordina and Ventsov (personal communication): the number of possible vowels in a word depends on the position of the syllable relative to the stressed syllable. Reduction phenomena in English are known to be very important (it is commonly believed that less reduction happens in Japanese and French).

A prosodic model, to be complete, should include Fo, duration and reduction phenomena.

1.6. APPLICABILITY OF THE MODELS

Since there is no ideal model, there exist a very large number of models, developed or under development. The models may differ just be having as a different **goal**. One model, judged as best by phonologists may look extraterrestrial to engineers and vice-versa...

A) Functional Model for Synthesis

Often, models developed for synthesis purpose imitates a single (professional or well trained) speaker. The input of such model is the orthographic form of the sentences. The goal is to produce intelligible speech and possibly natural.

B) Recognition

The use of prosody in Automatic Speech recognition has to cope with intra- and inter-speaker variability, and to rely on completely automatic process. The input are generally the Fo contours itself and the durations , the output being information about prominence, locations of stresses and degrees of boundaries. The program may be speaker-dependent or not, and the amount of information extracted depends on the sentence and the speaker. The goal is to produce extra information for the recognition process.

C) Prosodic data base labeling

There is also a growing need to label the prosodic events in data bases. While there is a rather good consensus on how to segment and label the segments, there still is no agreement for what concerns the labeling of prosody, particularly in spontaneous speech (with the pragmatic and emotional effects). There is not yet a clear idea on what to do exactly with such labeled data bases.

A transcription model has either to be simple to be used by human labelers or automatic. The TOBI (tones and break indices) association (electrical engineers, psychologists, linguists) represents a real hope in that direction (Beckman, 1993).

1.7 MODELS OF SUPERIMPOSITION

A model should allow to combine commands of different domain size, because speech is a structure with units of different sizes (the sentence, the phrase, the word, etc.) and studies have shown that each of these units are acoustic characteristics. The superimposition may be explicit such as in the Ohman-Fujisaki model (3 and 2 levels) , Maeda's modelisation for English, my own description for French (4 levels) or implicit, such as in the TOBI transcription (introduction of 4 break indexes).

More levels may be needed too. The syllables, the morphemes, the breath-group, the paragraph, have also acoustic correlates.

For example, the highest Fo values is generally found in the first sentence of the text, the lowest at the end of the last sentence, and there is a "resetting" of the Fo values at the beginning of each paragraph. A sentence extracted from the text can be identified by listeners as text initial, text medial or text final sentence (see Lehiste's work on that topic). Changing topic in a discourse is marked by higher Fo value, etc.

No need to say, not all levels are needed for a single sentence. The paragraph level plays no role in read isolated sentences. There is still an (unwanted) tendency to have the highest Fo value in the first sentence of the list, and the lowest in the last, when the speakers get tired of speaking unrelated, sometimes stupid, sentences. In short isolated sentences, there is no division into breath groups. In a very short sentences, composed let say by one or two meaningful words, there is no division into phrases. The longer the sentences the more divisions is needed, and a model working for fairly short sentences cannot be extended without damage into longer sentences without adding another level: there will be a small, but systematic error. A model with two inputs (like Fujisaki models) may work fairly well in medium size sentences with two level structures, a model with one input with very short sentences, and a model with a third level would be needed in longer sentences, because they may receive a three level structures.

In any case, a *superimposition model,* like the Ohman-Fujisaki model is convenient.

1.8 CONCLUSION TO PART ONE

No need to say, there is no ideal model satisfying all the above requirements. For that reason, it is very important to expose clearly the goal of each description.

The goal here is to propose a method for describing the prosodic structuring of the sentence uttered by the five speakers, and to put into light the speaker commonalties, the inter and intraspeaker differences. The five speakers are not professional speakers, and this fact makes the description more complex, as compared to trained speakers. They are reading, and this eases the description, as compared to spontaneous speech. What I have described so far is the many elements that must be taken into account when one wants to have a description model of Fo contours. In the next part, the basic principles underlying the descriptions are exposed in more details: the language independent division of the sentence into "acoustic" sense-groups by applying the hat pattern principle, the subdivision of the HPs into smaller units and their subgrouping into larger units, the principle of superimposition. The well-known (language-dependent) principle of the alignment of the Fo movements with the stressed syllables in English will be briefly reviewed.

PART TWO: BASIC PRINCIPLES

2.1) DIVISION INTO SENSE-GROUPS AND THE HAT PATTERN PRINCIPLE

A) The Dutch origin of the hat-pattern

The expression "hat pattern" (HP) is due to the work of Collier, t'Hart and Cohen. The basic HP pattern is composed by a rise from the baseline, a plateau, and a lowering to the baseline. It is represented in figure 1. The rise and the lowering are bounded to the lexically stressed syllable(s) in Dutch[8]. When words are acoustically grouped, the rise corresponds to the stressed syllables in the first lexical word, and the lowering to the stressed syllable in the last word.

B) Application to English by Maeda (1975)

In his analysis of texts read by 4 speakers, Maeda showed that the hat pattern principle is not only valid for Dutch, but also for English, also a stress language. Like in Dutch, the main Fo movements (rises and falls) along the sentences are bounded to the lexically stressed syllables[9].

A hat pattern in English may therefore also correspond to **one word**, or to more **semantically related words**. The rise takes place on the first stressed syllable and the lowering on the final stressed syllable. If the HP corresponds to only one word, the rise and lowering happen on that word. If the single word is a monosyllabic word, the rise and lowering characterize the only syllable. The function words and the unstressed syllable(s) before the rise and after the lowering tend to be on the baseline. The function words and the unstressed syllable(s) between the first rise and the final lowering tend to be on the plateau.

The basic principle (one rise and one lowering per sense-group) can be applied with modifications in non-stress language. Figures II.2 and II.3,

---

[9]In a way which is language-dependent. See the work by Nina Thorsen on Danish, a stress language, where the stressed syllable occupies typically a low Fo position, followed by rise located on following unstressed syllable(s).

(extracted from Vaissiere, 1982) illustrate the general tendency not (yet) contradicted by the data available for languages[10].

Figure II.2: General properties of Fo contours observed in a number of languages

Figure II.3: Fo contours of a sentence translated into English, Spanish and French.

C) Marking semantic boundaries by prosodic means

1) The acoustic and semantic definition of the hat pattern

The HP notion has both an acoustic and a semantic definition. It is acoustically defined as the succession of a rise starting from the baseline to the return to the baseline. it also corresponds to one "unite de sens", composed by one or more lexical words. The transcription of the Fo contours of the sentences will be therefore be based on the notion of Rise, Lowering and Baseline.

A transposition into a system like TOBI is probably feasible[11]. It will render however the task of describing speaker-differences more difficult. Moreover, TOBI transcription is partially based on the listener's impression.

2) The essential role of the baseline

The baseline is the acoustic cue used to define the limit of the hat pattern. Visual determination is an easy task.

The many examples illustrated in this report allow to "learn" how to draw the baseline. The baseline consists into a single line, or a series of segments. It consists in starting from the first Fo minimum in the sentence, drawing a line to the next most distant minimum, without crossing the Fo curve, recursively, up to the last Fo minimum. It is not always perfect, but I don't know of another strategy, at least for a speaker-independent, task-independent, even language-independent parsing (remember that the "baseline" can vary from one sentence to the next, ven when the speaker repeats the same sentence). The number of times where Fo values reach the baseline indicates the number of acoustic sense-groups

[10]Note that there are several thousand languages, and only very few are studied. The expression "language-independent" Iis "un abus de langage".

[11]As indicated in Figure II.1, the equivalent TOBI notation seems to H* for the first rise, H* for the lowering followed by the phrase accent L- tone. Both stressed syllable are interpreted as having an high target, and the second one is followed by a low target.

in the sentence, that is the number of HPs. Non factoring out microprosodic effects are responsible for errors, but as seen before, their complete elimination is not entirely justified on the prosodic point of view, since accentuation of the fall-rise pattern is a prosodic effect.

Each of the HPs delimited by the baseline may receive a similar Fo pattern.

They may also a slightly different pattern. More levels are marked, either by **dividing** a single hat pattern into smaller units, or by **regrouping** two or more Hat patterns into larger units. On one side, a single HP containing a large number of syllables or words is likely to be further divided. For example, the "long" HP corresponding to "Presbyterian minister" is likely to be further divided (the data will be shown in Part 3). On the other side, in sentences containing more than two HPs, two or more of the HPs are likely to be regrouped into larger units, for structuring the sentence prosodically.

The very interesting point is that dividing a long HP into smaller units or regrouping small HPs into larger units is done using mainly language-independent feature[12], which will be listed in the following.

2.2) SUBDIVISION AND GROUPINGS OF HAT PATTERNS

A) GROUPING OF HAT PATTERNS

Figures II.4 and II.5 illustrate the general means for grouping hat patterns. The grouping of two HPs can be done either by reducing the strength of the boundary separating them, and/or to augment the strength of the boundary between the other HPs in the sentence.

Figure II.4: General means for grouping hat pattern

Figure II.5: Examples

1) Increasing the boundary marking between two HPs.

1) A **pause** can be inserted between the two HPs. The length of the pause can be adjusted to the strength of the boundary (see the very interesting work of Grosjean and Deschamps on performance structure in English and French, and Kaiki and Sagisaka for Japanese, 1992).

---

[12]It will be interesting to systematically compare the use of the same means in Japanese and French, on the same 200 sentences translated in Japanese and French, or on another of sentences, or even better on a meaningful text.

2) There may be a **resetting of the baseline**[13] after the boundary.

3) The boundary can be marked by a **leveling** of the baseline after the boundary[14].

4) An **extra-low Fo values** or a **continuation rise** can be realized on the final syllable of the preceding HP[15].

5) Extra **phrase final** lengthening increases the strength of the boundary.

6) An increase in the magnitude of the **following rise on the next** HP indicates a break.

2) Decreasing the boundary marking between two sense-groups

7) Decreasing the **fall-rise**[16] excursion between the two HPs. This is done by decreasing the magnitude of the preceding fall and/or the following rise (see the work by Wayne Lea in English).

8) Decreasing or suppression the **word final lengthening**

9) Decreasing the height of the **plateau** of the second HP.

10) More dramatically, **grouping** HP1 and HP2 into a single HP.

B) SUBDIVISION OF THE HAT PATTERN

Figures II.6 and II.7 illustrate the means for indicating a boundary between the words embedded into a single hat pattern.

1) There may be an **extra Rise** on the plateau corresponding to the final syllable (Rp)

2) There may be a general **lowering of the plateau** between the two stressed syllables

3) or a Lowering of the **function words** on the plateau

4) The **lengthening** at the end of the final syllable of the preceding lexical word.

Figure II.6: Subdivision of the HP

Figure II.7: Illustration of the subdivision

---

[13]The resetting correspond to the Phrase Accent in Fujisaki's model.

[14]A leveling of the baseline happens in most long sentences, but the speaker can controlled to a certain extend its position.

[15]French uses a series of final tones. Two tunes are generally assumed in the description for English.

[16]The "fall-rise" expression comes from Lea's work in English, and describes adequately what is happening. The reading of the papers by Lea on the use of prosody for automatic speech recognition is highly recommended.

One, several or all means can be used to increase and diminish the strength of a boundary. Each of the preceding means is attested in the 1000 sentences and it will be illustrated in Part III.

C) PERCEPTUAL STRENGTH OF ACOUSTIC BOUNDARIES

It would be interesting to evaluate the **perceptual strength** of each boundary from the acoustic manifestations[17].

While it is reasonable to consider that a *large* pause marks more disjuncture than a *smaller* pause, and that an *higher* resetting, or a larger next rise marks more disjuncture that a *smaller* resetting or rise, and that a pause + an higher resetting + final lengthening mark more disjuncture than a pause or a resetting alone, it is very difficult to estimate the perceptual value when "means" of different kinds (pause, Fo, lengthening[18]) are combined. Systematic perceptual experiments could be done.

A rough estimation is however sufficient for our purpose. For the description of the various languages, my students[19] put one point for every means. What is important is not the exact value of the strength of each disjuncture, but the relative strength between pairs of HPs in each particular sentence. By basing the decision on **relative** strength inside the sentence to be analyzed, it is feasible to cope with style, rate of speech and speaker differences. The number of acoustic levels in short sentences is typically low, as expected, and can increase dramatically in long sentences. A study of **speaker differences** should evaluate the preference of each speaker for a particular disjuncture means or for a combination of means.

2.3) THE PRINCIPLE OF SUPERIMPOSITION

The sentence should not be considered as not a regular succession of independent sense-groups. The phonetic realization of a sense-group is influenced by its position in the sentence and in the breath-groups (if the

---

[17] In the TOBI transcription system, four or five levels of boundaries are marked by listening in the sentences: 0 (for the strongest perceived conjoining) to 4 (for the most disjoint).

[18] Lengthening or a strong fall-rise pattern are often perceived as "virtual pause" in French.

[19] As a matter of fact, the learn very rapidly how to read spectrograms and decode that prosodic profiles of the sentences in French since their performances are rated and are part of the final examination.

speaker takes a breath during a sentence). Figure II.8 illustrates the principle of superimposition[20].

Figure II.8: Superimposition Fo model

The basic sentence, breath-group and HP Gestalts[21] are slightly different for a particular language. For example, the model of superimposition for French postulates strong acoustic differences between final breath-group[22]. In English, the difference between final and non final breath group is much less dramatic. For example, to the perceptually prominent French continuation rise corresponds sometimes a small optional little hook in English (or an extra low value).

2.4) ALIGNMENT OF THE Fo MOVEMENTS WITH THE STRESSED SYLLABLES

A) The Rise

When the sentence starts by a vowel , the rise involves the whole vowel (figure 9). When it starts by a glottal stop, it tends to start higher (figure 10). When the stressed syllable starts by a sonorant, the rise encompass the sonorant and the vowel (figure 11). When it starts with an unvoiced vowels, most the whole rise occurs during the unvoiced stops.

Figure 9: Fo of the beginning of the sentence "Alf's brother..."
Figure 10: Fo at the beginning of the sentence " Henry..."
Figure 11: "They launched in to battle..."

B) The Lowering

When the final stressed syllable in an HP is word final, most of the lowering occurs during the vowel. In other cases, the stressed syllable is high, and the lowering occurs during the following unstressed syllables (see examples on figures 12, 13 and 14).

Figure 12: Superimposed Fo contours in the first stressed and the last stressed syllables in HP's. The Fo contours have been lined up with the onset of the stressed vowels.

---

[20]See the explicit superimposition models: Ohman and Garding for Swedish (1967), Fujisaki for Japanese (1971), Vaissiere for French (1971), Maeda for English (1975).

[21]In the German gestaltist theory, the basic gestalt if the "pregnant" form. To be "pregnant" (with the German meaning for the adjective "pregnant"), a form has to be simple, and frequent.

[22]Corresponding to the "marked" and "unmarked" breath groups in Ph. Lieberman's notation for English ((Lieberman, 1968).

Figure 13: Lowering on the stressed syllable of the monosyllabic words inserted in an hat pattern "It is difficult *to choose...", "... the rope...", "... have proof...",* " The length of *her skirt...",* "When forced to make *a choice.*

Figure 14: Lowering on words with stress on the initial syllable.

Figure 15: Summary

2.5) EXAMPLES OF PROSODIC PARSING IN ENGLISH

Figures II.16 and following illustrate the results of manual parsing of the sentences, directly from Fo contours. The figures illustrate the original Fo contours , the visually schematized contours, and the results of the parsing.

An "automatic" schematization of the Fo contour is under way at ATR (by P. Taylor, T. Hirai and H. Valbret[23] for English Japanese). I use to schematize Fo curves for French using directly the output of the segmentation and labeling module of the analytic speech recognition system, the Keal system.

Figure 16 and Figure 17 correspond to the same sentence spoken by two speakers in a slightly different manner. They both divide the sentence into four HPs. The first HP is undivided, while the other HPs are divided into 2 or three subunits. The leveling of the baseline was done between HP1 and HP2 for gsw and between HP2 and HP3 for the second, indicating a different grouping of the first HPs. As a consequence, the prosodic tree is slightly different, and the acoustic differences seem to correspond closely to the perceptual impression. The adding of the duration parameters may change however the strength of the boundaries.

I have applied manually the principles to a large number of sentences. The boundaries located from Fo contours correspond almost always (as far as I can judged) to the main boundary that I would have derived by listening only to the sentence.

There are errors. First, the **function words** may join with the preceding HP. Second, the **disyllabic words with stress on the second syllable** (only 20% of the disyllabic words are stressed on the second syllable in English), corresponding to a single HP pattern, terminated with a continuation rise cause a problem (there are very few cases in the data

[23]Helene Valbret is currently working on the automatic determination of the baseline, using the algorithm describe using (on Japanese data).

base, but the problem remains). Third, **long words** have caused an error here and there (there are very few long words in the data base) and some have divided into two subunits. Fourth, part of the sentences are spoken **without Fo fluctuations** (the speakers get a little bored uttering unrelated sentences) and could not be parsed. Five, there are some sentences with **focusing**, which disturbs the syntactic patterning, as exemplified later.

Figures II.16, 17 and 18: Manual prosodic parsing of some sentences in English

Such parsing could be compared and completed with Nick Campbell's results obtained by durational parsing[24] (a task that I had not time to do during my stay), in order to derive the basic correspondence between the durational and Fo profiles in English (see the French two profiles in Vaissiere, 1991). A good parsing in French requires both Fo and duration. It is probably also the case for English, and it may correct for some of the errors, I hope.

[24] on the same set of sentences

PART THREE: THE STRUCTURING OF THE ENGLISH SENTENCES

The purpose is to illustrate the manual prosodic parsing for the Fo curves of the sentences with a large number of figures.

3.1) DECLINATION LINE AND SPEAKER RANGE

A). The natural tendencies

As said before, the natural tendencies are to a declining Fo along the sentences, and a decreasing Fo range (see I.I b). Variations in declination and range allow to create prosodic information.

B) The use of variation in range

1) Speaker-dependent topline and baseline

Same sentences have a simple declination scheme: Fo is declining is a smooth way along the sentence, with a tendency to start with a rather fixed Fo value, and to end also with a rather fixed F0 value (both values are naturally speaker-dependent).

Figure III.1 illustrates the superimposed Fo curves for the first sentences in the list, as spoken by two speakers, a female (top) and a male (bottom). The Fo contours have been times aligned on the first stressed syllable. As it can be seen in that figure, the Fo values for each tend to fluctuate between two lines, called the top-line and the baseline, which are characteristics of the speaker (but which fluctuate from one sentence to the next). The baseline tends to decline more in the first part of the sentence, at least when the data are plotted with a linear scale (language-independent feature). The first point above and under the topline and the baseline corresponds all to errors. Some of the typical halving and doubling of the Fo values are illustrated in Annex 1.

Figure III.1: Speaker range
Superimposed Fo contours for the first sentences in the list, as spoken by a female (top) and a male speaker (bottom). (because of an unsolved problem with the software Waves, and because of a lack of time, the Fo contours in this figure are not time-normalized and the time scale is disturbed).

2) Pattern of reduction of the Fo range along the sentence

a) Inter-speaker variation in the use of his(her) pitch range

While reducing the Fo range from the beginning to the end of the sentences should be considered as a normal tendency, there are speaker differences in the pattern of range reduction. Figure III.2 illustrates such differences. The speaker on the top usually continues to use about 2/3 of his total range all along the sentence, while the second speaker reduces it dramatically after the first Hat Pattern. There are two consequences. First. parsing from beginning to the end of the sentence is easier and done with much more confidence in sentences in which the speaker continues to use a larger pitch range. No need to say, parsing from Fo only is not possible when there are no Fo fluctuations. Second, after looking at the Fo curves of the first few sentences in the list, it is possible to predict how the speakers may utter the rest of the sentences. When the expected pattern of Fo range for a sentence is different from expectations, the sentence is heard as marked (focus, desaccentuation of a part, bored, or more alive, etc.).

Figure III. 2: Speaker-difference in the Reduction of Fo range along sentences Fo contours of two sentences " The length of her skirt causes the passers-by to stare." (left) and "It is strange that I slept so long since I was not feeling tired." (right) for two speakers (gsw: top and mlp: bottom).

b) Intra-speaker variations and focusing

Inside a sentence, reduced range is generally used for parenthesis (BL= low, flat and reduced range), for information considered as less important by the speaker, for expression boring in reading the sentences, for focusing another part of the sentence. Increased range is known to be used for excitation, interest, focusing, etc. (see the nice work of Garding for Swedish). An extra high or an extra low values may be used occasionally by the speaker, for specific purposes. Focusing will illustrated later.

C) The use of variation in the baseline

1) Declination as a general Declarative sentence characteristics

Figure III.3) illustrates the general use of the baseline as attested in a number of languages. represents the general tendency. The speaker can exert a certain control: adjustment of the slope to the length of the sentences (work on Danish, English, French), the leveling of the baseline

(not the leveling, but the point where leveling starts generally represents a break, as seen previously), resetting, preceded or not by a pause (and eventually by a continuation rise. and downstepping of the BL (attested in a number of African languages).

For the description of the (declarative) CSTR sentences, only BL leveling, and BL Resetting was needed. Leveling and resetting are disjuncture marking and they correspond to breaks in the sentences, but not all breaks are accompanied with leveling and resetting.

Figure III.3: Declination tendency in declarative sentences.

2) Marking the Sentence type

Figure III.4 illustrates the global control of the baseline for contrasting sentence types as used in a number of languages (including naturally English): suppression of the tendency on yes-no question, and exaggeration of the tendency in orders.

There are no yes-no questions, and no imperative sentences in the list.

Figure III.4: Declination tendency and sentence types.

D) How to decide that a resetting has occurred?

How to decide whether or not a resetting has occurred inside of sentence? (equivalent to the phrase command of Fujisaki) ? Is it possible to find a formal criteria?

The answer is yes and no. For deciding whether or nor a resetting had occurred in French and German, I personally compared the height of the **function word**(s)[25] preceding and following the boundary. If the height of the following function word(s) is higher than the height of the following function, then a resetting is hypothesized. If it is equal or lower, there is not resetting hypothesized. If there are no comparable function words before and after the boundary, no decision can be taken (rule 1) or it may be decided that a resetting has occurred if the Fo value on the first syllable after the vowel on the baseline is higher (rule 2) The same criteria cannot be applied to all languages, because not all languages have functions words.

[25]function words re articles, auxiliaries, etc... which tend to be uttered on the lower Fo range.

For example, in Japanese it could be the first (non accented) syllables of two successive lexical words: if one of the word is unaccented, or the one of the word carries accent on the first syllable, the criteria cannot be used. Each language has its own problem one has to cope with.

The criteria based on the height of the function words has been temporally applied for English, but English has less function words than French, and when there is no function words between the two HPs, I hesitate between rule 1 and rule 2. Rule 2 has failed for a very few examples, because of the occurrence of an Fo extra low value. An extra value inside a sentence is a disjuncture cue, and it naturally causes problem or the delimitation of the HPs after the break. It has been almost never used by the 5 speakers, but there may be other speakers which may more used of this extra low.

In the same manner, I calculate the **resetting for the topline**: when the highest value in a HP is higher than in the preceding HP, I assume a resetting of the topline (it could be interpreted also as an larger rise)[26].

3.2) DIVISION OF THE SENTENCE INTO SENSE-GROUPS

A) Marking prosodically the clauses

1) General rule

When the sentence is composed of two clauses, all speaker end the first clause with an low Fo value. Let us illustrate speaker differences.

2) Speaker differences in marking main boundary between clauses

Figure III.5 illustrates the Fo contours of the (long) sentence composed by two clauses "The government triumphed four years ago and we have every reason to believe that it will triumph again." by four speakers. As expected, the main boundary marks the end of the first clause "...ago": the *lowest Fo value* inside the sentence correspond to the syllable "go" of the word "ago", followed by a *continuation rise* within the same syllable for the second speaker. The second and the fourth speaker insert a *pause*. The Fo minimum is as low or even lower than the sentence final value. There is a *baseline resetting* : *For* all speakers, the function word after the boundary ,"and" is higher than the preceding function word "ago". The maximum Fo values in the second part of the sentence is lower

---

26 In the TOBI system, a non downstepped high will be assumed. For our purpose, downstep is the default behavior, because there is a "natural" tendency of reduced the Fo range form the beginning to the end of the sentence.

than the first maximum, at the sentence onset: no obvious resetting of the topline can be assumed, if the speaker is not identified.

However, when speaker-dependent reduction of the Fo range is taken into account, a resetting of the topline can be hypothesized with confidence. Compare for example the typical range reduction scheme for speaker gsw and mlp (top and bottom on figure III.2).

2) Influence of focusing

In most cases, the lowest values correspond to the final syllable of the first clause. The focusing of one word may contradict this general rule. Figure III. 6 illustrates the division of three sentences into two parts by a single speaker. The first sentence has been selected to illustrate the problem of focusing: "They launched into battle with **all** forces they could master". Because of the focusing on the word "all", the reaching of the BL happens on the preceding function word "with". The second and third sentences illustrates the general behavior: "The smell of the freshly grounded coffee // never fails to entice me into the shop.", and "The government triumphed four years ago // and we have every reason to believe that it will triumph again."

Figure 6: Main boundary marking: the problem of focus
Fo contours by one speaker of the sentences "They launched into battle with // all forces they could master", "The smell of the freshly grounded coffee // never fails to entice me into the shop.", and "The government triumphed four years ago // and we have every reason to believe that it will triumph again." (// indicates the position of the reaching of the BL).

B)) Division of the sentences into 2 sense-groups, 3 and 4

A sentence is typically divided into a number of HPs, by reaching the baseline. As explained before, I have tried to use a fixed criteria for deciding whether or not the baseline has been reached, which is illustrated on all the figures. The number of HPs is the number of line segments used to go from the beginning of the sentence to the end, without crossing Fo detected values[27]

The next figures illustrates the division of sentences into two, three and four HPs:

---

[27] Helene Valbret at ATR is testing the same order of idea to define precisely the baseline in Japanese sentences and her help and discussions have been greatly appreciated.

(It is futile)$^{Rc + Reset}$ (to offer any further resistance).

(Amongst her friends)$^{Rc + Reset + Pause}$ (she was considered beautiful.)

(The smell of the fleshly grounded coffee)$^{Rc + Reset + Pause}$(never fails to entice me into the shop.)

(John could lend him)^{Reset} (the latest draft)(of his work).

(Form forty love)$^{Reset + Pause}$(the score was new deuce)$^{Reset + Pause}$(and the crowd grew tense).

(I am often perplexed)$^{Extra\ low + Reset + Pause}$(by rapid advances)^{Reset} (in the state of the art technology).

Figure III. 7. Fo contours of the sentences:
t is futile to offer any further resistance.
Amongst her friends she was considered beautiful.
The smell of the freshly grounded coffee never fails to entice me into the shop.
They asked if I want to come along? on the barge trip.
The government triumphed four years ago and we have every reason to believe that it will triumph again.
When foed to make a choice Sarah chose ping pong as her favorite game.

Figure 8: Fo contours of the sentences
John could lend him the latest draft of his work.
Form forty love the score was new deuce and the crowd grew tense.
I am often perplexed ^{low} *by rapid advances in the state of the art technology.*

Figure 9: Fo contours of the sentences
They asked if I want to come along? on the barge trip.
The government triumphed four years ago and we have every reason to believe that it will triumph again.
When forced to make a choice Sarah chose ping pong as her favorite game.

3.3) COMPOUNDS WORDS AND ADJECTIVES

As expected, compound words form a single HP, and (adjective + noun) tend to form a single HP.

According to the Chomsky/Halle compound and nuclear stress rules, compound words tend to be initially stressed (1bath 2plug) and (adjective + noun) finally stressed (2Presbitarian 1minister)

A) Compound words: a single HP

The few compound words in the 200 sentences form a single HP (more data with larger compound words are needed to confirm that).

Figure III. 10 illustrates the beginning of the sentence Fo " The bath plug is missing so you have to take a shower.", spoken by four speakers. For the two first speakers, the compound word "bath plug" has been included into a single HP, and the BL is only reached at the of the following lexical word "missing", while it is reached before for the two last speakers. The single HP for the two first speakers has been further divided of described later.

spk gsw: (The bath plug is missing) so you have to take a shower.
spk jok: (The bath plug is missing)$^{Rc + Reset}$ so you have to take a shower.
spk mlp: (The bath plug) (is missing)$^{Rc + Pause+Reset}$ so you have to take a shower.
spk sab: (The bath plug is)(missing) so you have to take a shower.

As expected, Fo[28] is rising during the first (primary) stressed syllable "bath" and low (or slightly lowering) during the last (secondarily) stressed syllable plug. Fo height on "bath" is considerably higher than the height of Fo on "plug". For the top speaker (P2 tendency) the reaching of the baseline marks a boundary between "The bath plug is missing..." and "so you have to take a shower."

B) Adjective + noun: one or more HPs

Adjective+noun may compose one or more HPs.

Figure III.11 illustrates the Fo contours for the beginning of the sentence " The Presbyterian minister...). To the (adjective + noun) sequence correspond two HPs (according to our acoustic criteria) for spk gsw, and a single HP for spk mlp and jok. The grouping of the adjective and noun for the first speaker is done by resetting of the topline (Fo height in "managed" is higher than Fo height in "minister").

spk gsw: [(The Presbyterian)(minister)]^{Resettop}(...
spk mlp:(The Presbyterian *minister)(...
spk jok: (The Presbyterian minister)...

For the bottom speaker, the two successive words are regrouped into a simple shape, hat-pattern like, with a long plateau on the intermediate syllable: there is no subdivision of the HP. For the middle speaker, the rise

[28]The S shape of Fo may be due to the open back long vowel [a] and jaw opening gesture.

on the plateau correspond to the stressed syllable "mi", gives the second word a certain amount of independence. The so-called "Rise on the Plateau" (Maeda's expression), leads to the subdivision of mlp's HP into two subunits (see Part II.B and figure II.6).

3.4)FUNCTION WORDS

The function words tend to be uttered on the speaker lowest Fo range, as a general tendency.

They are low or lowering, but may even exhibit a rising Fo contour when located at the beginning of the sentence or after resetting.

If they are located in the middle of an HP, they are uttered in an higher range, as expected. They tend to lower toward the baseline .

A) Beginning of the sentence

Figure III.12 illustrate the variations in Fo height and movements in functions words. It displays the beginning of the sentence "You ought to brush your teeth..." for three speakers. The Fo contours during the pronoun "you" is falling (speaker on the top), or level (speaker on the middle) or rising (bottom speaker).

B) Inside a sentence

Inside the sentence, they may be falling, level or sometimes rising (see the function word "to" in figure III.12: lowering for jok, lowering then low for mlp, and rising for gsw. Listening to the sentences indicates that for the two first speaker "to" is grouped with "ought" (= you ought to + brush), while for the third speaker, a slightly stronger boundary is perceived between "ought" and "to" than between "to" and "brush" (= You ought + to brush).

C) Inside of an HP

When they are inside of an hat-pattern, they tend to be uttered on the plateau, i.e. the upper Fo range of the speaker. To create a subdivision of the hat-pattern, they may start to shape a fall-rise pattern on the plateau of the hat-pattern.

3.5) SENTENCE INITIAL RISE

A) WITH OR WITHOUT AN EXTRA PEAK?

The first rise in the sentence is often the largest rise in the sentence. It may or may not be accompanied by an extra peak (inter-and intra-speaker variations). Only comparison with the speakers habits allow to decide whether or not an extra peak has occurred.

Figures III. 13 and III. 14 illustrate the beginning of three sentences spoken by the speaker (mlp) " It's silly...", "Opportunities ..." and "It was important...", \"The smell ...", "It is futile...". The Fo maximum is reached, as expected, during the first stressed syllable. The maximum value is about 250 Hz for the third first illustrated sentences (figure III.13) and about 280 Hz) for the two last sentences. Spk mlp is inconsistent in associating an extra peak to the first rise. The exact location of the peak in the stressed syllable can be predicted from the identity of the underlying segments, with some speaker-dependent tendencies, as seen later).

B) PEAK LOCATION

Figure III.15 illustrates the Fo contours corresponding to the beginning of the sentence "I yearn for the day...", spoken by three speakers (sab, gsw and mlp).

spk sab: (I yearn for the day) $^{Rc + Reset + Resettop}$(when ...
spk gsw: (I yearn for the day)(when ...
spk sab: (I yearn for the day)^{Resettop} (when ...

The grouping of the two lexical words "yearn" and "day" are indicated by different cues by the speakers: Continuation Rise, reset of top- and base- lines for sab, and by regrouping the two words in the same HPs for gsw and mlp. As it can be observed, the Fo peak in "yearn" appears early for sab, between the nucleus and the coda for gsw and at the end of the coda for mlp. This reflect in part general speaker tendencies.

3.6) FOCUSING

A) How to detect focusing in a sentence?

The problem with focusing is that it is not possible to predict it. Moreover, focusing a part in a sentence renders the parsing difficult, and it may be not possible to emphasize a part of speech and to demarcate it in

clear chunks: the culminative function of prosody is used at the expend of its demarcative function ("culminative" and "demarcative" follow the terminology of Troubetskoy).

Focus is realized by contrast: generally, focusing a part of the sentence can be done into three ways: either focusing that part of the sentence to be focused by higher Fo and/or longer duration and/or higher energy , or destressing the other parts of the sentence or both (the final recipe if language-, word-, word position-, style- and speaker dependent...).

In neutral sentences, the maximum Fo is expected to happen during the first lexical word. Any other word, if focused will attract the maximum of Fo. If the maximum of Fo does not happen in the first lexical word of the sentence, a focus should be hypothesized. Focus can also be determined using normalized and smoothed measures of duration and energy (see Campbell).

B) Examples

Figure III. 16 displays the Fo contours for the beginning of the sentence "The world is becoming increasingly dangerous". as spoken by two speakers (gsw and mlp). The second speaker emphasizes the word "increasingly" by lowering the Fo values on the first word "world".

Figure III.17 illustrate the Fo contours of the beginnings of the sentence "It is futile to offer any further resistance", as spoken by two speakers. For gsw, the word "any" is treated as a function word and it is uttered in the lowest range of the speaker. For mlp, "any" receives a rise.

Figure III.16: "The world is becoming increasingly dangerous", as spoken by two speakers.

Figure III.17: Fo contours of the sentence "It is futile to offer any further resistance", as spoken by two speakers.

PART FOUR: SPEAKER FAVORITE PATTERN

When comparing the Fo contours of the same sentences pronounced by different speakers, it appears that they have often the same way of parsing them, but they realized different Fo pattern shapes.

To describe such variations in the realization of the HP in the present set of sentences, I use here the terms used for the description of the speaker differences in French: the P3 tendency, i.e. the typical hat-patterns (the magnitude of fall tends to be equal to the amplitude of the

rise), P4 tendency (the rises are more prominent than the falls), and the P2 tendency (the falls are more prominent than the rise, see Vaissiere 1975)[29].

Figure III.18 illustrate the basic variations in the realization HP .

Figure III.19 illustrates the same sentence "The candidate felt that a trapezoidal badge would have a more visual impact than the usual rosette.". The sentence has been divided into 4 HPs by the three speakers.

spk mlp: (The candidate felt) ^{Rc}(that a trapezoidal badge)$^{Reset+Leveling}$(would have a more visual impact)(than the usual rosette).

spk sab ; (The candidate)$^{Extra\ Low}$ (felt that a trapezoidal badge)^{Reset}((would have a more visual impact)^{Reset}(than the usual rosette).

spk gsw: (The candidate) $^{Rc+Reset}$(felt that a trapezoidal badge)$^{Rc+Reset}$((would have a more visual impact)(than the usual rosette).

The first speaker has realized P3/P4-like, the second P3/P2-like, and the third P2-like pattern.

For the first speaker (mlp), the rise is followed by slowly decreasing values on the plateau, and then by a lowering of smaller amplitude of the rise. The global shape is an hat pattern (P3), but the speaker tends to realize more clearly the rises than the fall (P3 tendency P4). For the second speaker (gsw), the rise and the lowering are equally important (P3), but there is typically a peak on the plateau (tendency P2). For the third speaker, the lowering are more prominent than the rise.

Figure III. 20 illustrates further examples. In the sentence "I am obliged to tell you that most women", gsw superimposes a P2 like pattern, and "stressed" more the final fall (lowering on "tell" larger than the rise on "bliged", lowering on "women" larger than the rise on "most"). The reverse is true for mlp (P3-P4 shape).

How far such tendencies are useful to classify natives of English remains to be tested. Such patterns may be characteristic of the style used by the speaker in the particular sentences. In French, different typical patterns are used in telling stories, journalistic news, etc...

[29]According to this classification, French is a P1/P4 language, Japanese is a P3 language, and English a P2 language....

CONCLUSIONS

There is no ideal method to describe the prosody of speech. The method described here is the best I could think of for describing the inter- and intra speakers commonalties and differences in the Fo contours in the set of sentences as spoken by the five speakers.

The **hat pattern approach**, with the use of a series of acoustic cues (pause, extra low or continuation rise, resetting of the baseline and/or the plateau, rise on the plateau) seems enlightens the demarcative function of prosody in read English in an acceptable way and it has the advantage of being very close to the Fo contours in the sentences.

A **patternist approach** eases the description of the inter-speaker variations.

Statistics on large data base and automatisation should be done. The **visual** examination of the 1000 sentences is a necessary step, to discover regularities and differences. Further thinking is needed for the integration of the durational profile of the sentence and for automatisation of labeling using pause, Fo and duration information.

Our approach in the present report is insufficient, since the **durational** profile of each sentence should have been studied in parallel with the Fo contour. As known, much of the specificity of a given language is due to the particular relationship between duration, Fo and energy. The independent study of a single aspect is not adequate to characterize the prosody of a language. There are indeed few multiparametric studies, but I am not aware of all the work done, and English is probably the most studied language. Nick Campbell's works on duration in sentences should be integrated to have more firm conclusions.

It would be also of highest interest to compare in details the profiles of pause + Fo + duration + energy + reduction in Japanese, English and French. It may be too early to assess directly real **spontaneous speech**, There are so many ways of spontaneously uttering a single sentence, from angry speech with a nuance of pity, and a bit of irony and fatigue to a soft voice, with a nuance of sadness and a bit of underlying agressivity, from self-adressed speech to conference.

FIGURES

REFERENCES

Beckman, M.E. and Gayle, N.A., (1993),
"The tobi handbook", (in preparation)

Bolinger, D.L.. (1972),
"Accent is predictable (if you are a mind reader", Language 48, 633-644.

Campbell, N. and Sagisaka, Y., (1992),
"Automatic automation of speech corpora", Proceedings of the Fourht Australian Interntional Cionference on Speech Science and Technology, Birsbane, 686-691.

Campbell, N., (1993),
" Predicting segmental durations for accommodation within syllable-level timing framework", ???
Campbell, N, (in prepatration), "Stress and focus detection in a speech corpus".

Camptell, N, (in press),
"Automatic detection of prosodic boundaries in speech" accepted for publication in Speech Communication

Chomsky, N. and Halle, M. (1968),
THE SOUND PATTERN OF ENGLISH, Harper and row.

Cohen, A., and t'Hart, (1967),
"On the anatomy of intonation", Lingua 19, 177-192.

Delattre, P., (1962),
"Comparing the prosodic features in English, German, Spanish, and French", Int. Rev. Applied. Linguistics., I, 193-210.

Di Cristo and Hirst
on microprosody for French, Phonetica.

Fujisaki, H. and Sudo, H., (1971),
"A generative model for the prosody of connected speech in Japanese", Ann., Res. Eng. Inst. Logopedics Phoniatrics, Univ. Tokyo 30, 75-80.

Garding, E. (1979),
"Sentence intonation in Swedish", Phonetica 36, 207-215.

Grosjean, F., Grosjean, L., and Lane, H., (1979),
"The patterns of silence: performance structures in sentence production." Cognitive Psychology, 11, 58-81.

Hirai, H., and Honda, K. (1993),
"Analysis of magnetic resonance images on the physiological mechanisms of dundamental frequency control", ATR Report, (in Japanese, summary in English).

Honda, K. (1991),
"A statistical analysis of tongue muscle EMG and vowel formant frequencies", Paper presented at the 122nd ASA meeting, Fall.

House, A.S. and Fairbanks, G., (1953),
"The influence of consonant environment upon the secondary characteristics of vowels" JASA 25, 105-113 (reprinted on Lehiste's book).

Kaiki, N. and Sagisaka, Y., (1992),
"Pause characteristics and local phrase-deendeny structure in Japanses", Proc. ICSLP 92, at Banff.

Kubozono, H., (1993)
THE ORGANISATION OF JAPANESE PROSODY, Studies in Japnaese Linguistics, Series Editoir, Kurisio Publishers.

Ladefoged, P., (1968).
on larynx lowering to maintianing voicing, Gloria thesis

Lea, W.A., (1974),
" Prosodic aids to speech recognition: a summary of results to date", Sperry Univac Technical report, No. PX 11087.

Lehiste, I., (1970),
READING IN ACOUSTIC PHONETIC. MIT Press

Lieberman, Ph, (1968),
INTONATION, PERCEPTION AND LANGUAGE, Mit Press.

Maeda, S., (1975),
"A characterization of Fundamental frequency of Speech", Res. Lab. of Elect., (MIT), Quarterly Progress Report, 193-211.

Maeda, S., (1976),
A CHARACTERIZATION OF AMERICAN ENGLISH INTONATION, Ph> D. Thesis MIT.

Ohman, S.E.G., (1967),
"Word and sentence intonation. A quantitative model", Speech Trans. \Lab. (Stockholm), Quarterly Progress Report, 20-54.

Pierrehumbert, J.B., (1981),
"Synthesizing intonation", JASA 70, 985-995.

Sagisaka, Y. and Sato, H., (1986)
" Accentuation rules for Japanese Text-to-Speech conversion", Review of the Electrical Communication Laboratories, Vol. 32, 188-199.

Scherer, K.R., (1981),
"Speech and emotional states", in the EVALUATION OF SPEECH IN PSYCHIATRY AND MEDICINE, ed. by J. Darby, (Grube and Stratton, \New York).
Sagisaka

Silverman

Simada, Z. and Hirose, H., (1971),
"Physiological correlates of Japanese accent patterns", Ann. Bull. RILP, No 5, 41-49.

Taylor, P.,A., (1992)

A PHONETIC MODEL OF ENGLISH INTONATION, Ph. D. Thesis, Edinburgh.

Vaissiere, J, (1985),
" The use of prosodic parameters in automatic speech recognition', in COMPUTER, SPEECH AND LANGUAGE, ed. by F. Fallside and W, Woods, Prentice Hall International.

Vaissiere, J. (1971),
Contribution a la synthese par regles du francais, These de troisieme cycle, Grenoble.

Vaissiere, J., (1974),
"On French prosody", Res. Lab of Electronics., MIT, Quarterly Progress Report 114, 212-223.

Vaissiere, J., (1975),
"Further note on French prosody", Res. Lab of Electronics., MIT, Quarterly Progress Report 115, 251-162.

Vaissiere, J., (1982)
"Language-independent prosodic features"
in PROSODY, MODELS AND MEASUREMENTS.
edited by Cutler, A., and Ladd, D.R.,
72-86.

Vaissiere, J., (1992)
"Rhythm, accentuation and final lengthening in French", in Music, Language and brain.

ANNEX :Fo detection

Except for the second speakers, the data base is far of being perfect. In many cases the microphone was to close of the speaker mouth, resulting in a burst of energy following the release of the labials (sab) There are large changes of recording levels form one sentence to the next, and also inside a sentence.

The use an accelerometer, placed on the throat of the speaker, may be necessary for recording the data base which are going to be labeled. Fo can be later tracked from either the accelerometer signal or from the speech signal.

The main errors are halfing the Fo values in female voice, doubling or halfing of the values for male voices. The figures on the next pages illustrates such errors.

CRICOTHYROID ACTIVITY IN AN ENGLISH SENTENCE

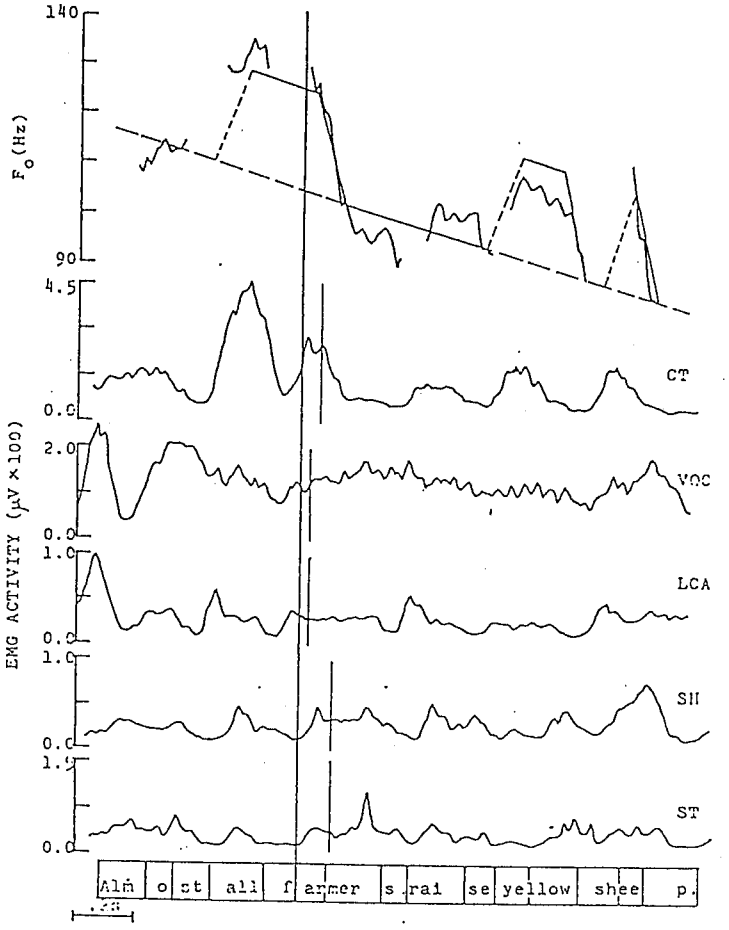


Fig. 3.6 (c) KS S60

Fig. I.1

CRICOTHYROID ACTIVITY IN DIFFERENT PITCH ACCENT POSITIONS

Japanese speaker

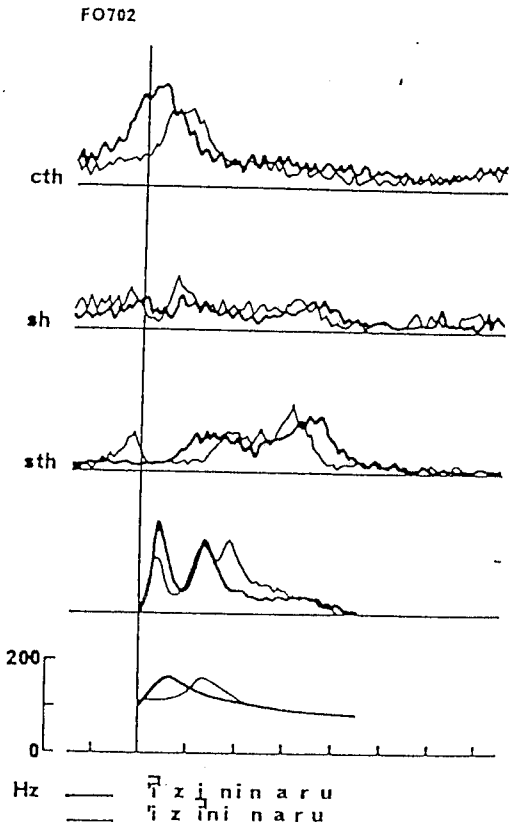


Fig. I.2

DECLINATION TENDENCY

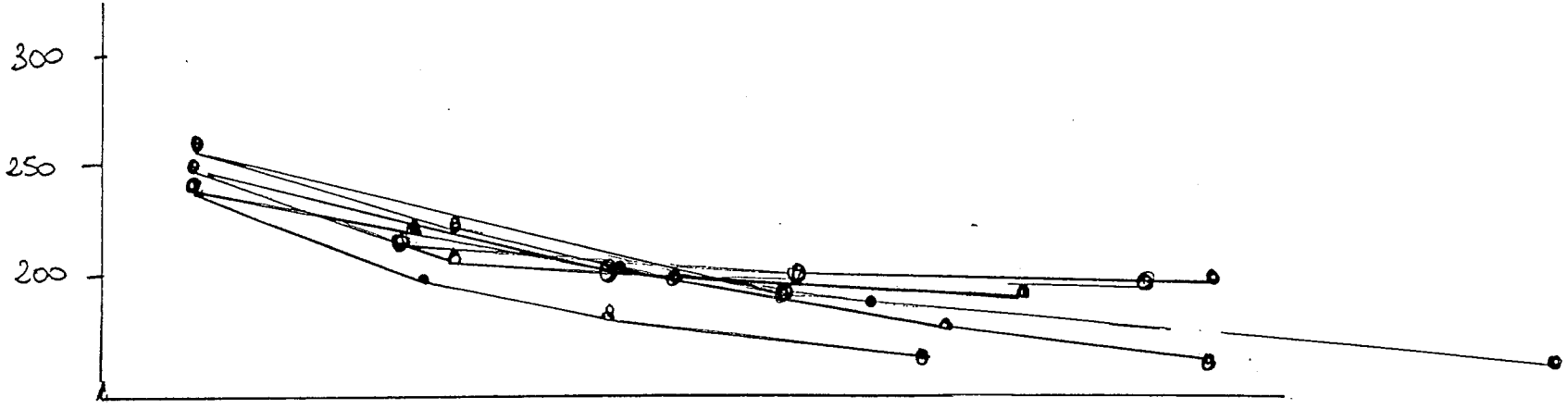
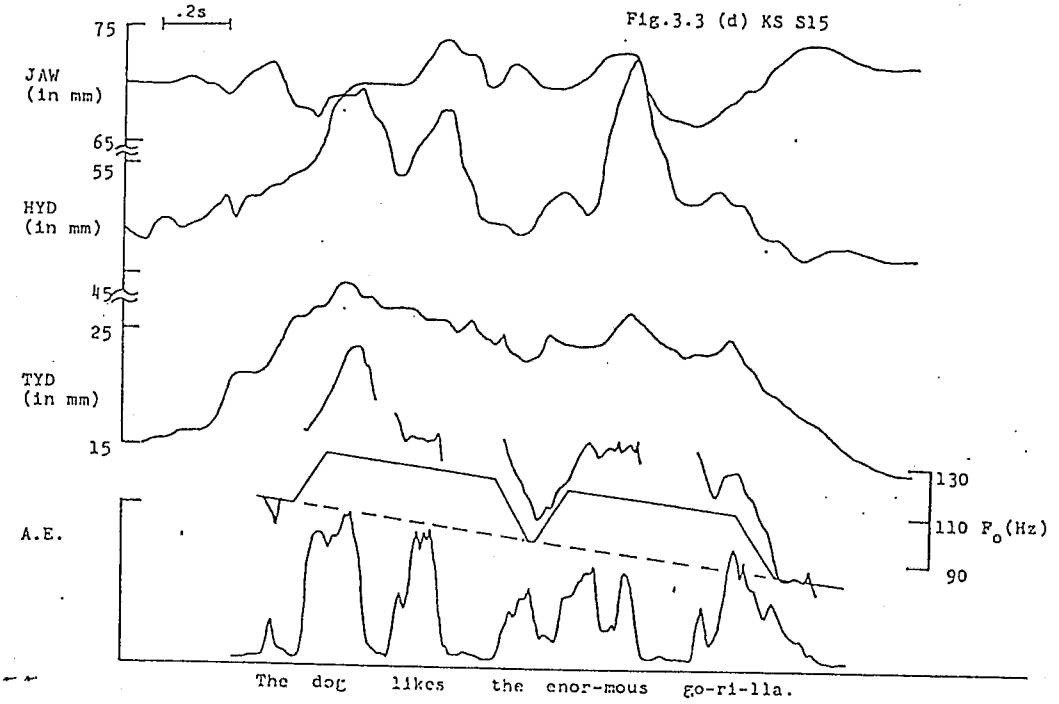


Fig. I.3

LARYNX HEIGHT AND FO CONTOURS



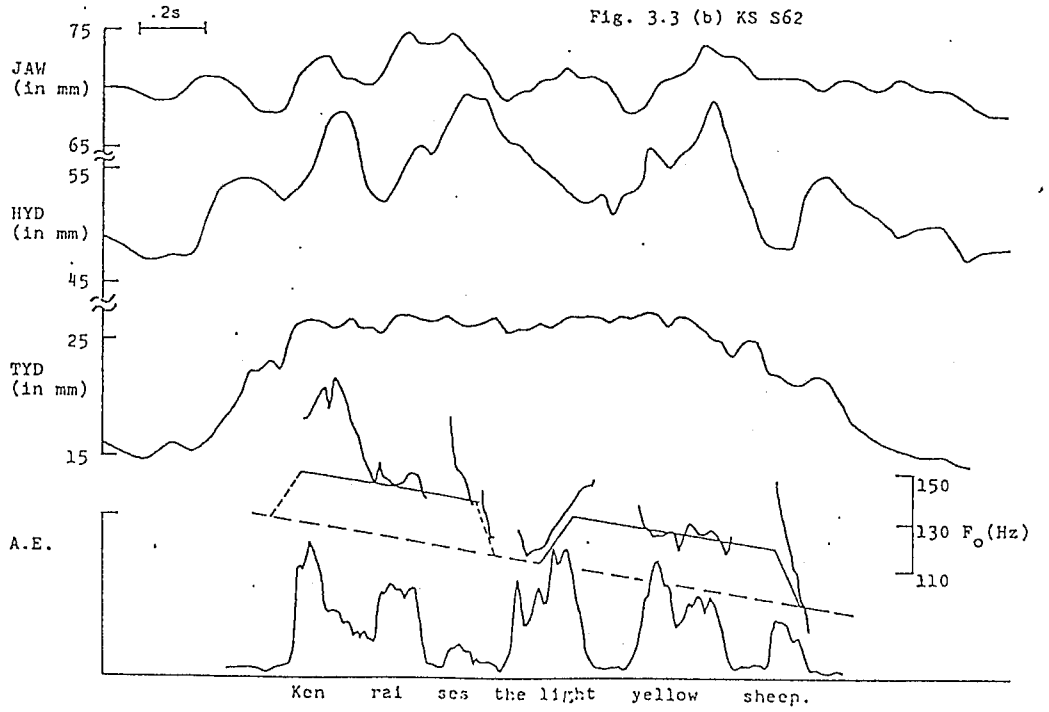


Fig. I.4

VENTRICLES LENGTH AND FO CONTOURS

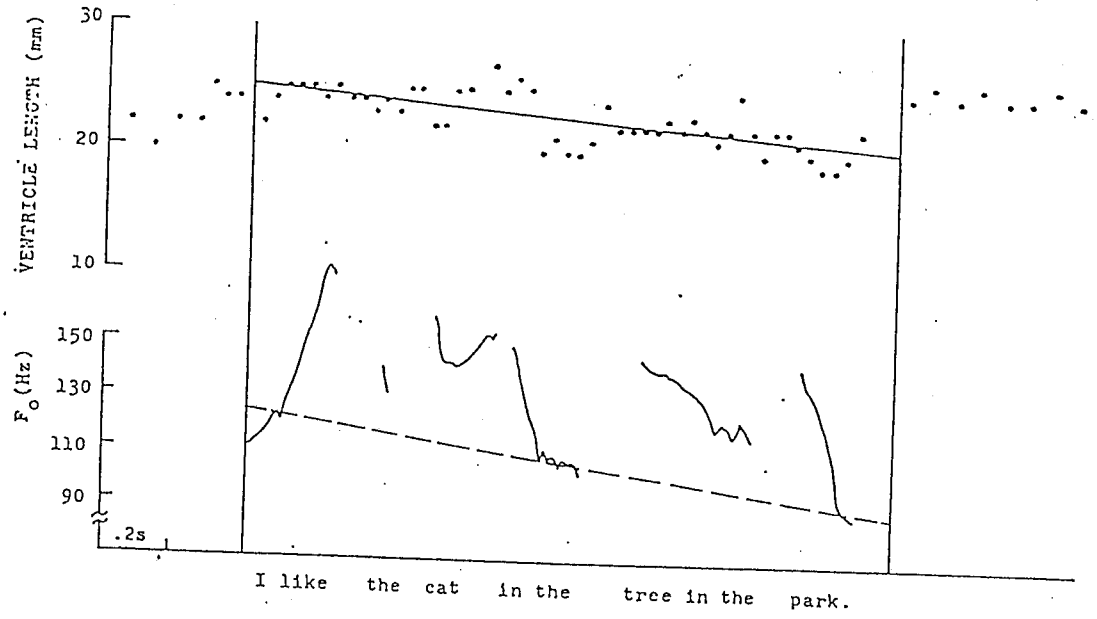


Fig. 3.4 (c) KS S77

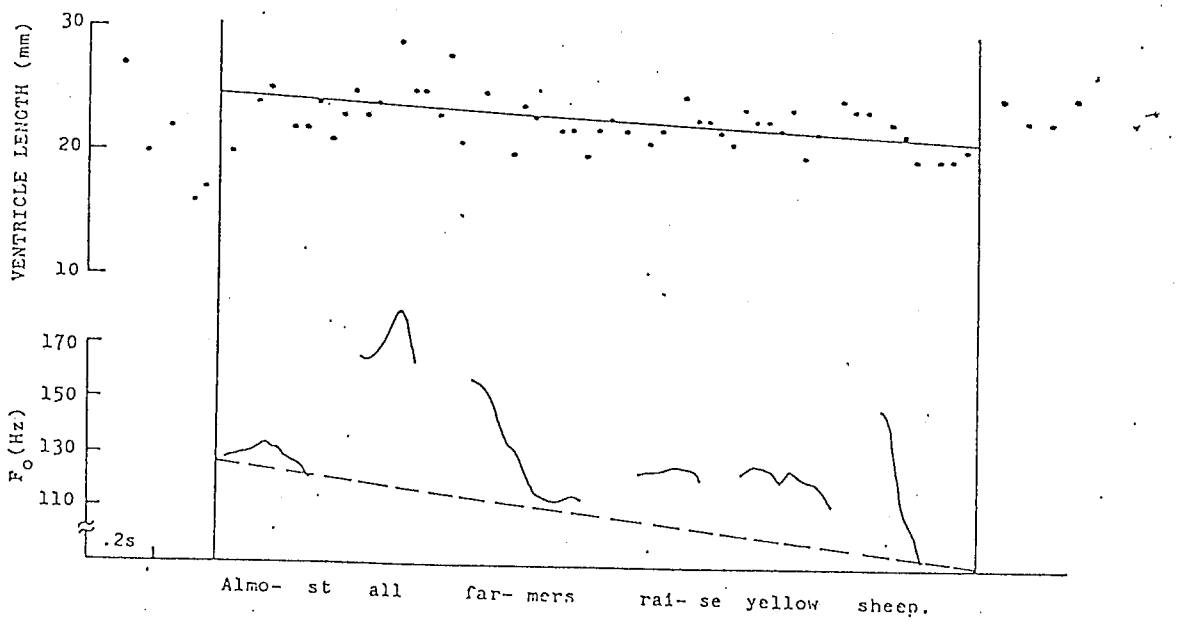
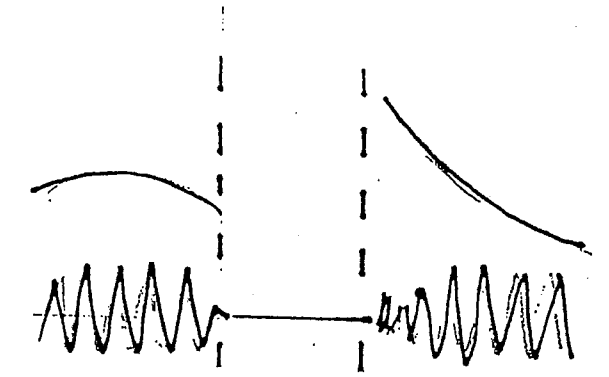


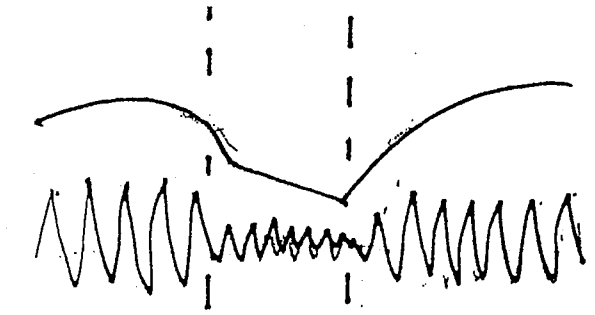
Fig. 3.4 (a) KS S60

Fig. 1.5

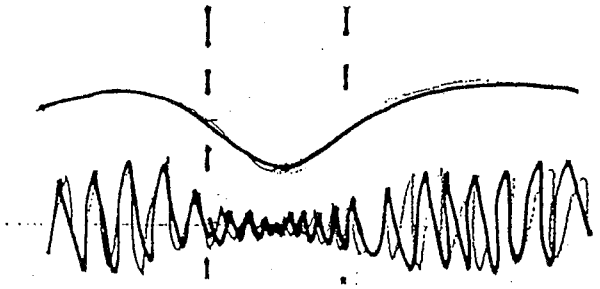
MICROMELODIC FLUCTUATIONS



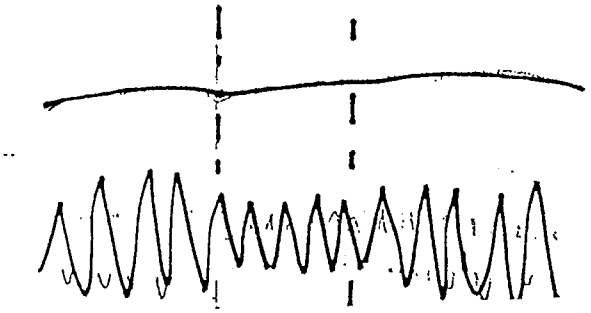
1) unvoiced stop



2) voiced stop



3) voiced fricative



4) sonorant

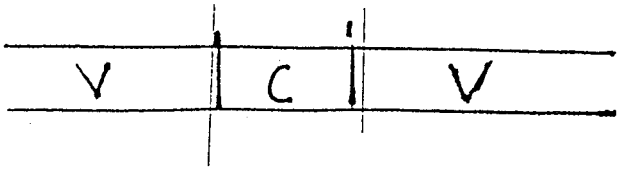


Fig. I.6

CRICOTHYRID ACTIVITY
IN VOICED AND UNVOICED STOPS
English speaker

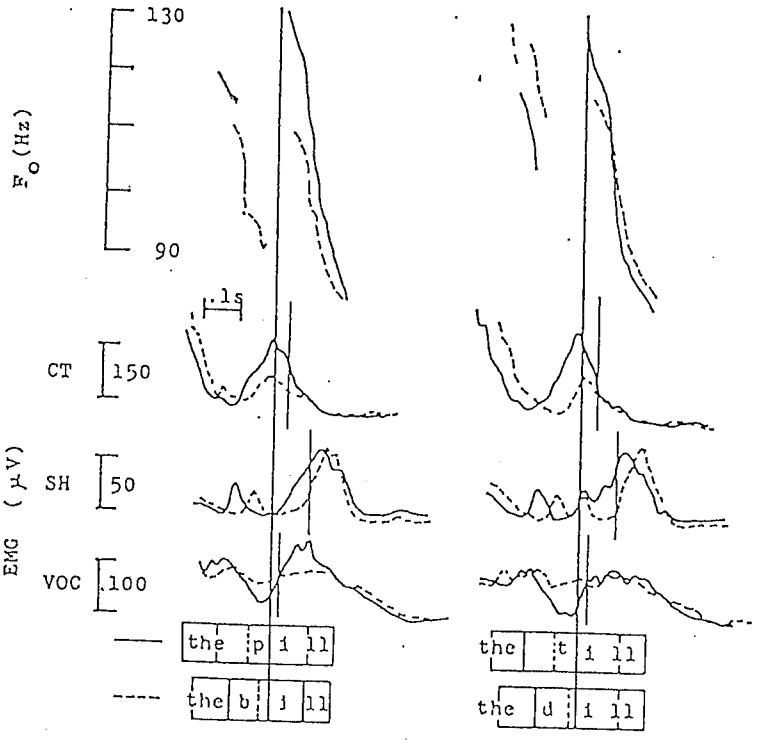


Fig. 3.10

Fig. I.7

VOICED/UNVOICED
AND TENSE/LAX CONTRAST

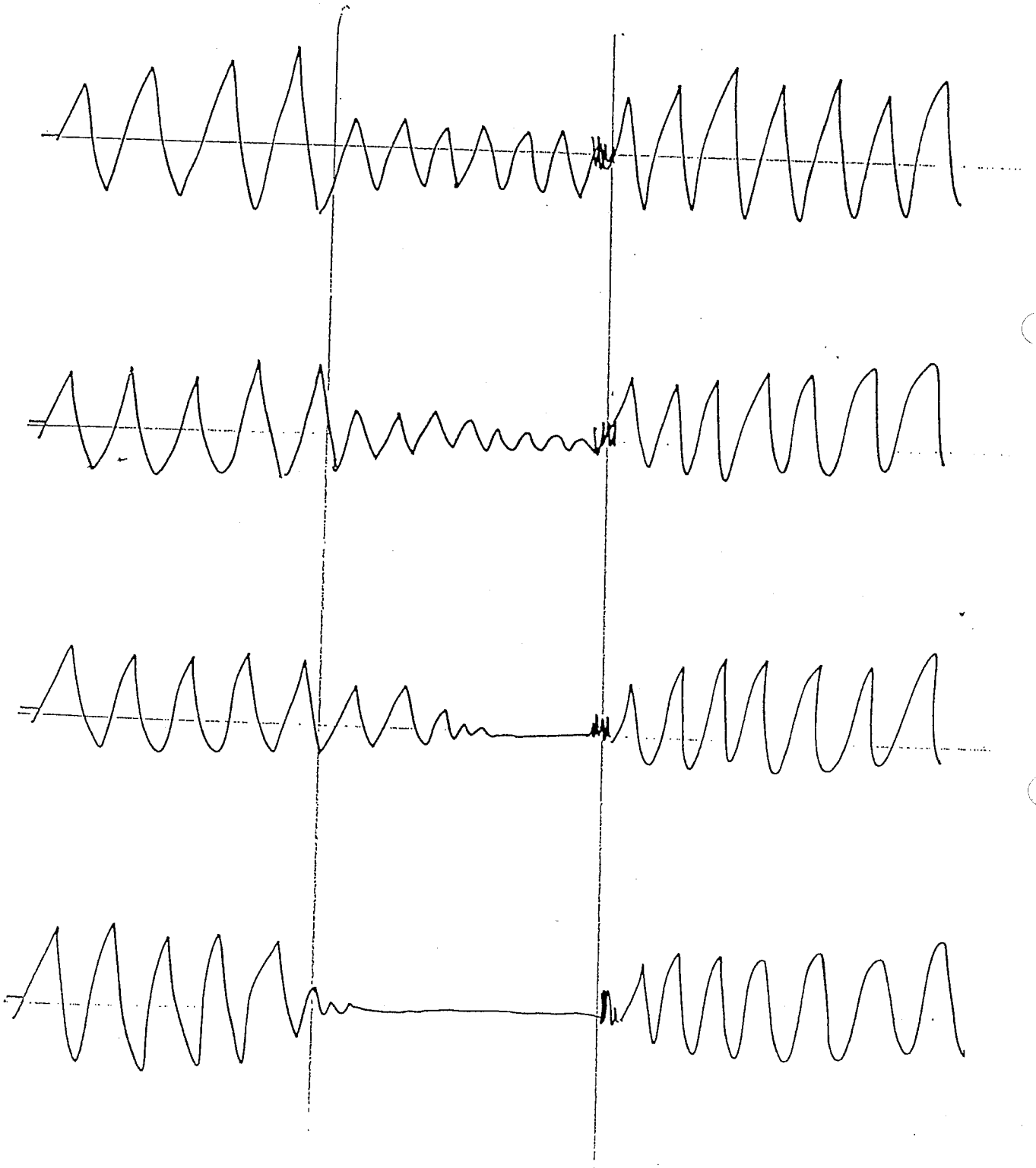
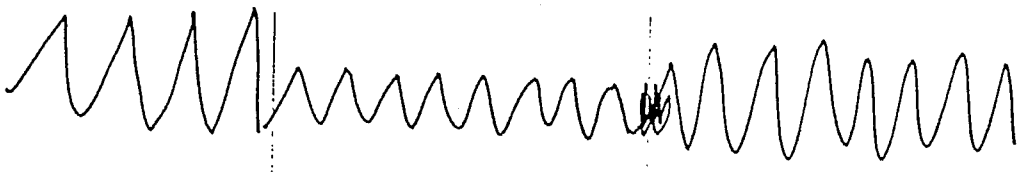
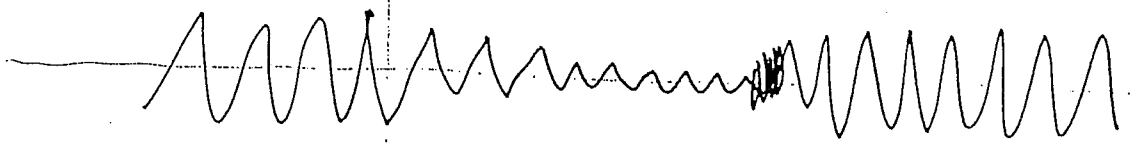


Fig. I.8

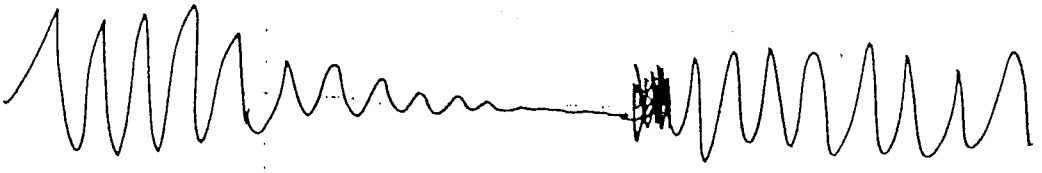
THE EFFECT OF PLACE OF ARTICULATION ON VOICING



labial



dental



velar

Fig. I.9

DIFFERENT MANEUVERS FOR MAINTAINING VOICING

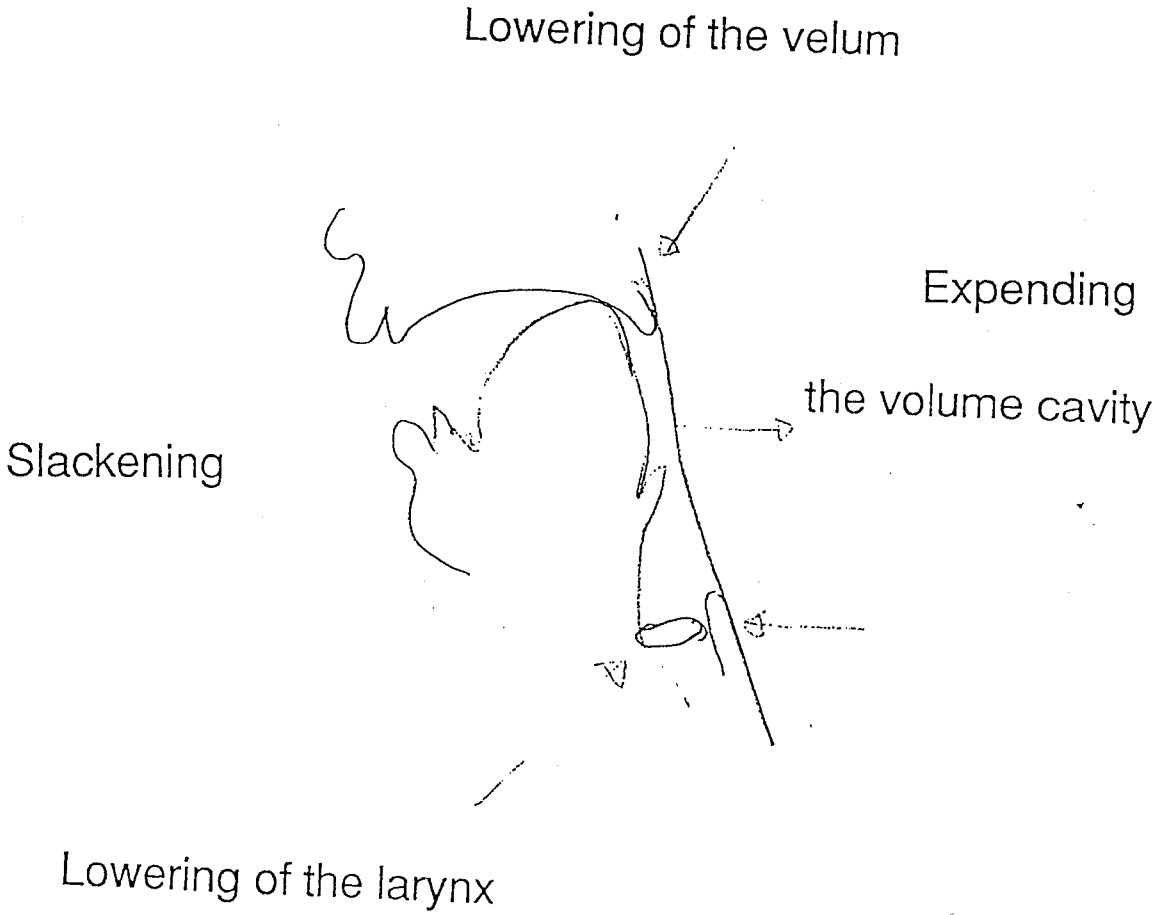
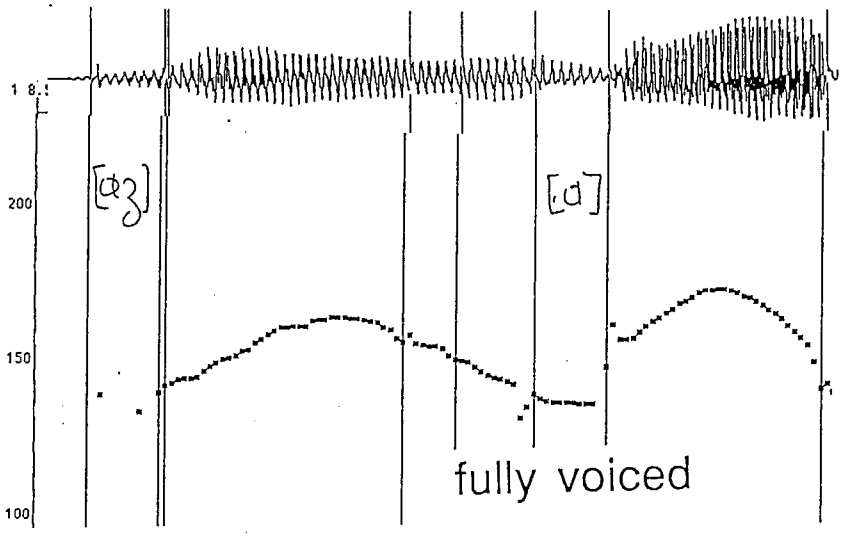


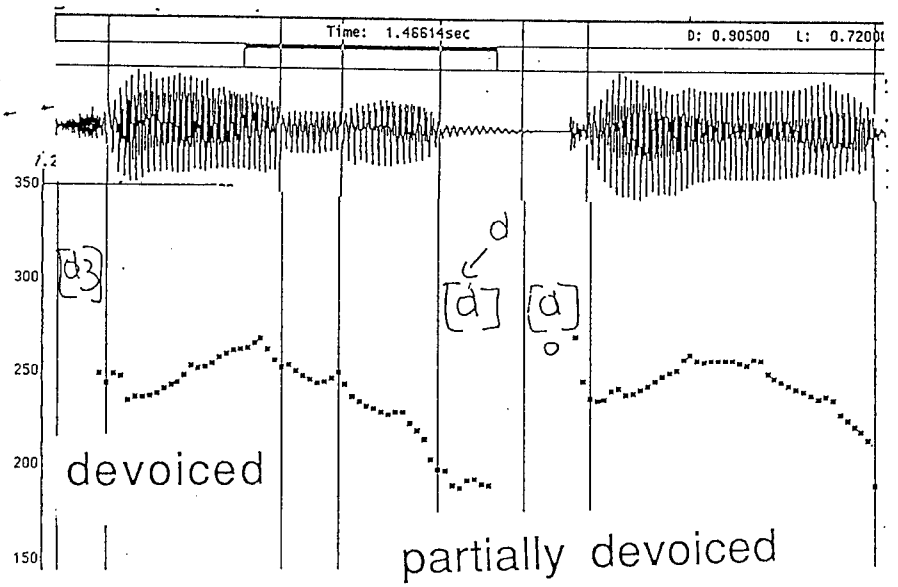
Fig. I.10

INTERPSPEAKER DIFFERENCES

SPEAKER DIFFERENCES IN VOICING



Jane adores



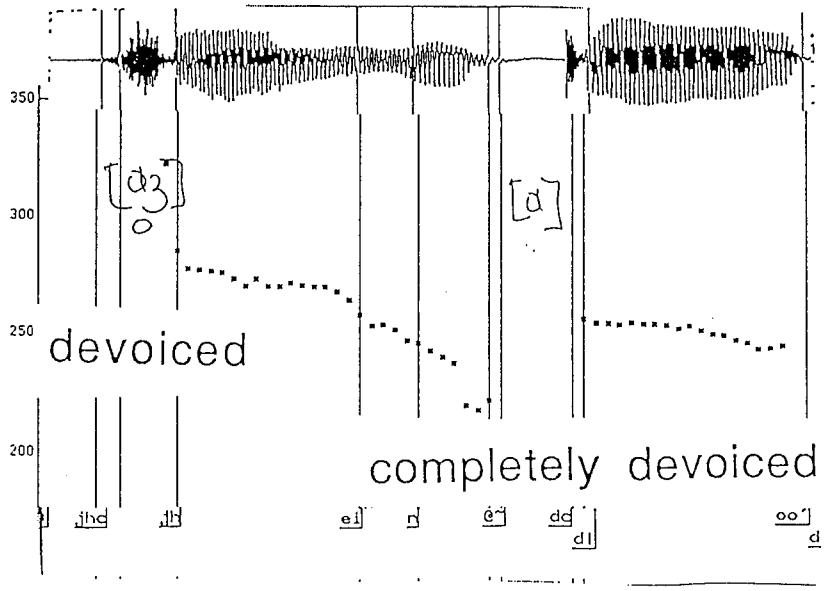
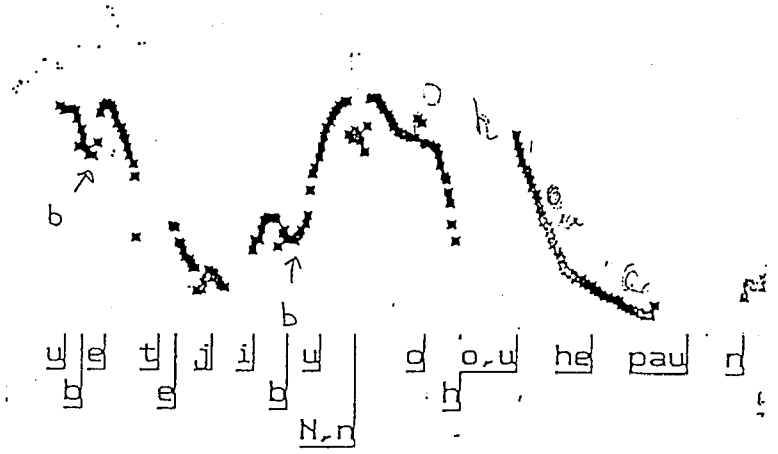
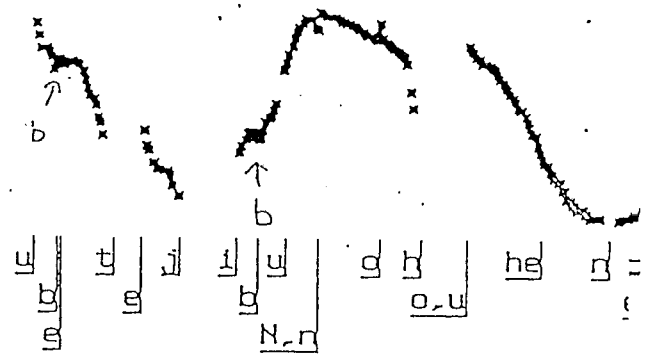


Fig. I.11

INTERSPEAKER VARIATIONS:

Fo drop in voiced stops





A phrase spoken by Japanese speakers

Fig. I.12

MICRO Fo FLUCTUATION IN VOICED STOPS AND VOICED FRICATIVES

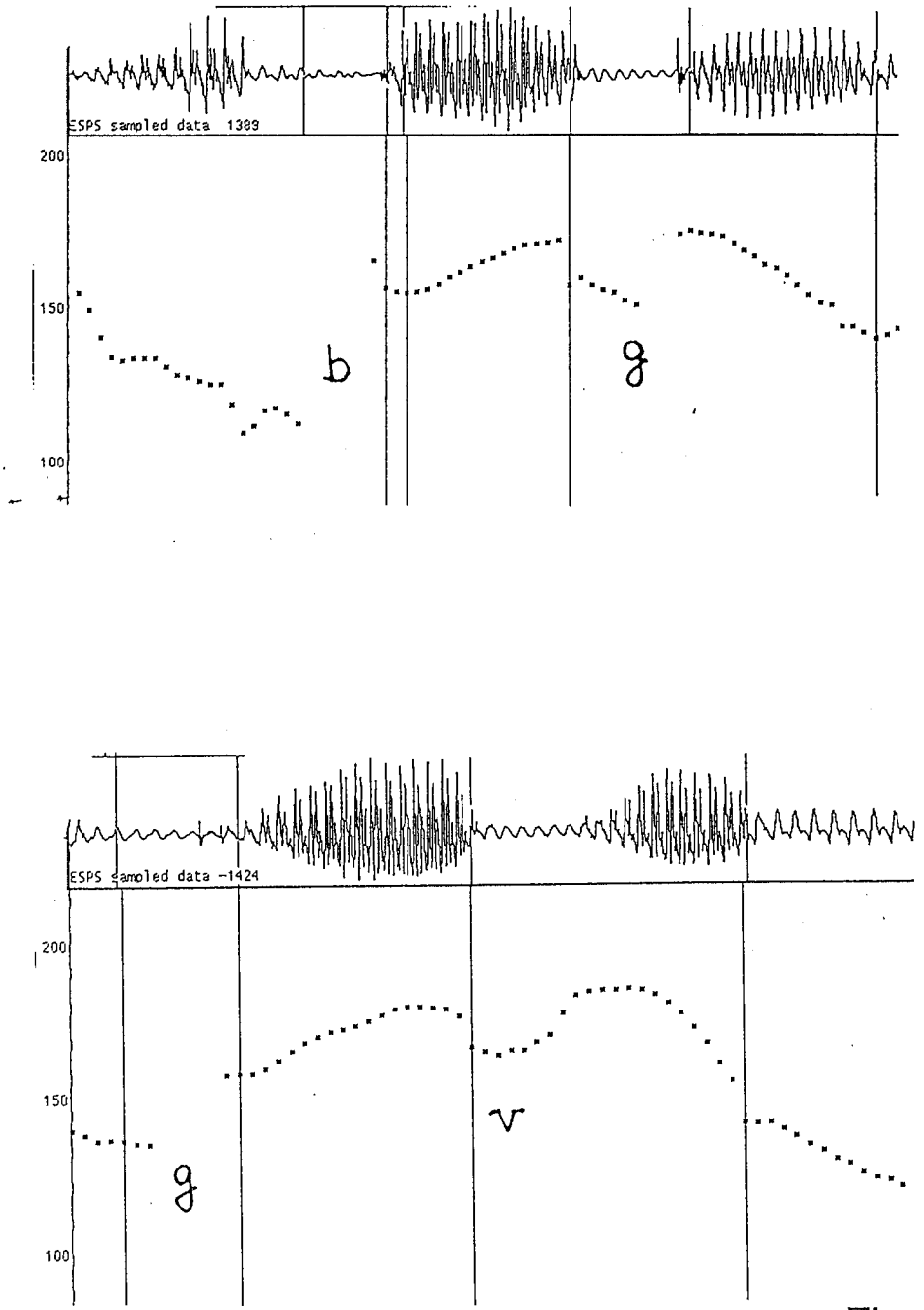
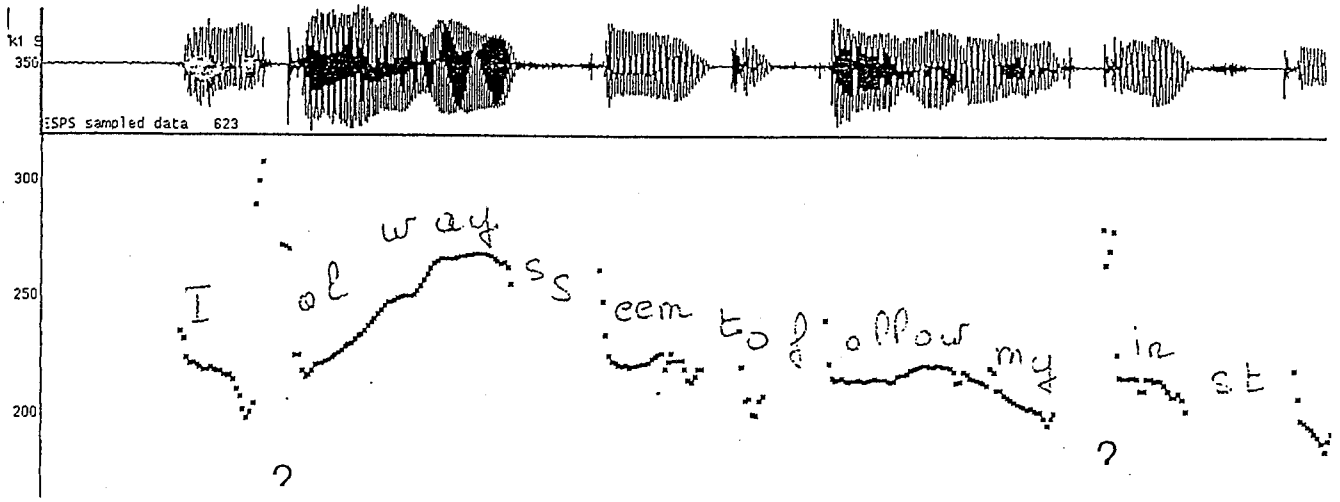
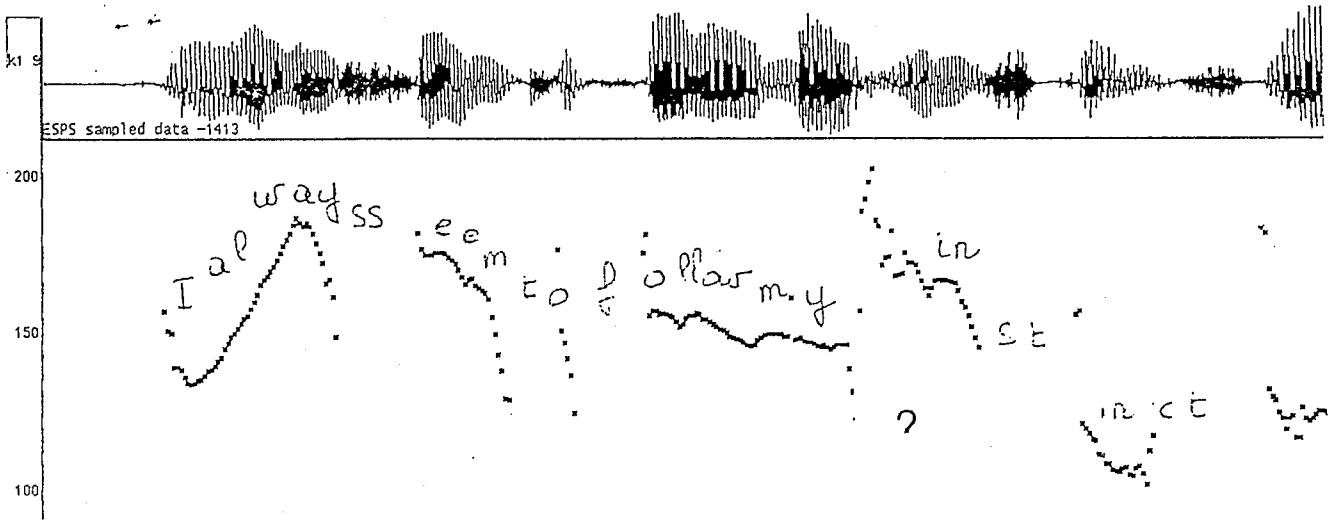


Fig. I.13

INTER-SPEAKER VARIATIONS
OPTIONAL GLOTTAL STOP INSERTION
BEFORE WORD STARTING BY A VOWEL



(I always seem to follow my instinct)



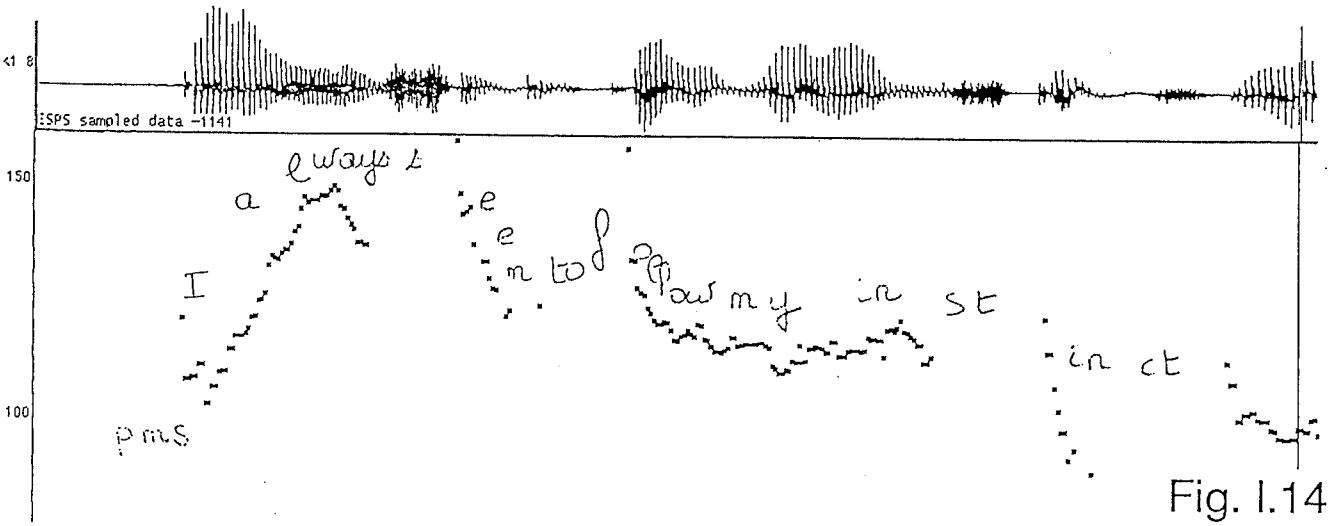


Fig. I.14

SPEAKER VARIATIONS IN VOICING

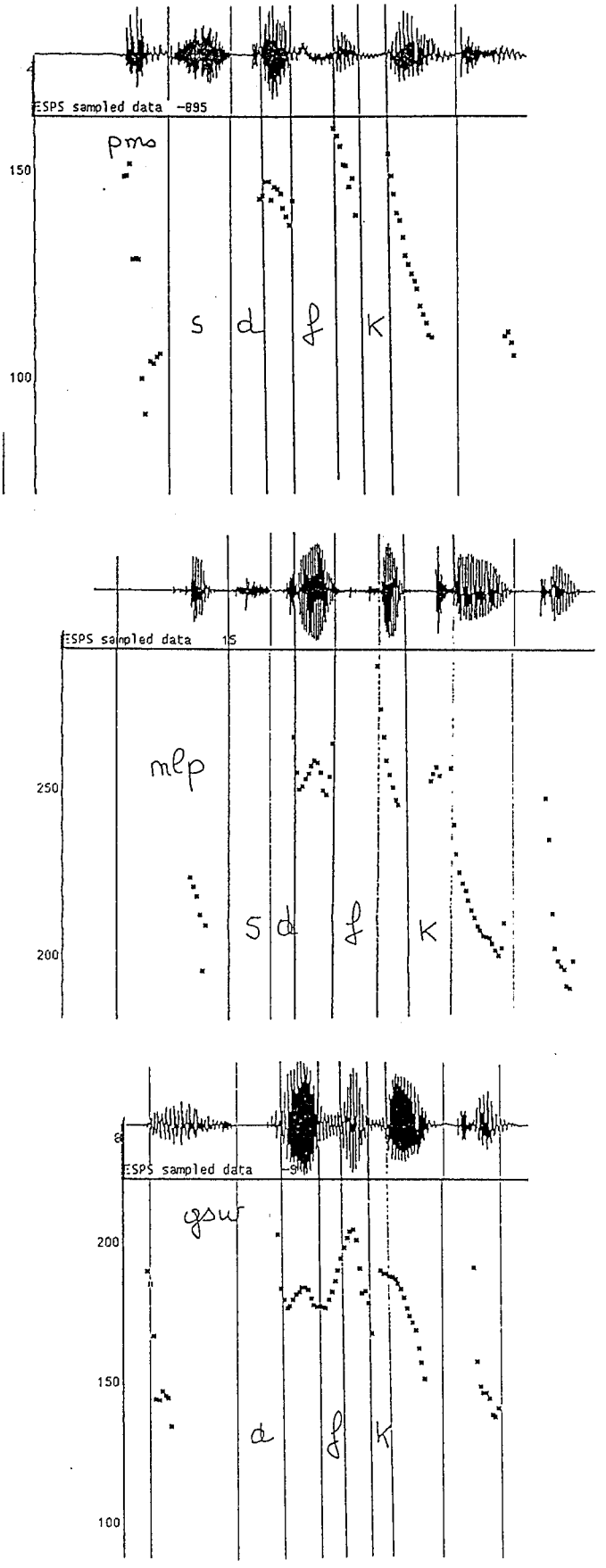


Fig. I.15

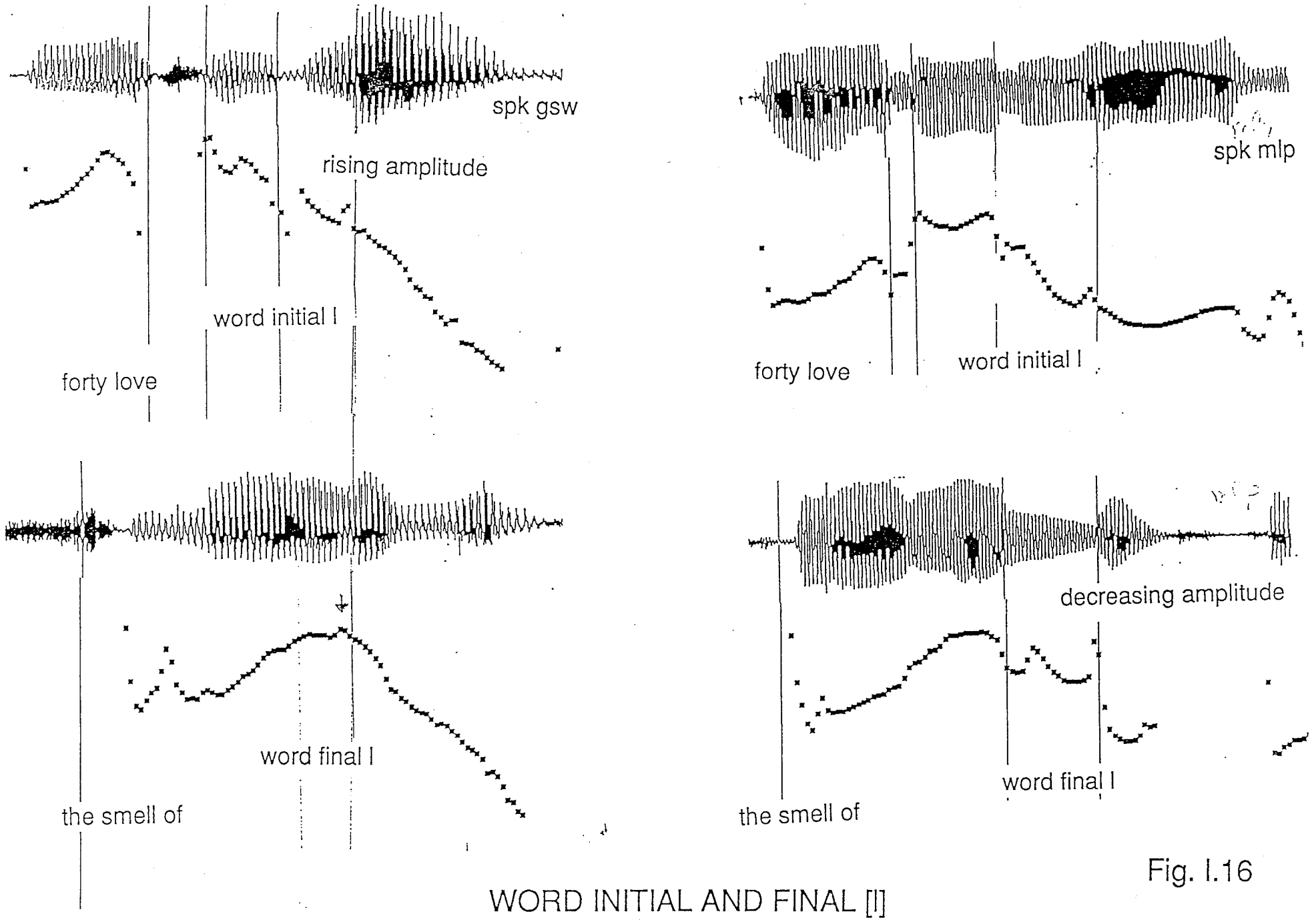


Fig. I.16

WORD INITIAL AND FINAL [l]

THE BASIC HAT PATTERN

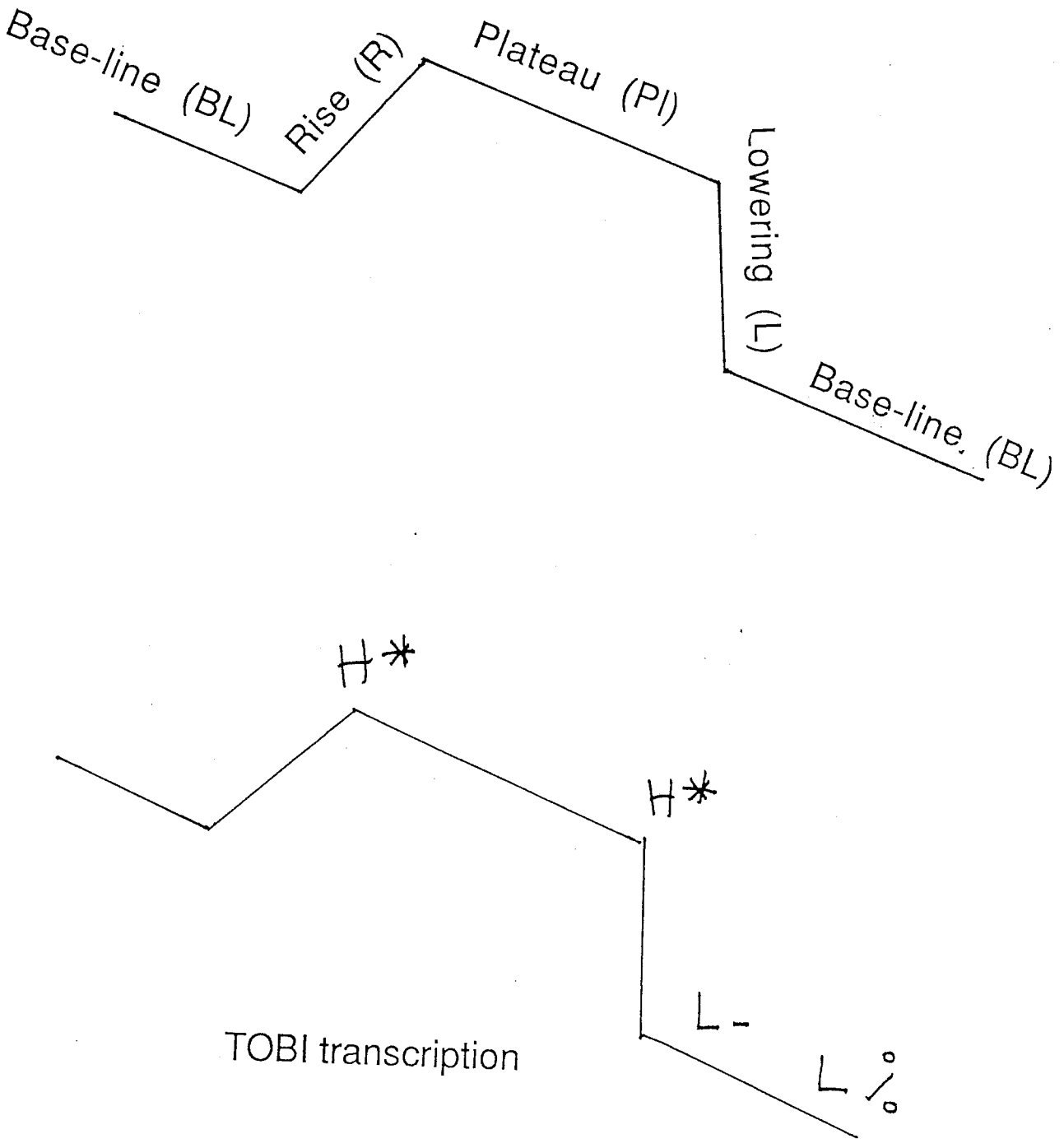


Fig. II.1

LANGUAGE-INDEPENDENT PROSODIC TENDENCIES

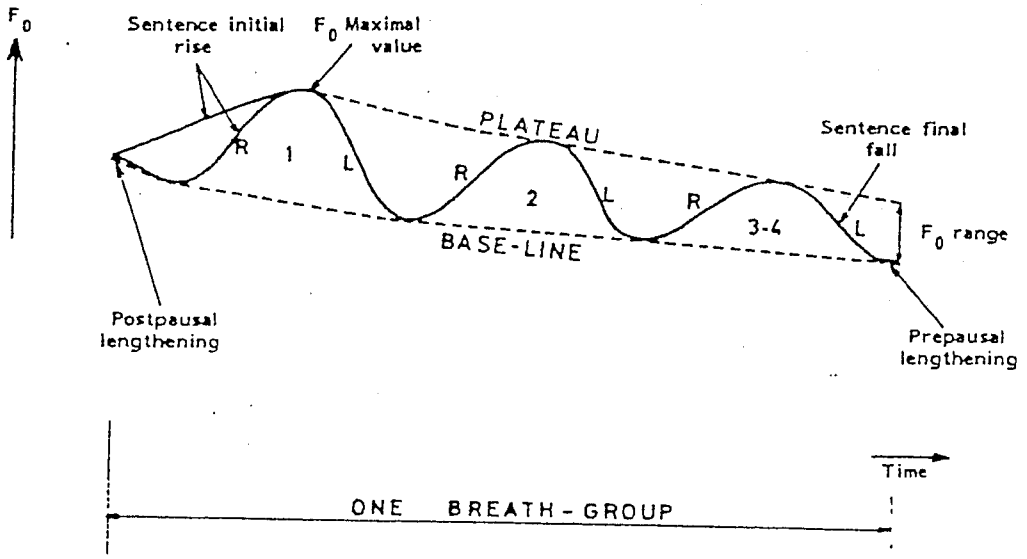


Fig. 5.1. General properties of F_0 contours observed in unmarked sentences in a number of languages. The common tendencies are the following: (i) a tendency for F_0 values to fluctuate between two abstract lines: the *plateau* and the *baseline*, which delimit the speaker's F_0 range; (ii) a tendency for the F_0 *range* to diminish as a function of time; (iii) a tendency to start sentences with a large *sentence-initial rise* in F_0, located on one of the first syllables, or spread over the first few syllables; (iv) a tendency to repeat a succession of F_0 rises (R) and lowerings (L): a pair of opposing movements indicates a prosodic word (see text); (v) a tendency for the *maximal value* of F_0 to be located on the first prosodic word of the sentence; (vi) a tendency to lengthen the duration of the last syllable at the end of the breath-group (*prepausal lengthening*), and of the first phoneme at the beginning of the sentence (*postpausal lengthening*)

LANGUAGE-INDEPENDENT PROSODIC TENDENCIES: DIVISION INTO SENSE-GROUPS

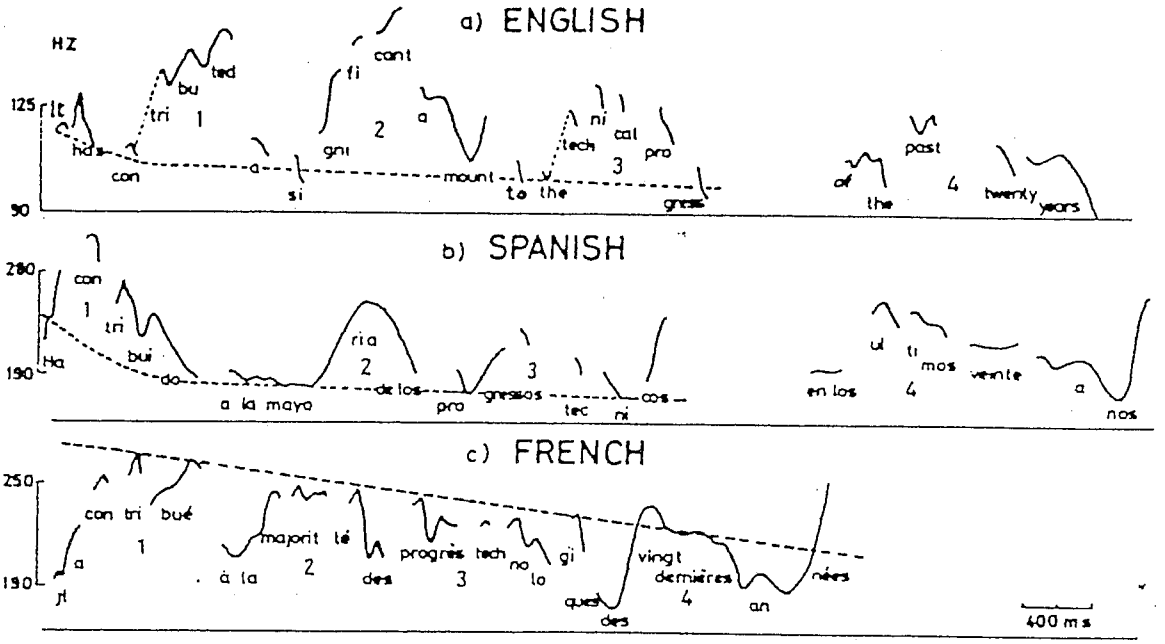


Fig. 5.2a–c. F_0 contours for the same clause translated into three languages and pronounced by native speakers: (**a**) "It has contributed a significant amount to the technical progress of the past twenty years, . . ."; (**b**) "Ha contribuido a la mayoria de los progressos tecnicos en los ultimos veinte anos, . . ."; (**c**) "Il a contribué à la majorité des progrès technologiques des vingt dernières années, . . .". The pauses and the rise-fall pairs in F_0 combine to divide the three clauses into four prosodic words (numbered 1 to 4)

Fig. II.3

GROUPING OF HAT PATTERNS

A) Increasing boundary strength

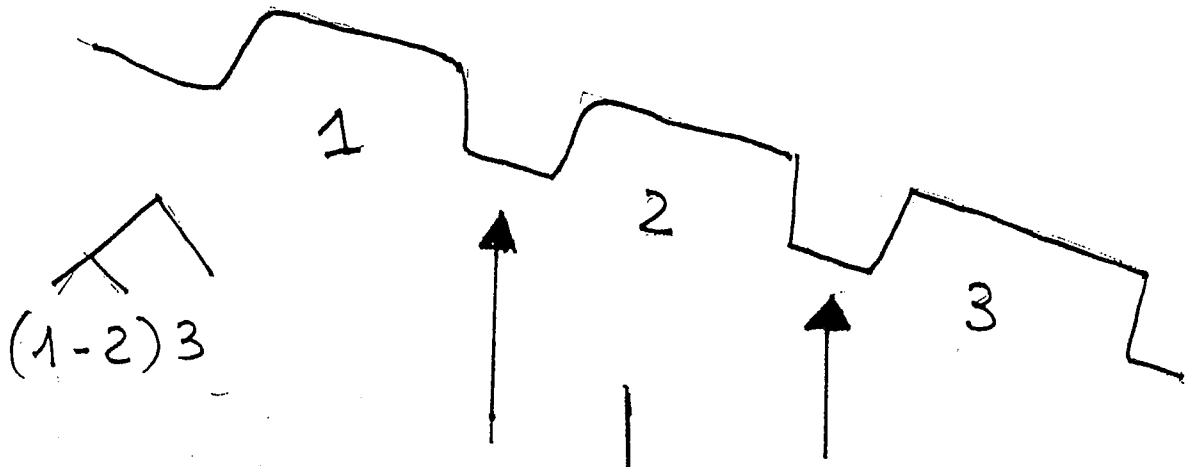
1) Inserting a *pause*
2) *Resetting* of the baseline
3) *Leveling* of the baseline
4) *Extra-low* and *continuation rise*
5) Extra phrase final *lengthening*
6) Increasing the magnitude of *Rise* after boundary

B) Decreasing the boundary strength

7) Decreasing the *fall-rise pattern*
8) Decreasing word final *lengthening*
9) Decreasing the height of the *plateau*
10) *Grouping* into a single HP

Fig. II.4

GROUPING OF HAT PATTERNS

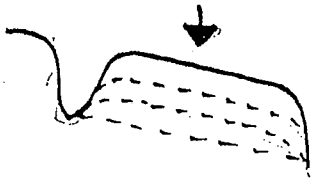


B) Decreasing the boundary

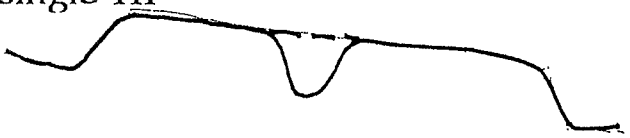
A) Increasing the boundary

7) Decreasing the fall-rise pattern

9) Decreasing the height of the plateau



10) Grouping HP1 and HP2 into a single HP



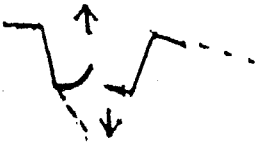
8) Decreasing word final lengthening

1) Inserting a pause

2) Resetting of the baseline

3) leveling of the baseline

4) Extra-low and continuation ris



5) Extra phrase final lengthening

6) Increasing the magnitude of HP3 first Rise

Fig. II.5

SUBDIVISION OF THE HAT PATTERN

1) Extra Rise on the plateau on the final stressed syllable (Rp)

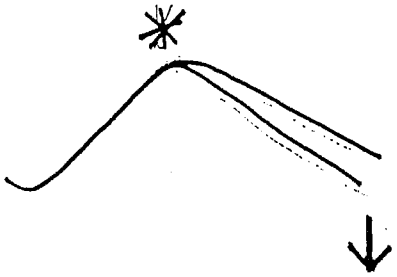
2) General lowering of the plateau between the two stressed syllables

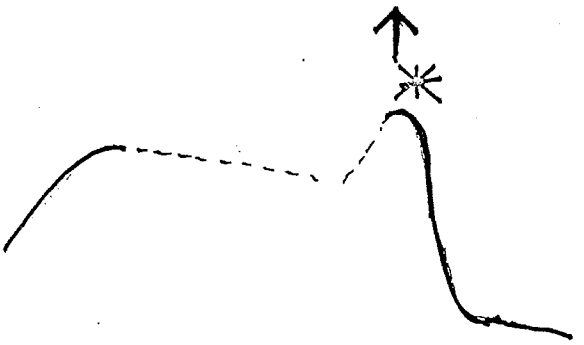
3) Lowering of the function words

4) First word final lengthening

Fig. II.6

SUBDIVISION OF THE HAT PATTERN





2) General lowering of the plateau

1) Extra Rise on the plateau

3) Lowering of the function words

4) First word final lengthening

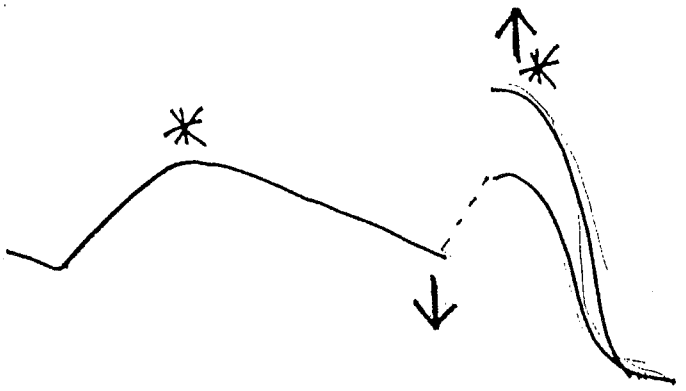


Fig. II.7

SUPERIMPOSITION FO MODEL

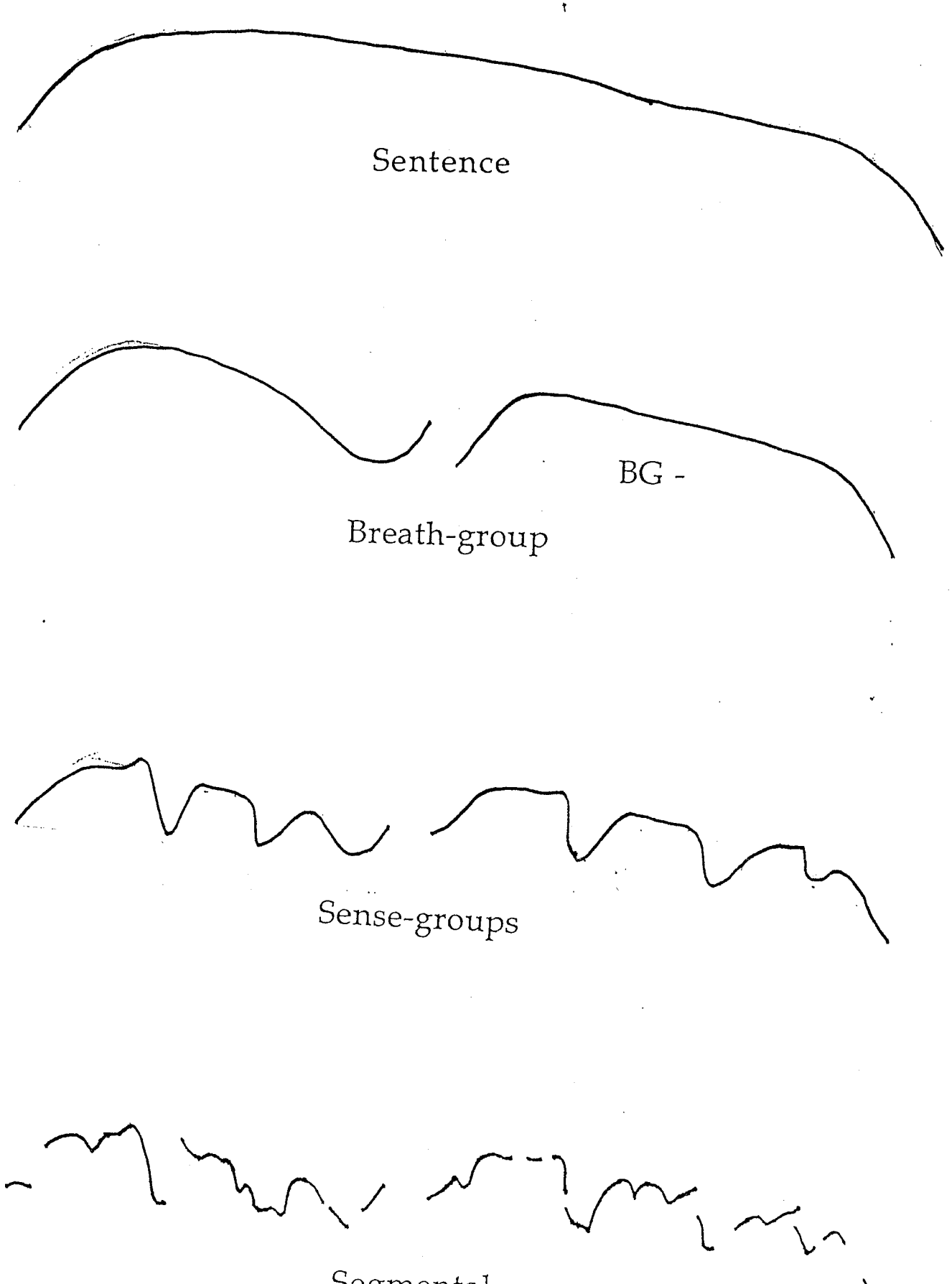
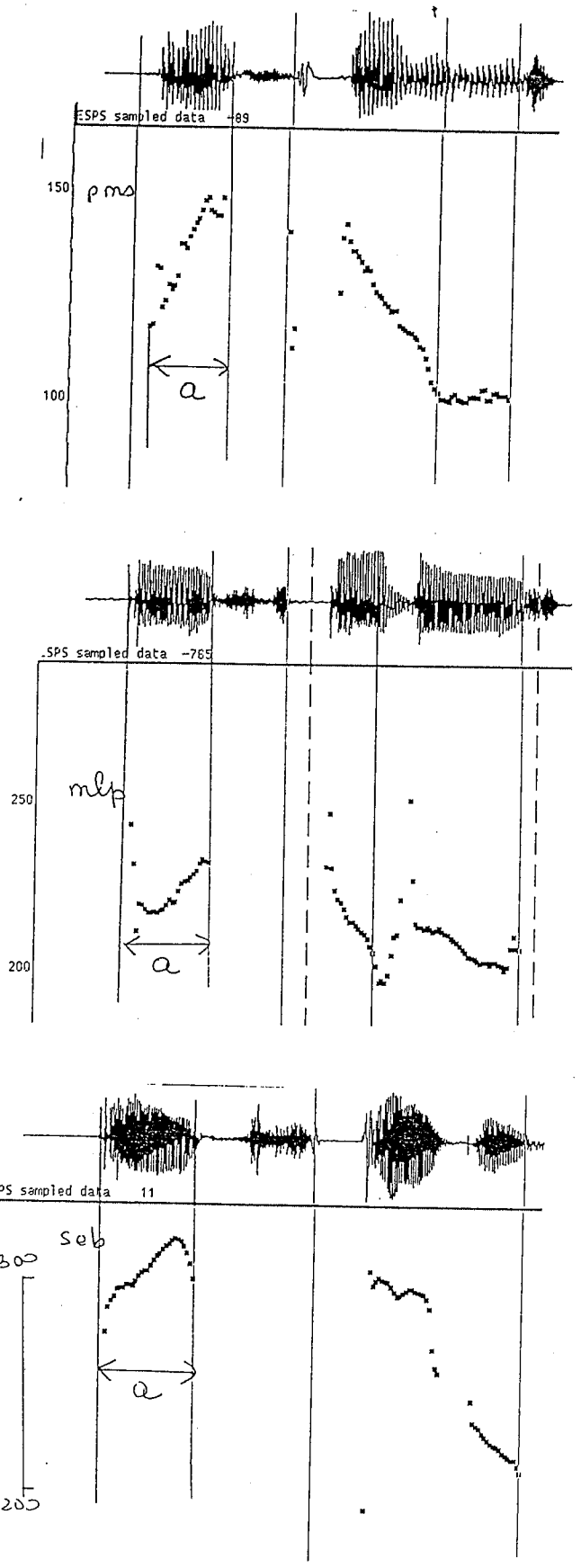


Fig. II.8

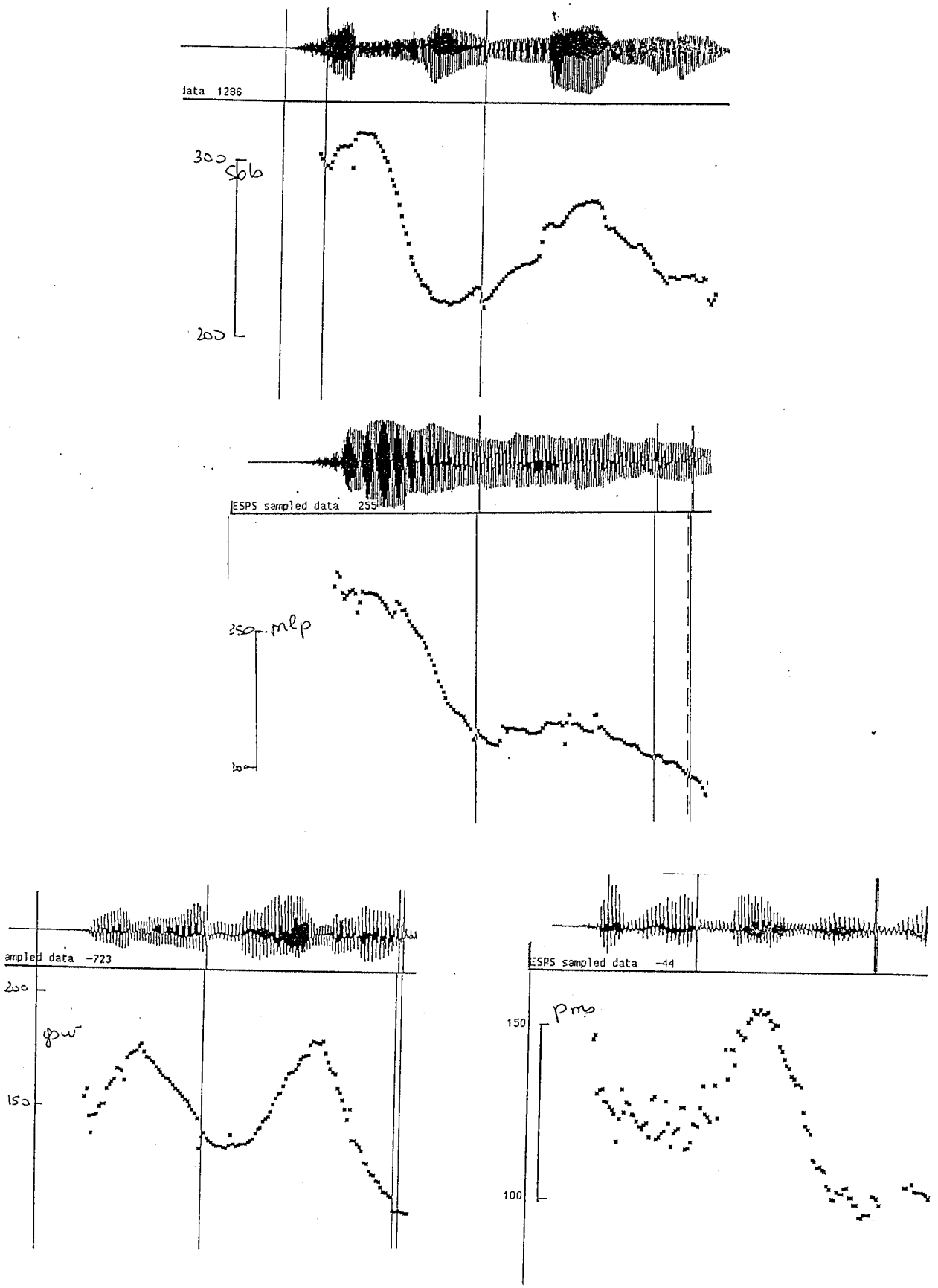
##[V*...



Alf's brother...

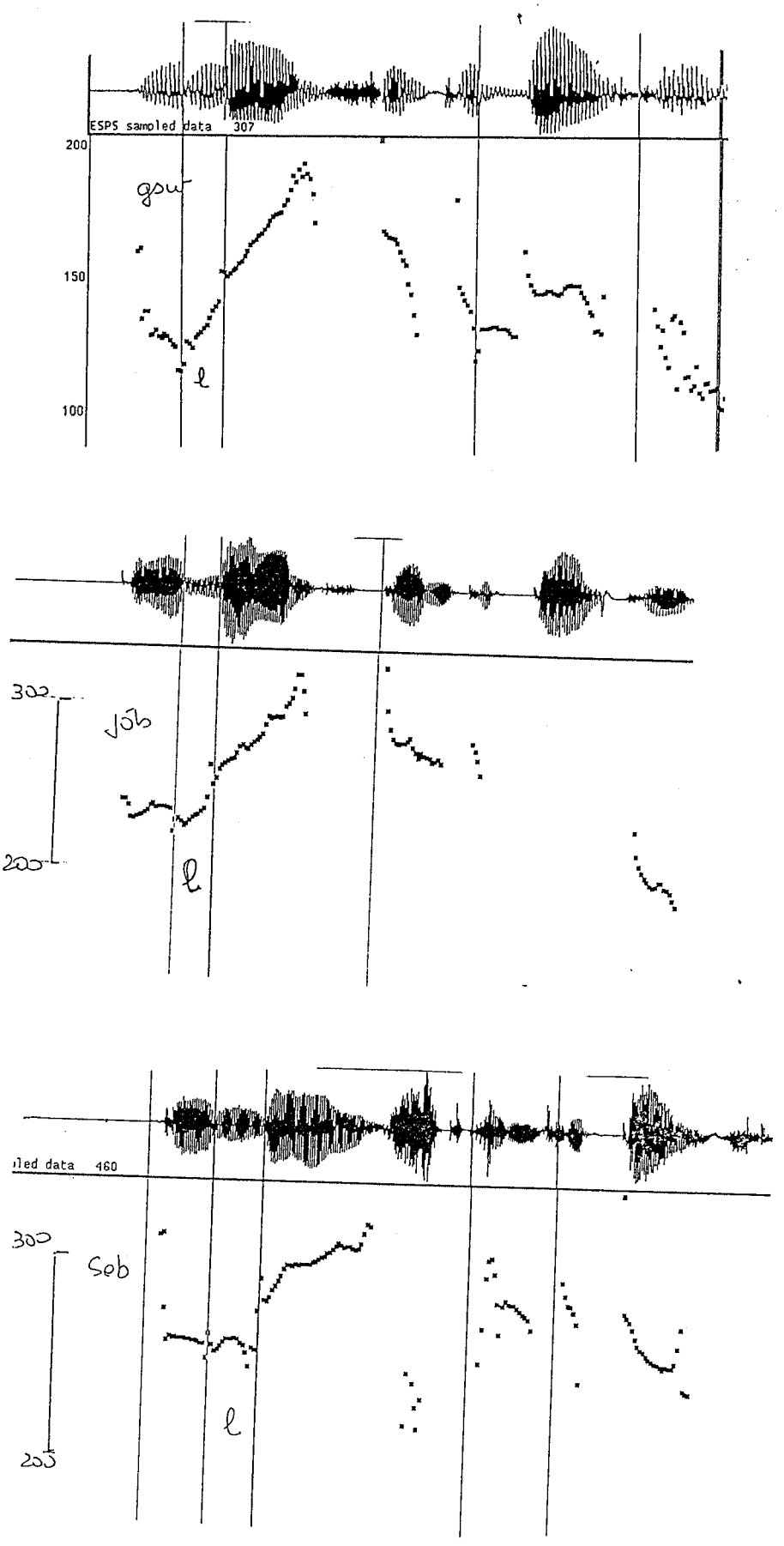
Fig. II.9

##[?V...



Henry normally...

Fig. II.10



They launched into battle...

Fig. II.11

RISE AND LOWERING

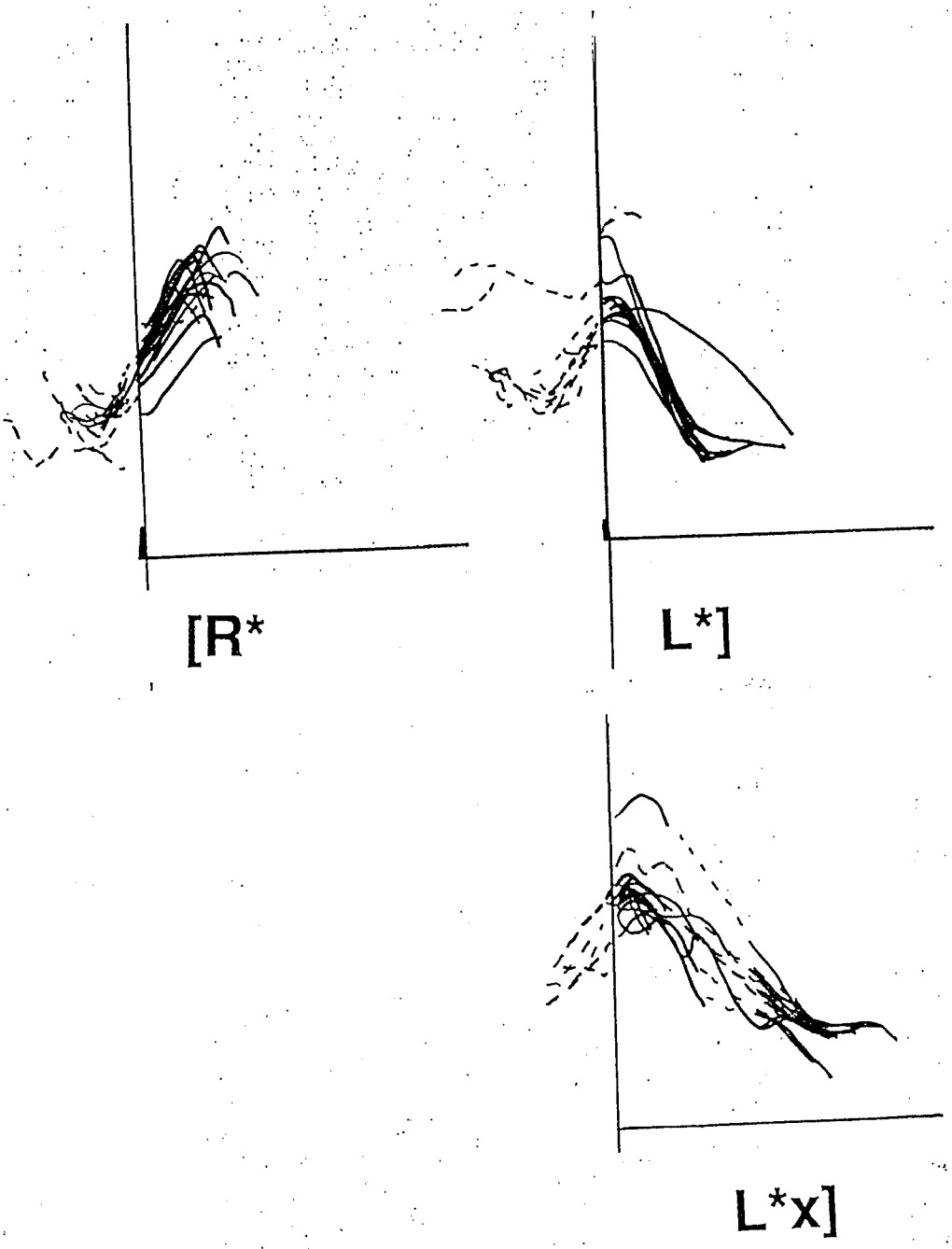


Fig. II.12

SPEAKER-DEPENDENT PATTERN
MONOSYLLABIC CONTENT WORD
WITH THE ATTRIBUTE L

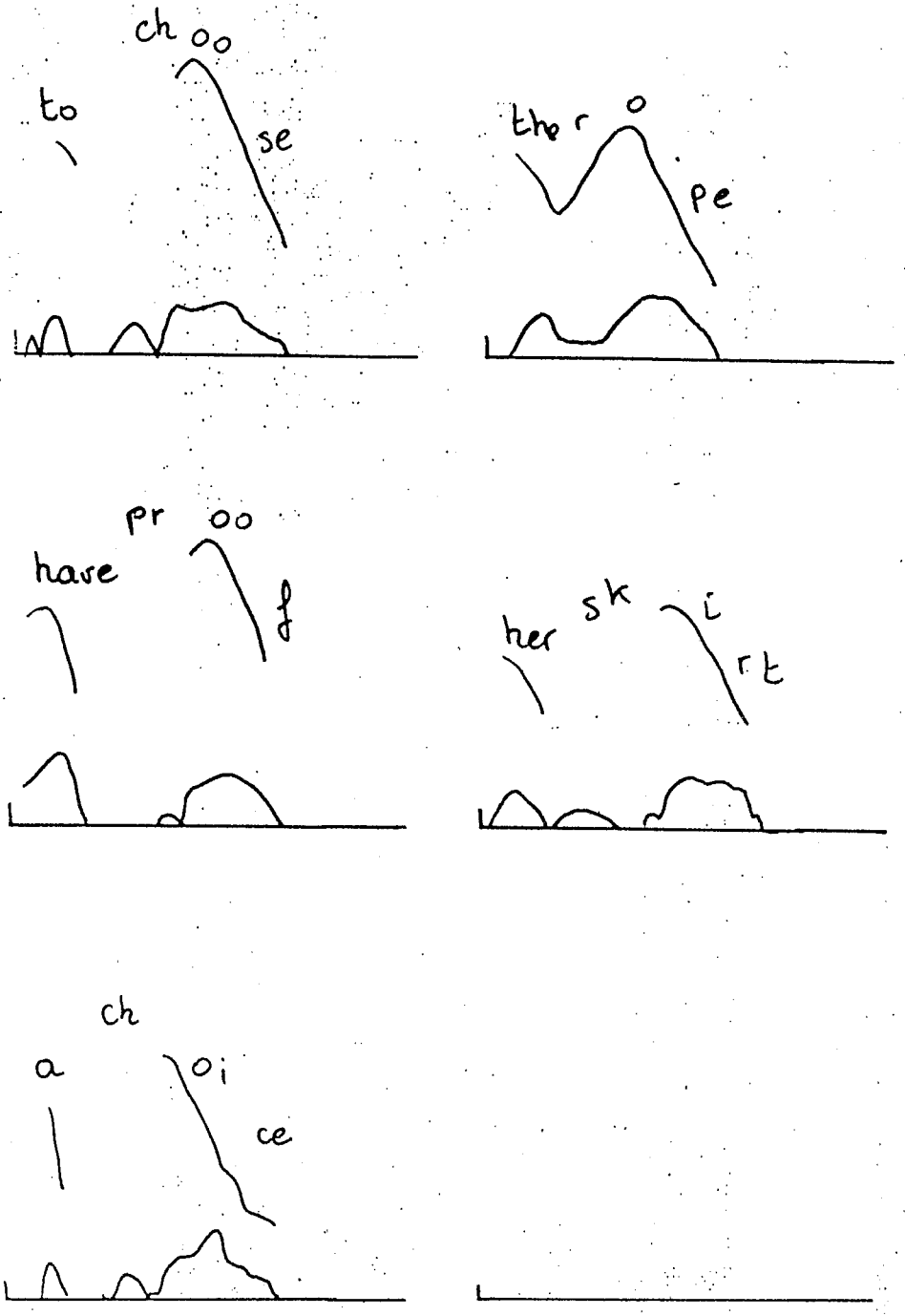


Fig. II.13

SPEAKER-DEPENDENT PATTERN
TWO-SYLLABES WORDS
WITH THE ATTRIBUTE L

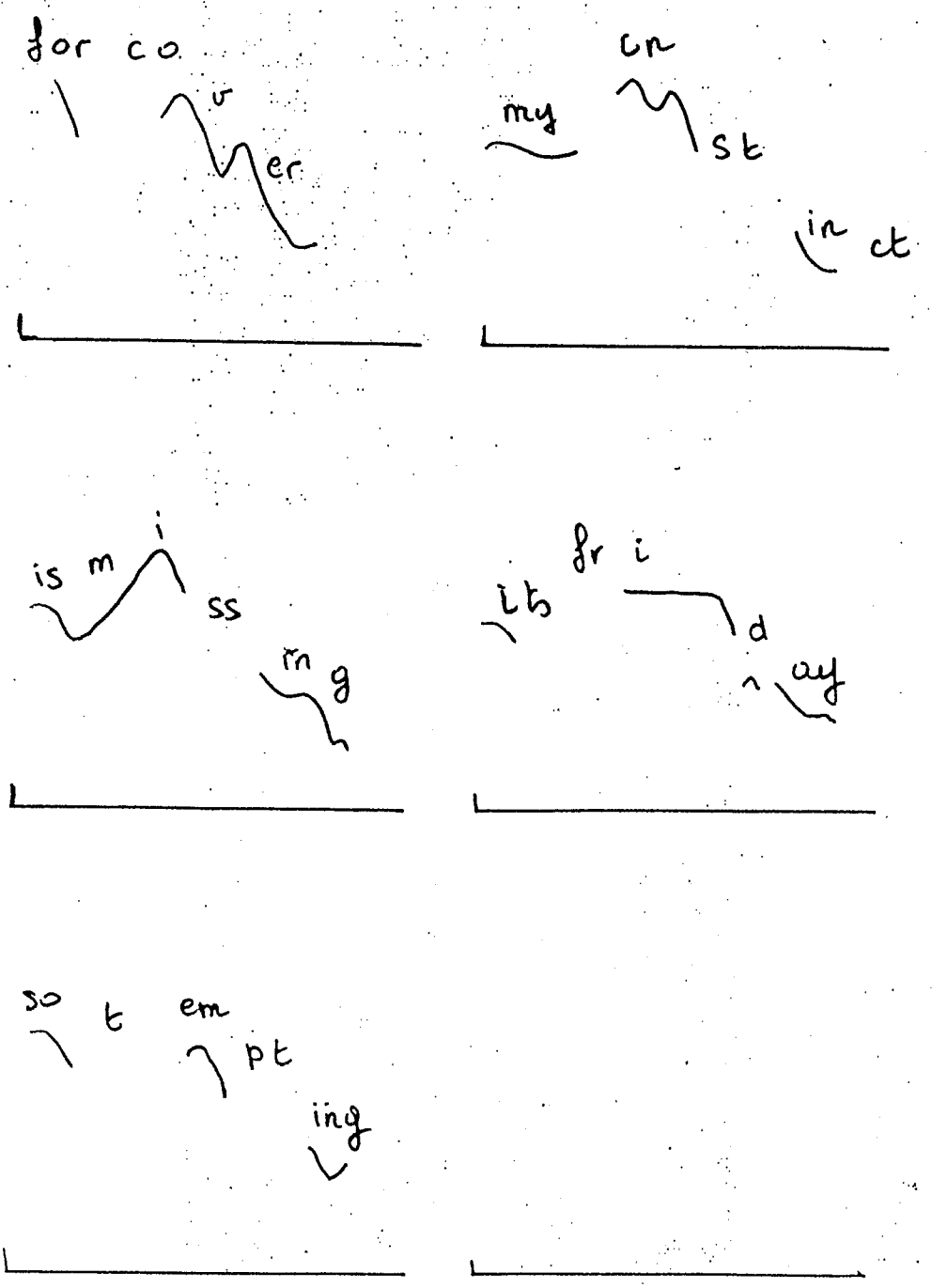


Fig. II.14

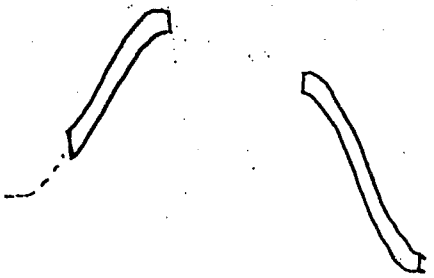
SCHEMATIC HP SUPERIMPOSED
TO ONE, TWO OF MORE WORDS

One word

Monosyllabic

Bisyllabic [cv *CV(C)]

two words



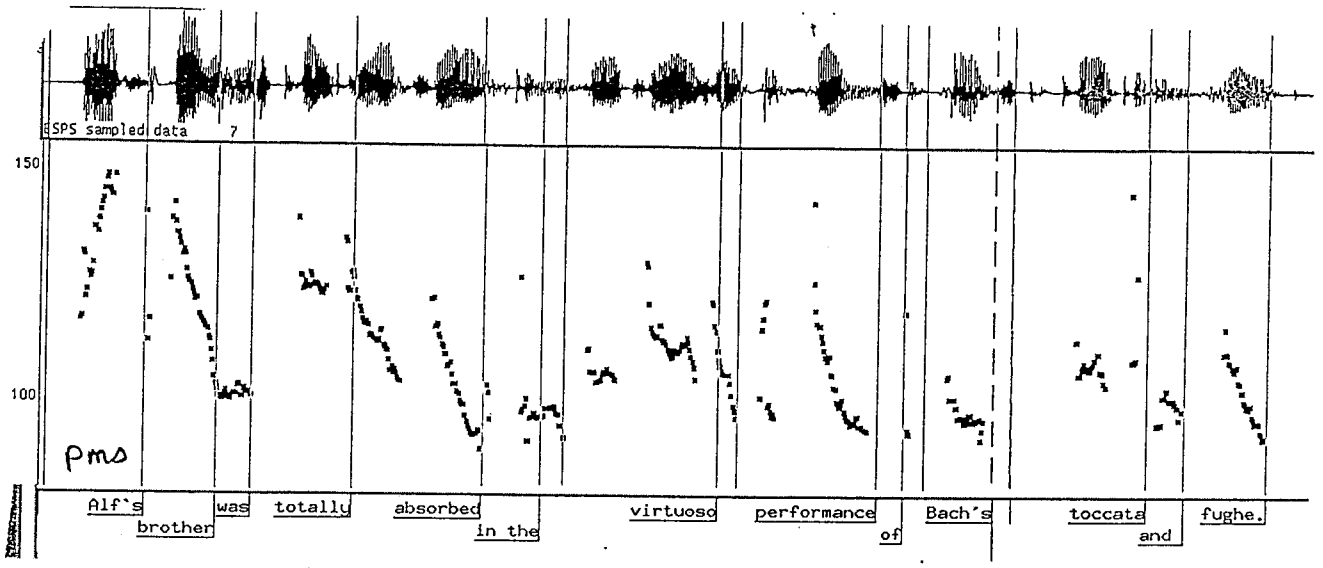
More words

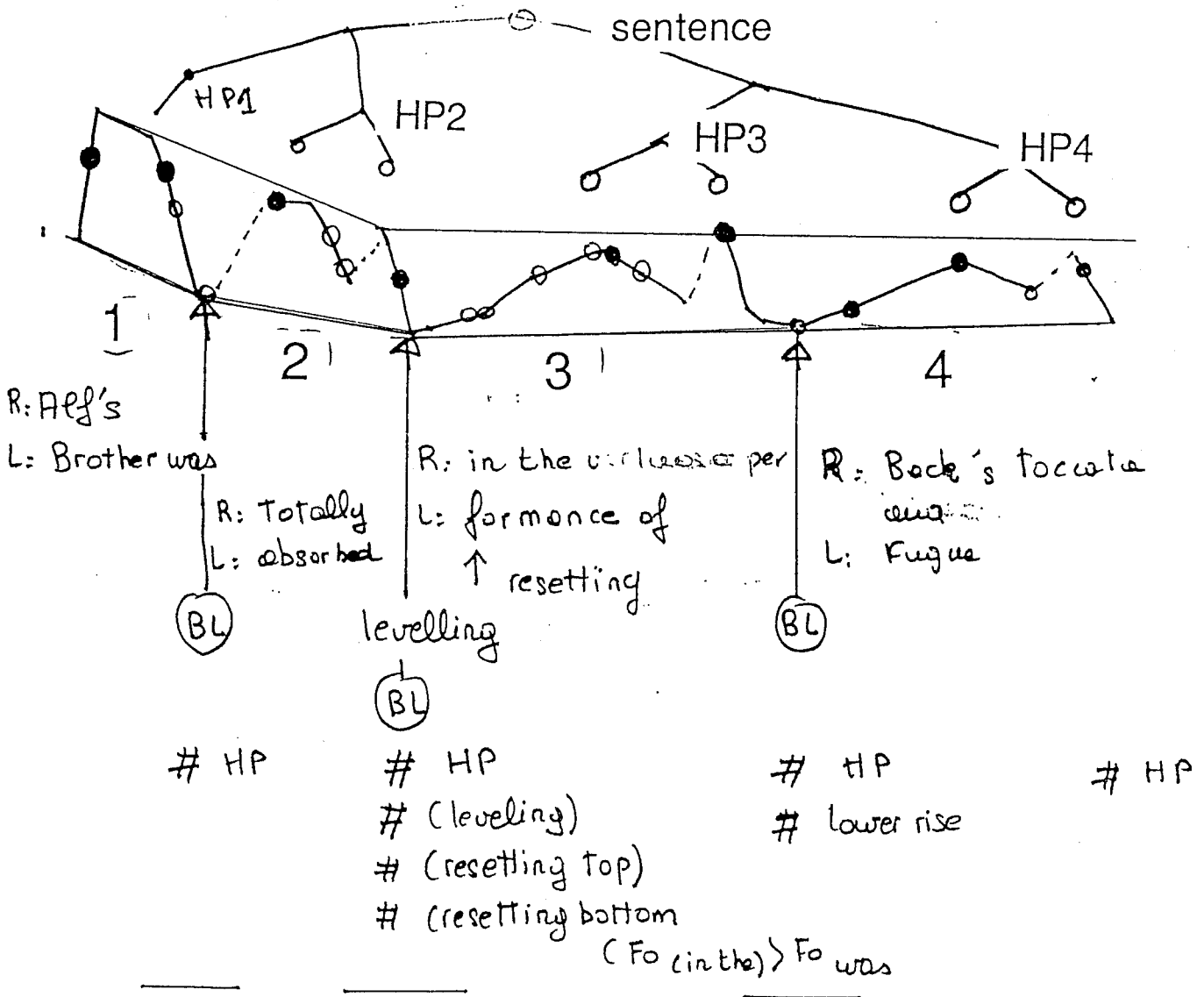
Rise in the stressed syllable of W1

Lowering in the stressed syllable of last word

Fig. II.15

PROSODIC TREE (1)

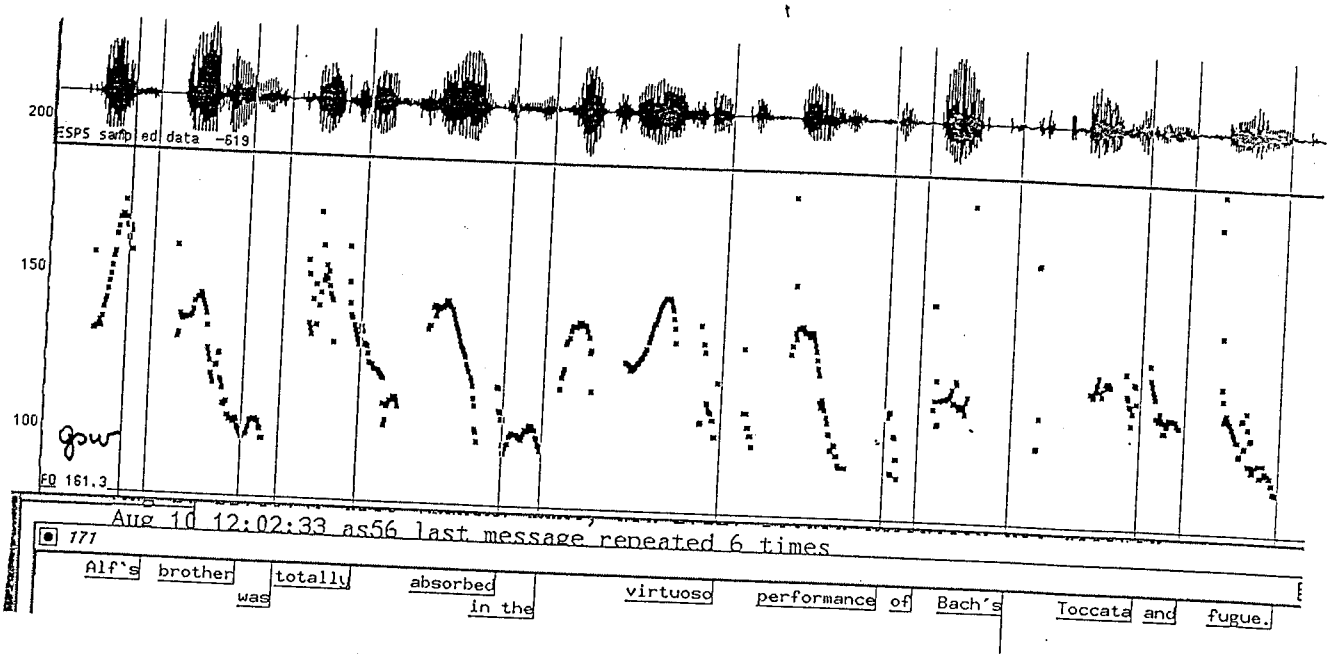


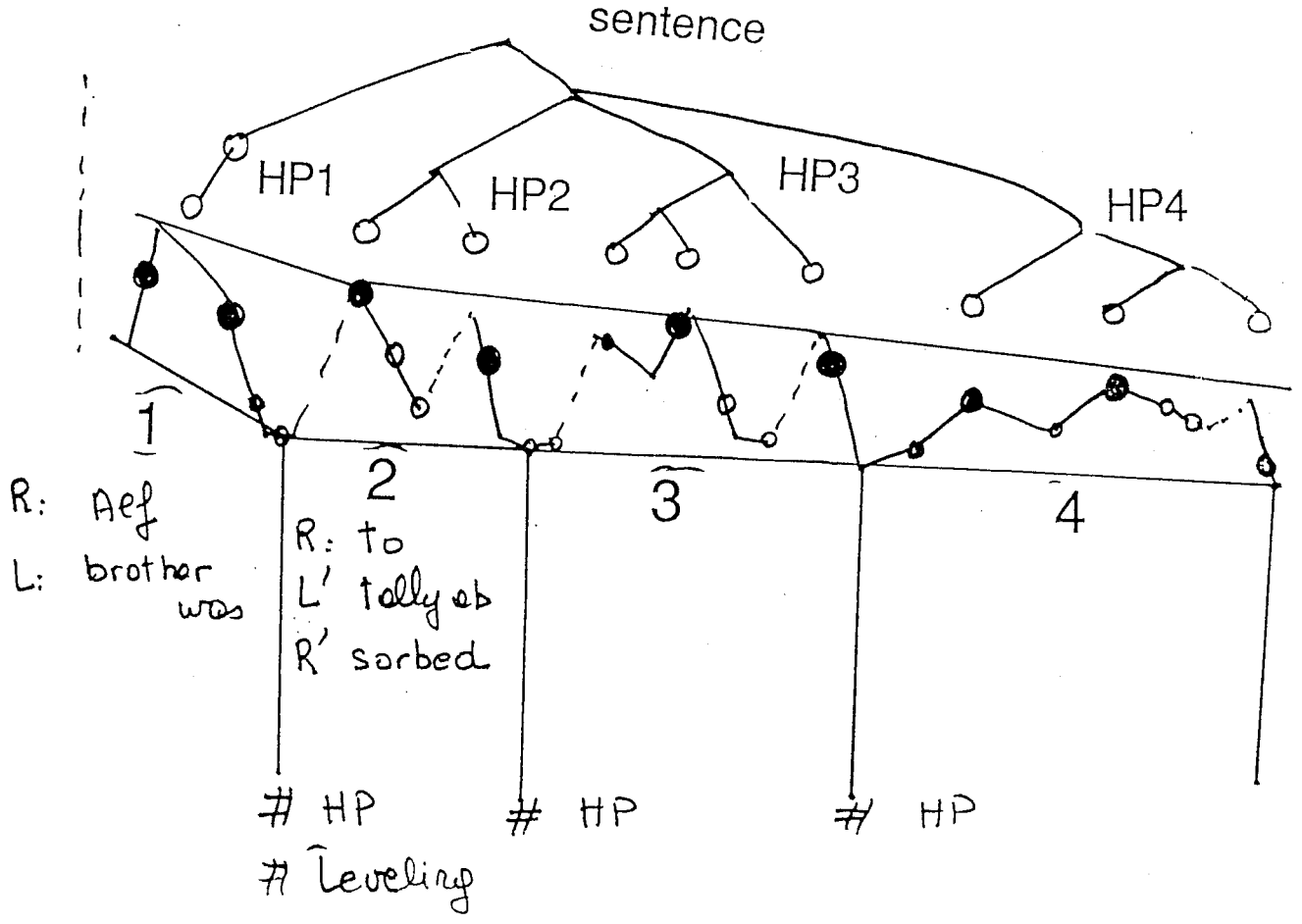


one cue four cues two cues

Fig. II.16

PROSODIC TREE (2)





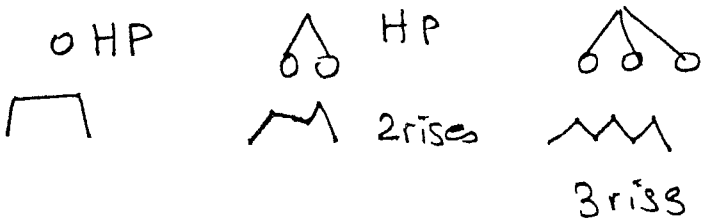
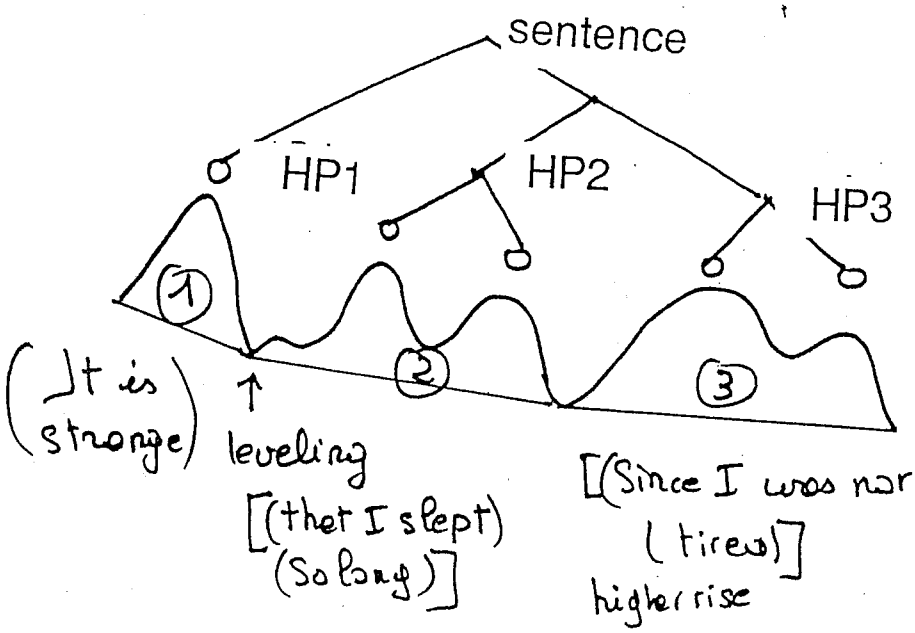
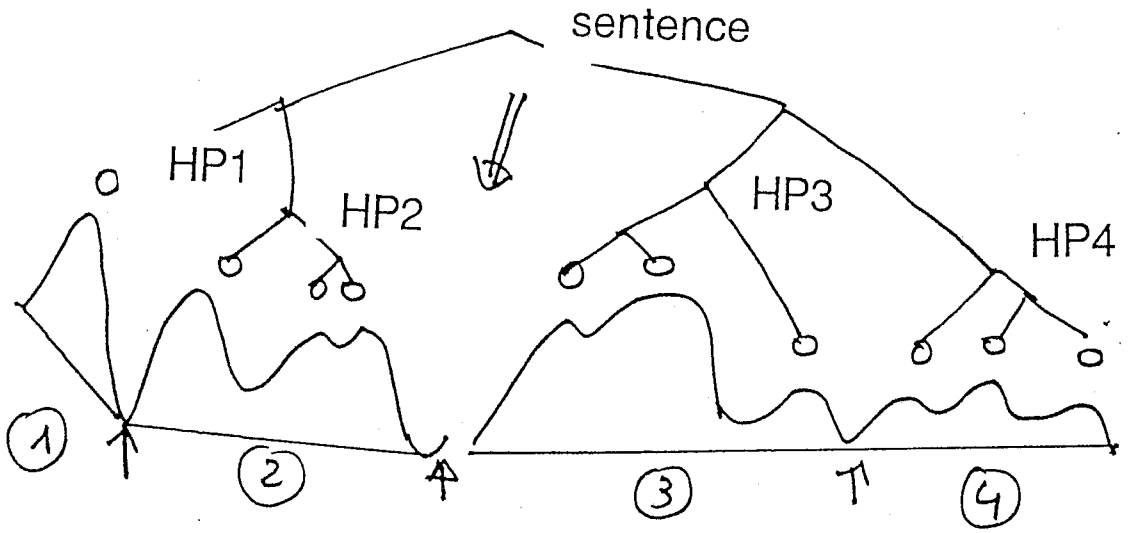


Fig. II.17

PROSODIC TREE (3)



conflicting cues:
- leveling (1)(2-3)
- higher rise (1-2)(3)



#
—
one cue

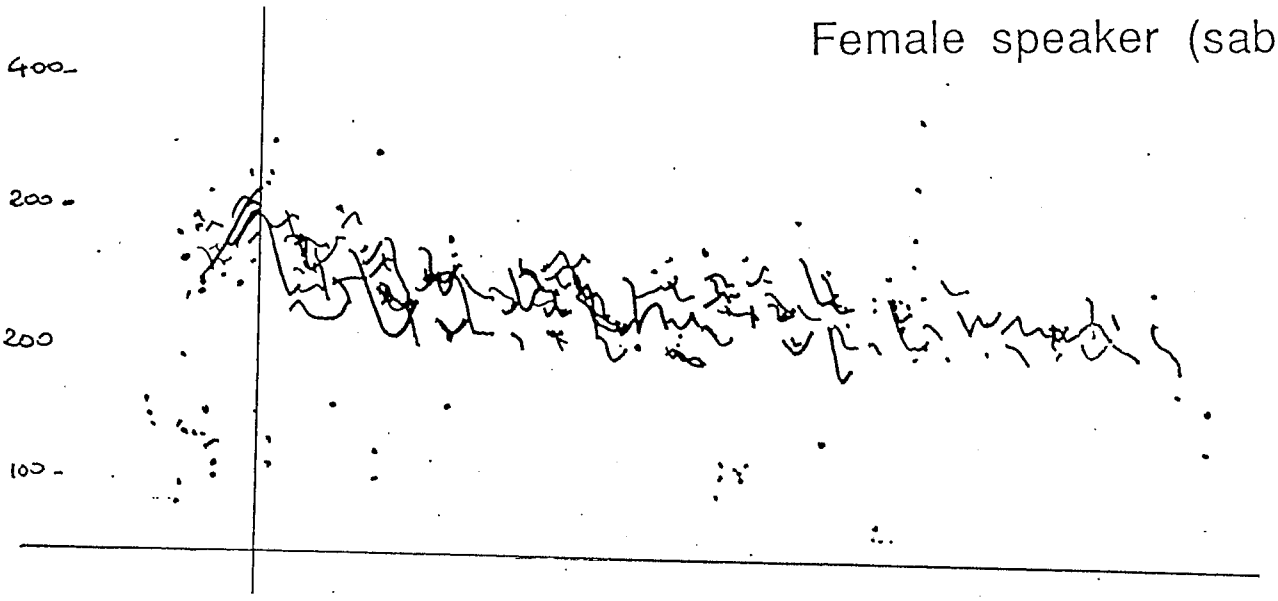
#
#
#
#
—
four cues

#
—
one cue

Fig. II.18

SPEAKER RANGE

Female speaker (sab)



Male speaker (gwm)

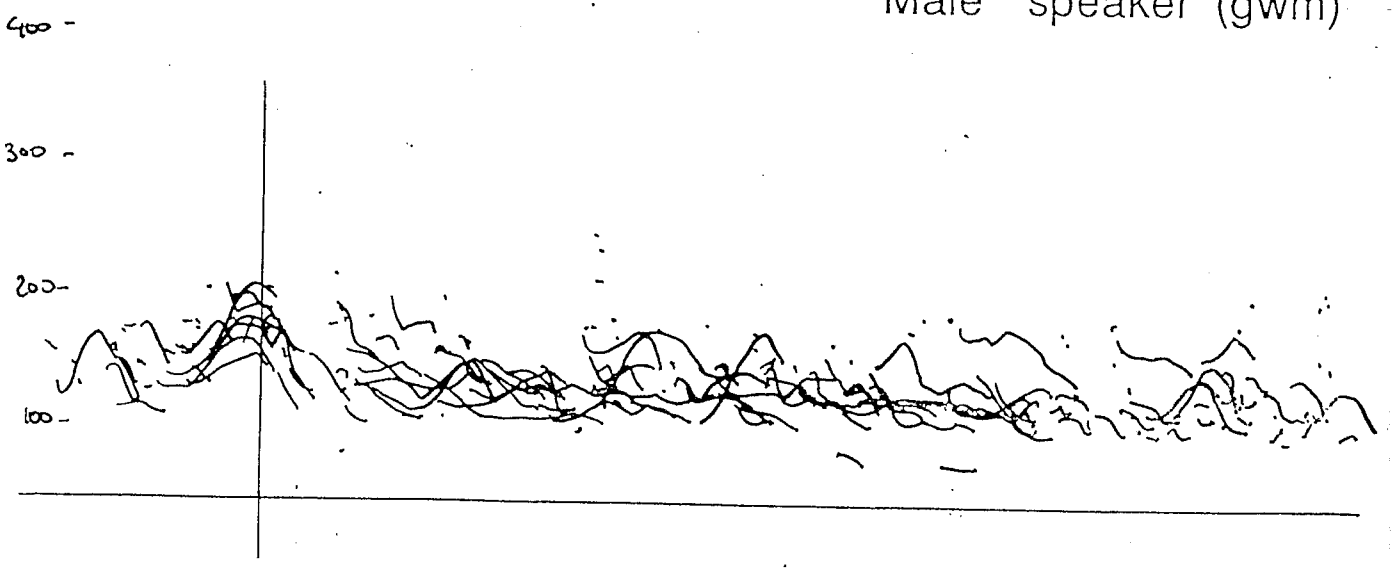
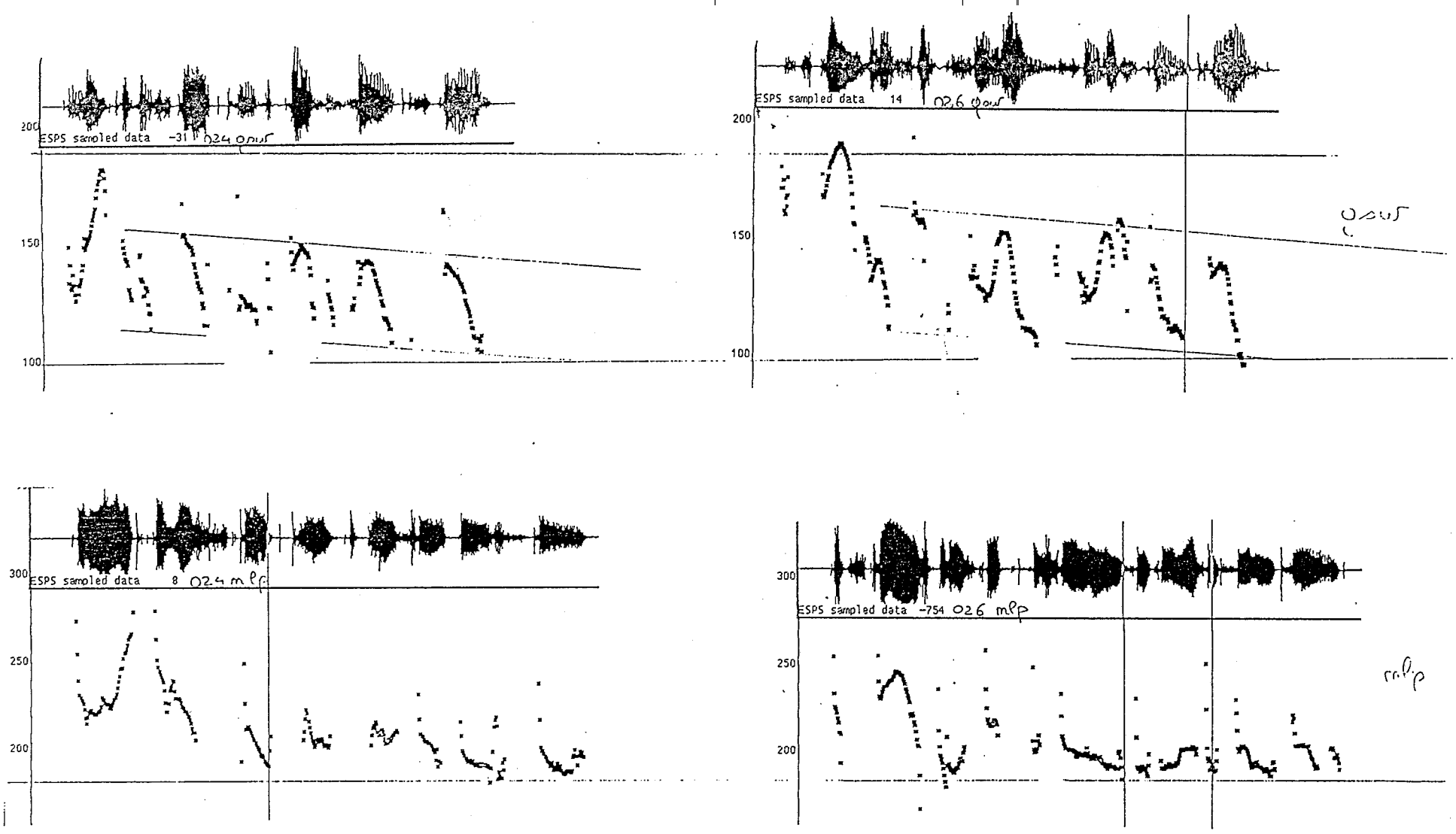


Fig.III.1

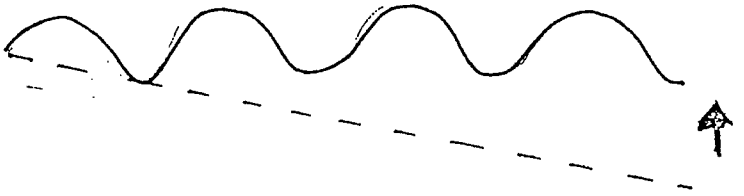


Fo RANGE VARIATIONS

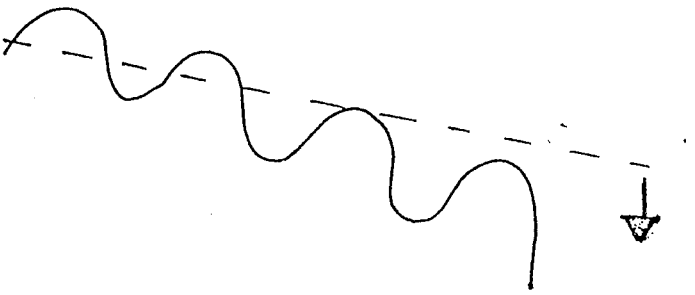
Fig.III.2

DECLINATION TENDENCY AND SENTENCE TYPES

a) Normal tendency (declarative)



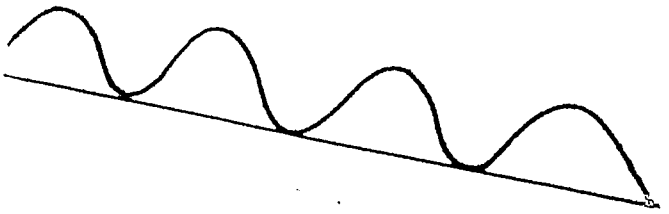
b) Suppression (yes-no question)



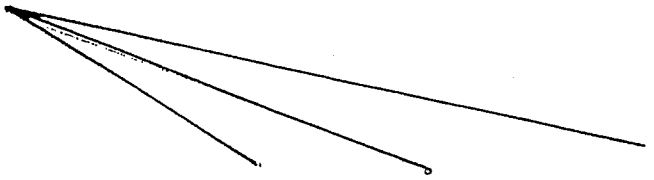
c) Exaggeration (order)_

Fig.III.3

DECLINATION TENDENCY IN DECLARATIVE SENTENCE



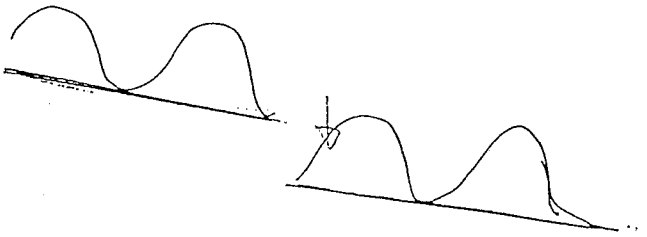
a) Normal tendency



b) Adjustment of the slope

c) Leveling

d) Resetting

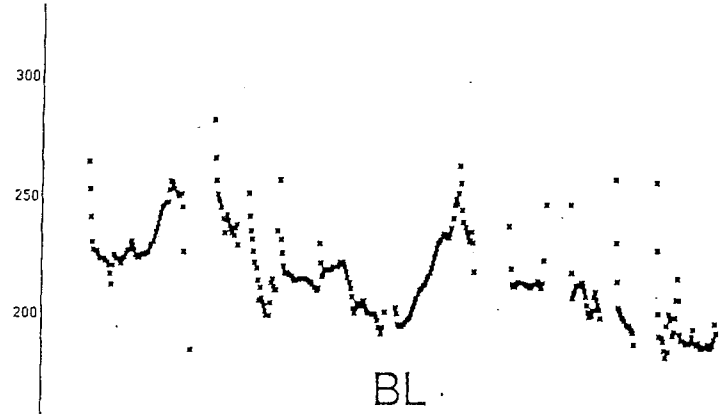


e) Downstepping

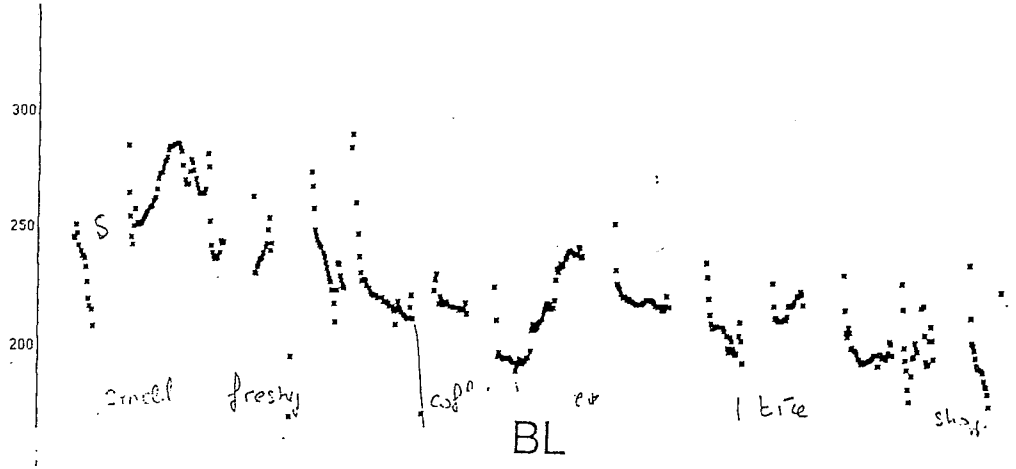
Fig.III.4

MAIN BOUNDARY MARKING

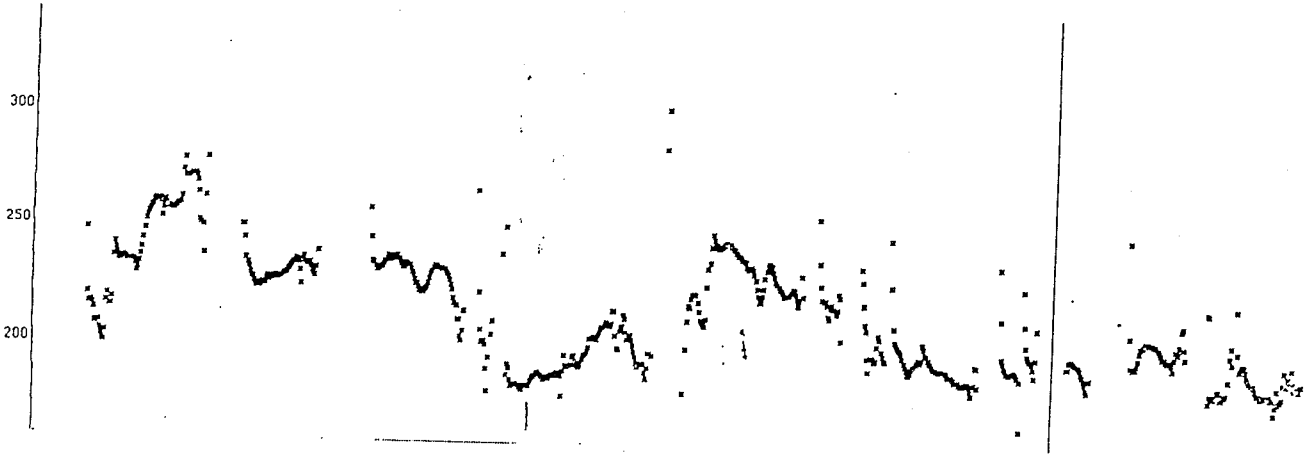
mlp 01



mlp 004

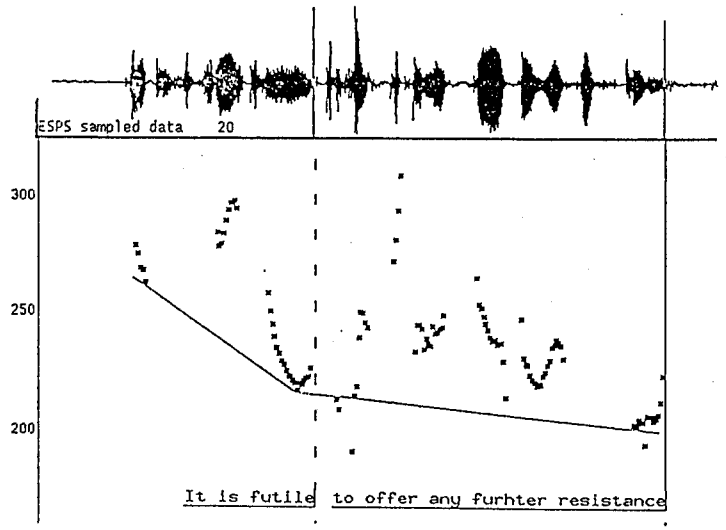


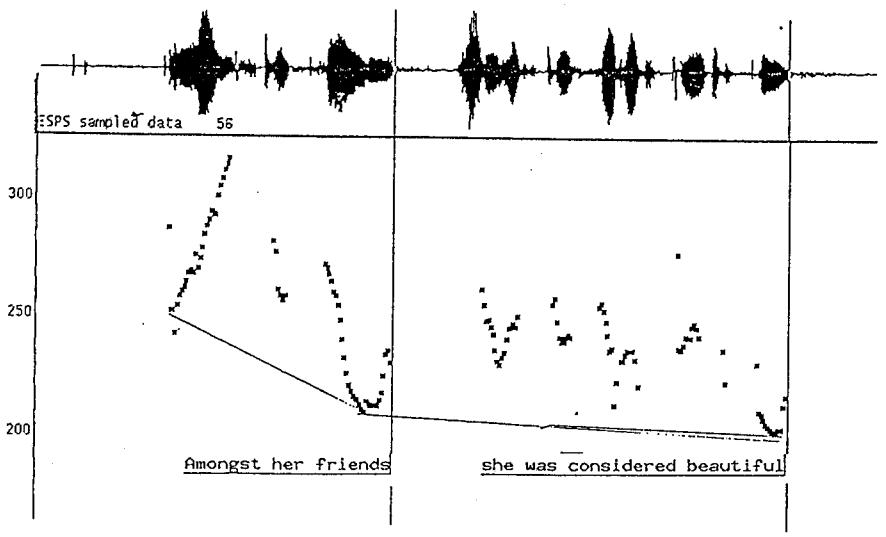
mlp 013



speaker mlp

DIVISION INTO 2 SENSE-GROUPS: EXAMPLES





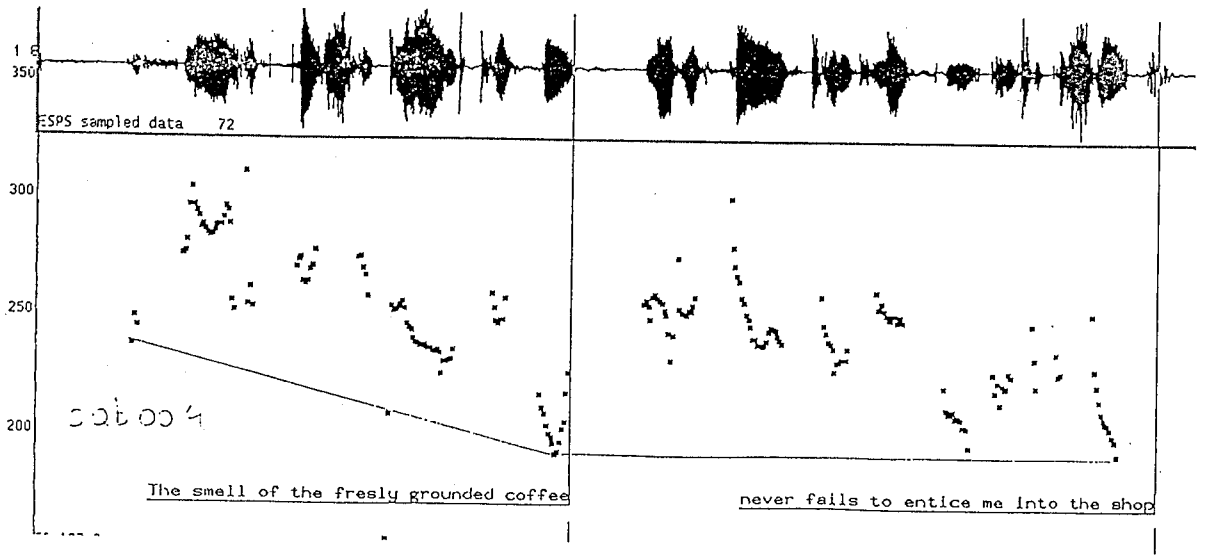
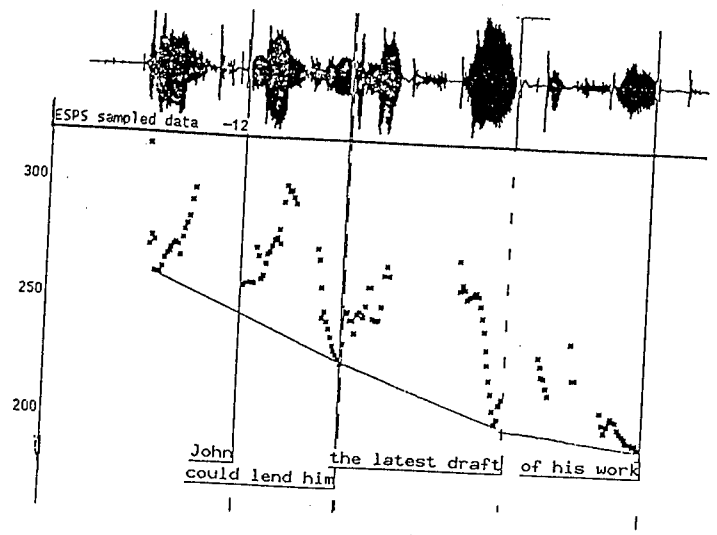
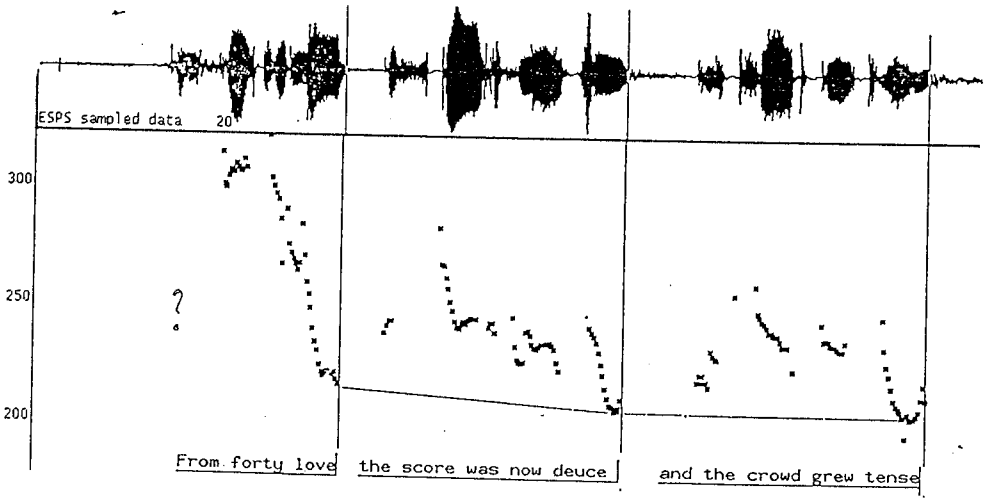


Fig.III.7

DIVISION INTO 3 SENSE-GROUPS: EXAMPLES





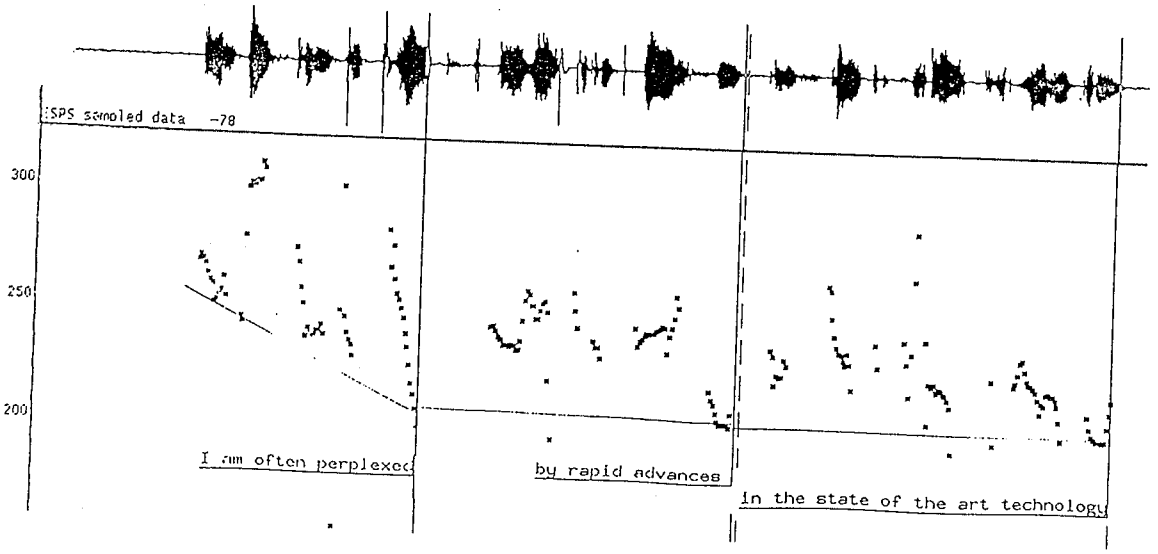
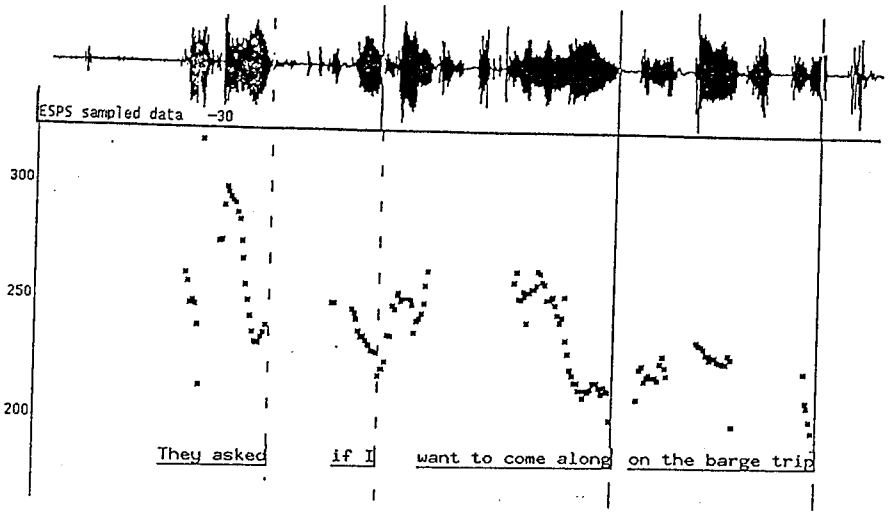
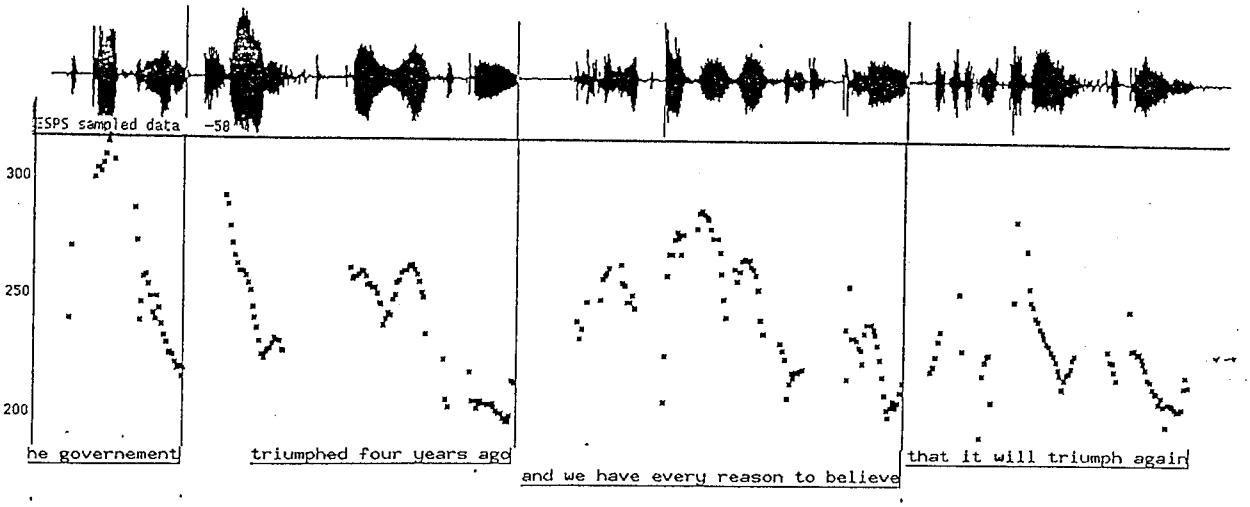


Fig.III.8

DIVISION INTO 4 SENSE-GROUPS: EXAMPLES





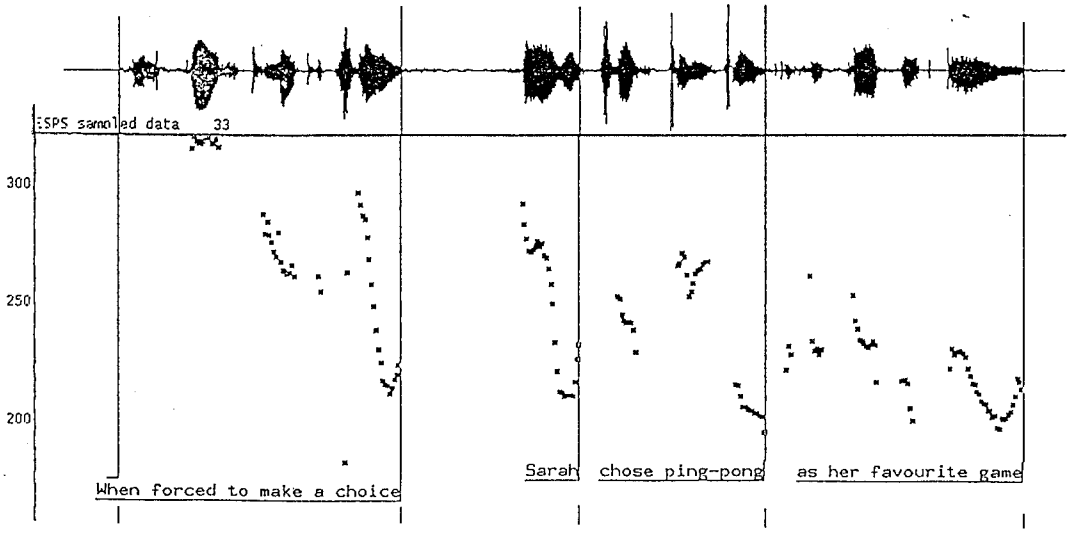
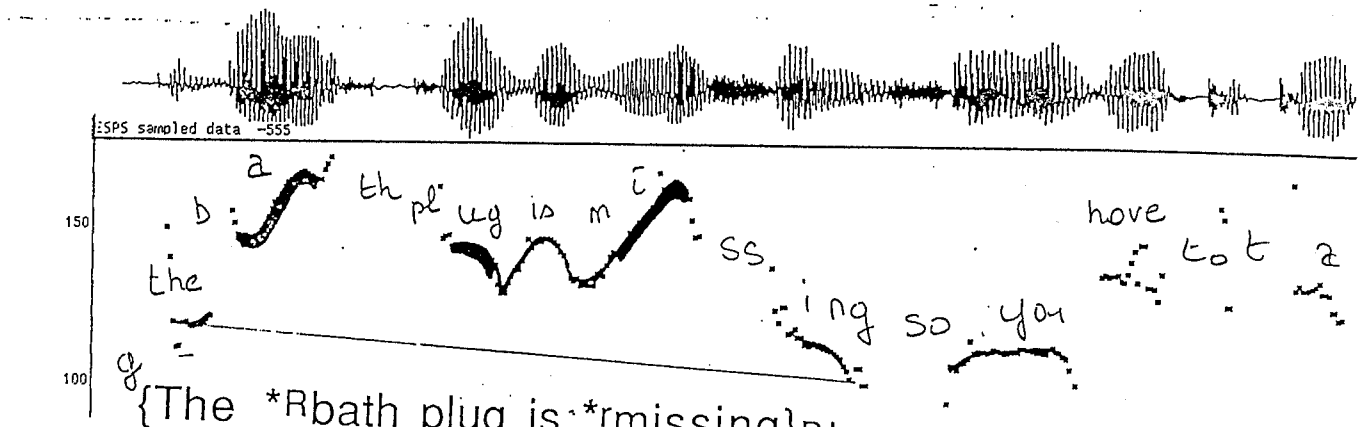


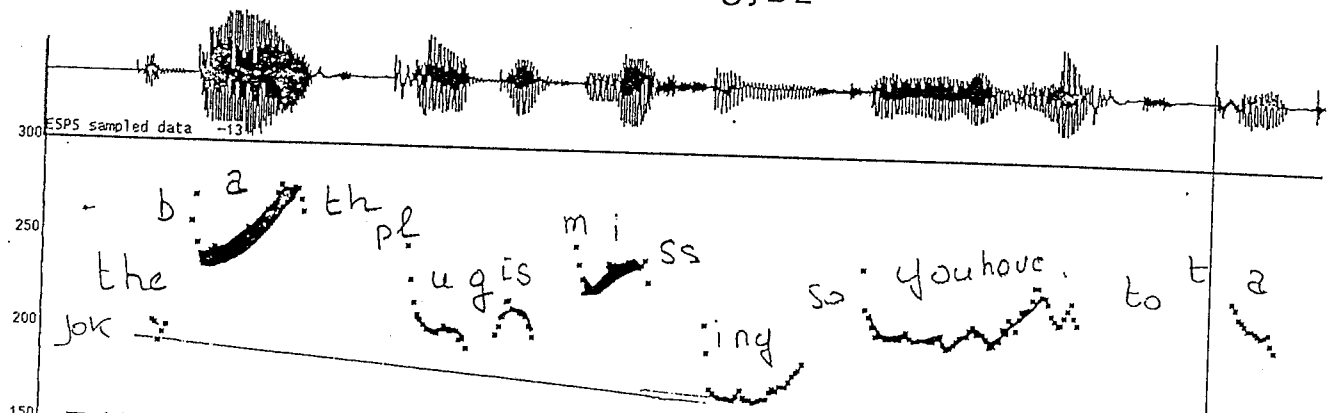
Fig.III.9

INTERSPEAKER VARIATIONS:

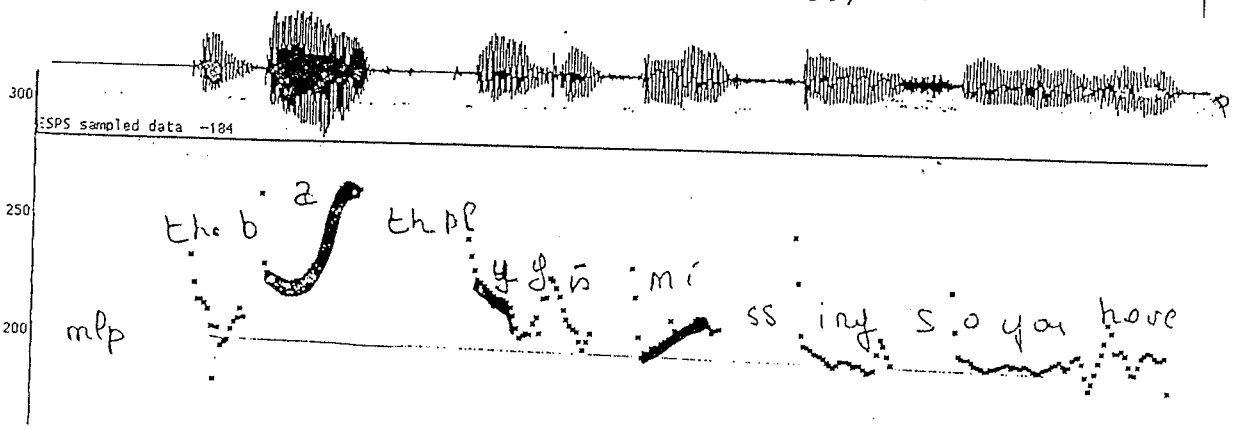
Compound word

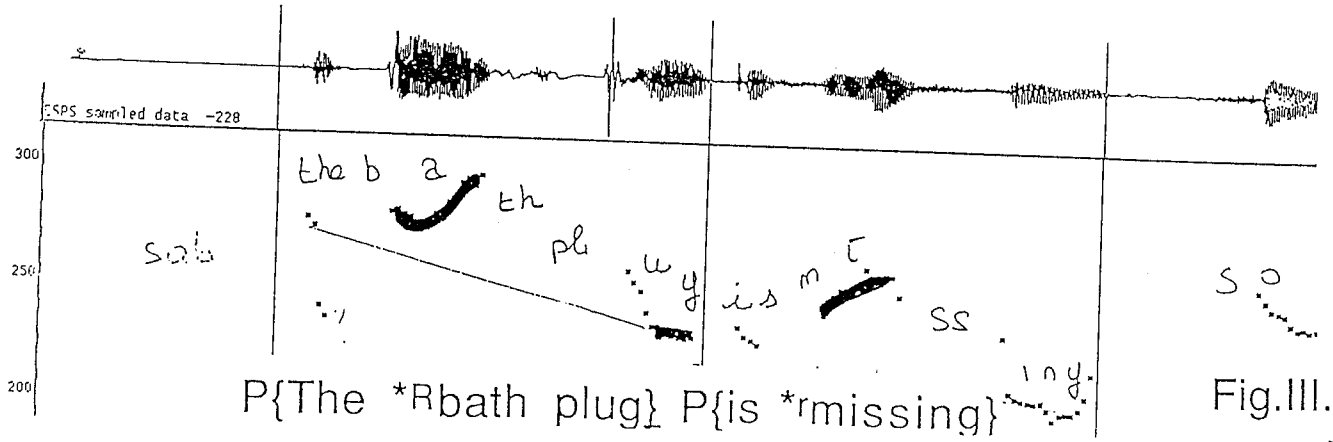


{The *Rbath plug is *rmissing}BL



P({The *Rbath plug} P{is *hmissing})+Rc



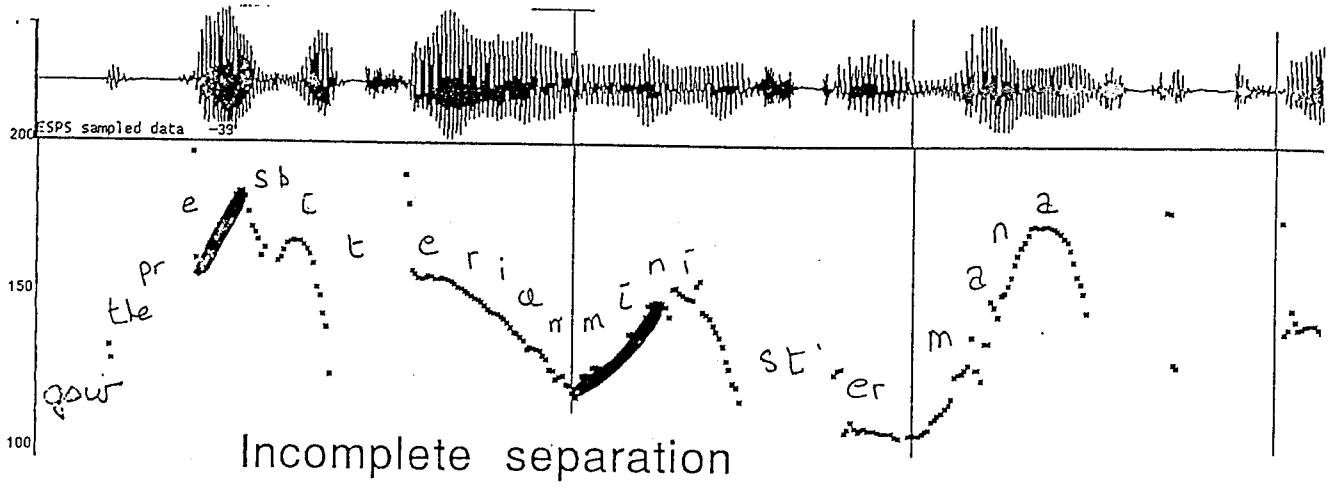


P{The *Rbath plug} P{is *rmissing}

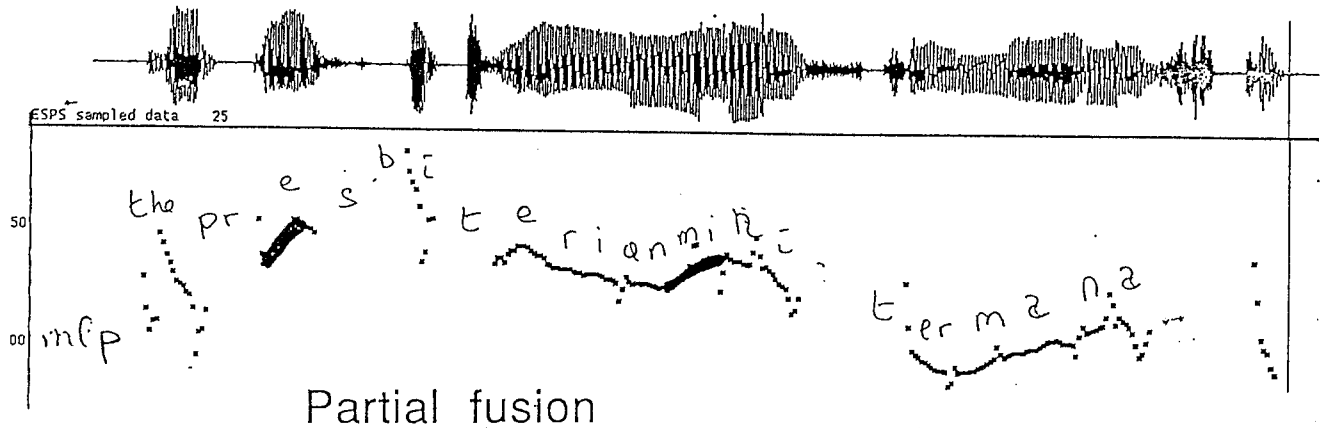
Fig.III.10

INTERSPEAKER VARIATIONS

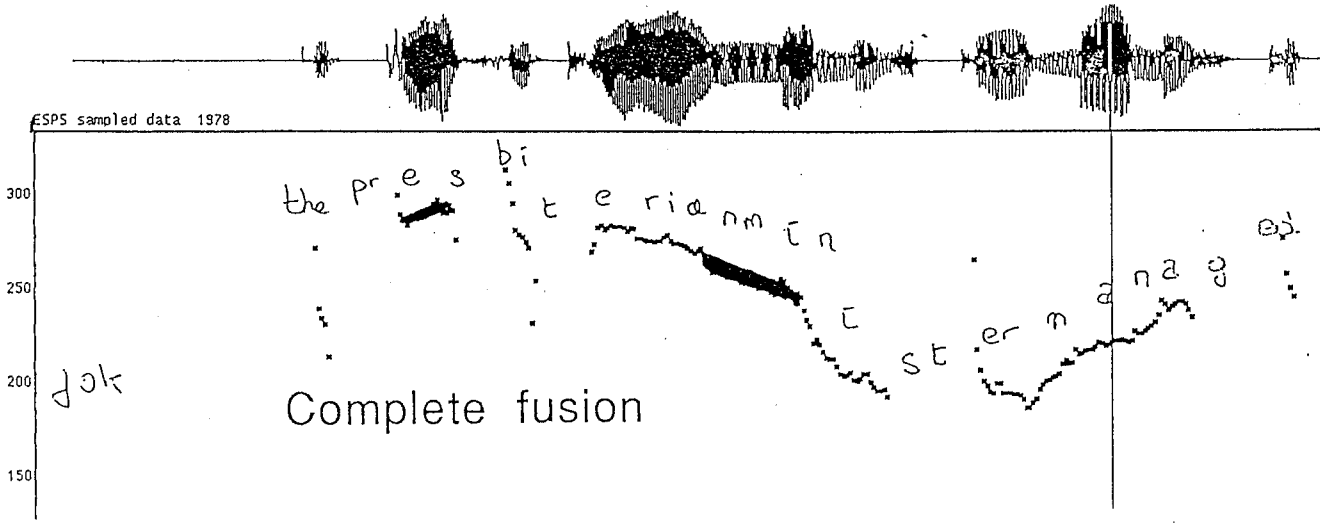
ADJECTIVE + NOUN



Incomplete separation



Partial fusion



Complete fusion

The presbitarian minister:
R (Rp) L
(the Rise on the plateau is optional)

Fig.III.11

INTERSPEAKER VARIATIONS

FUNCTION WORDS WITH LOWERING OR LOW F0 CONTOURS

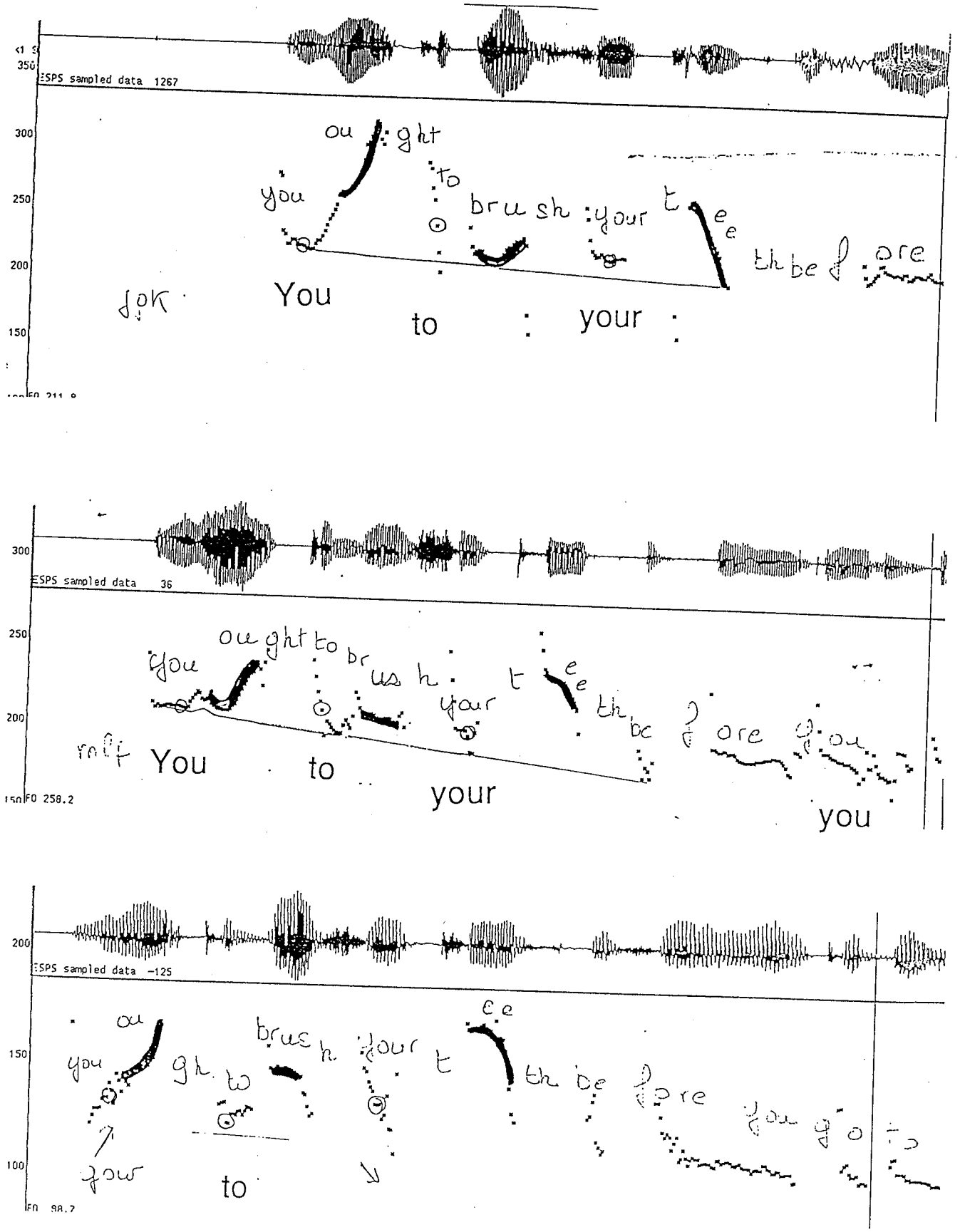


Fig.III.12

You ought to brush your teeth before you go to ...

INTRASPEAKER VARIATIONS IN SENTENCE INITIAL RISE

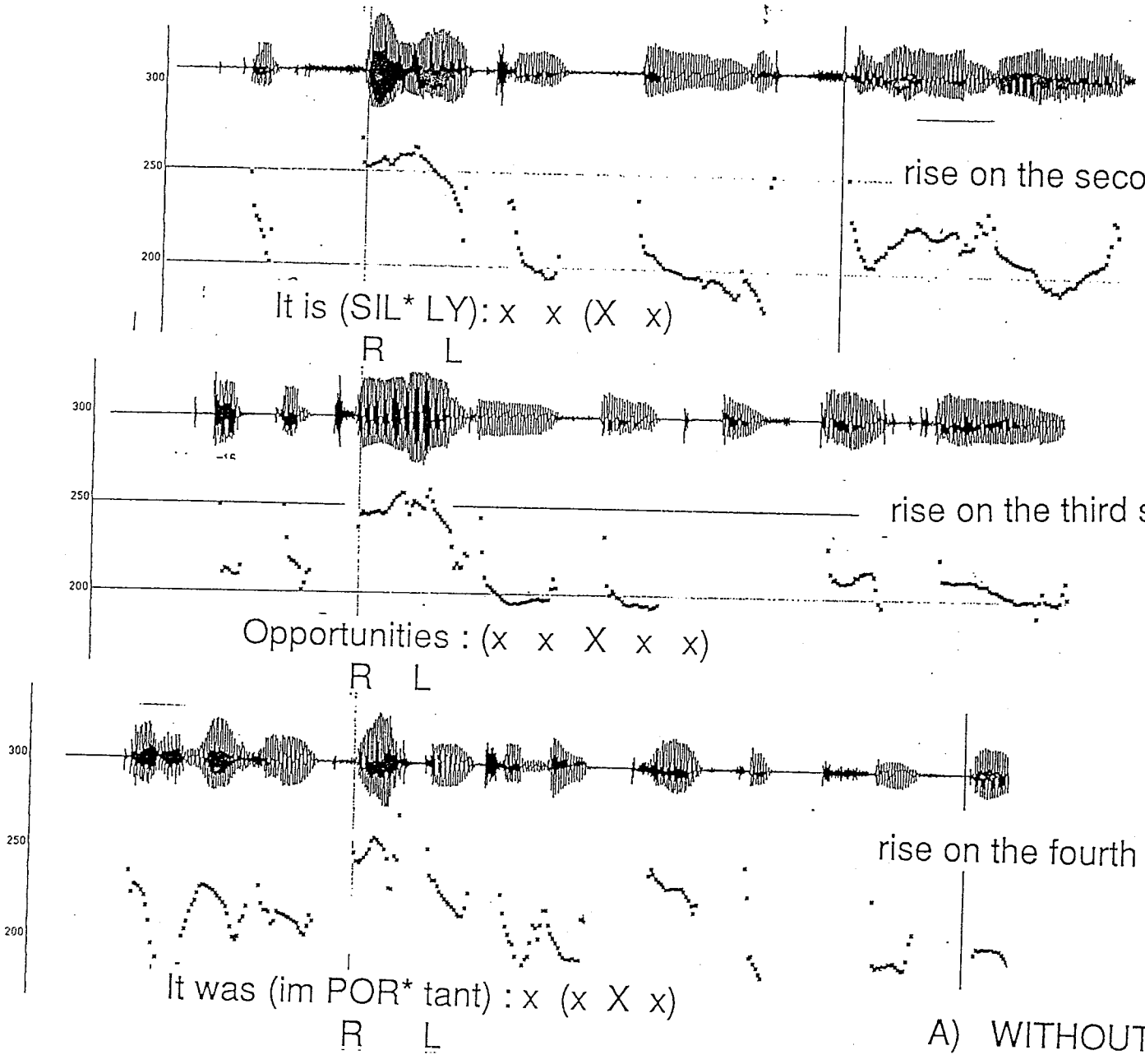
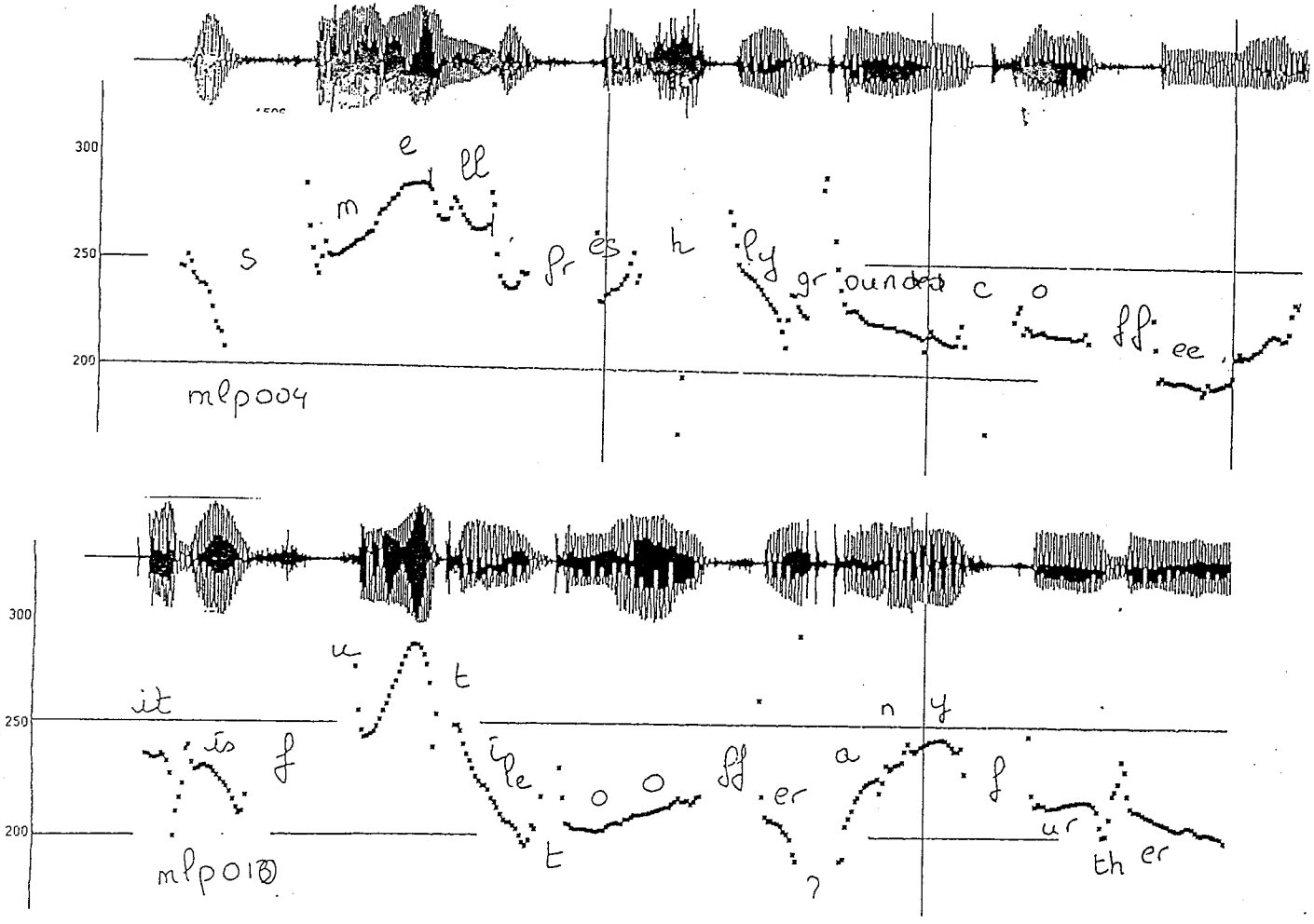


Fig.III.13

A) WITHOUT AN EXTRA PEAK

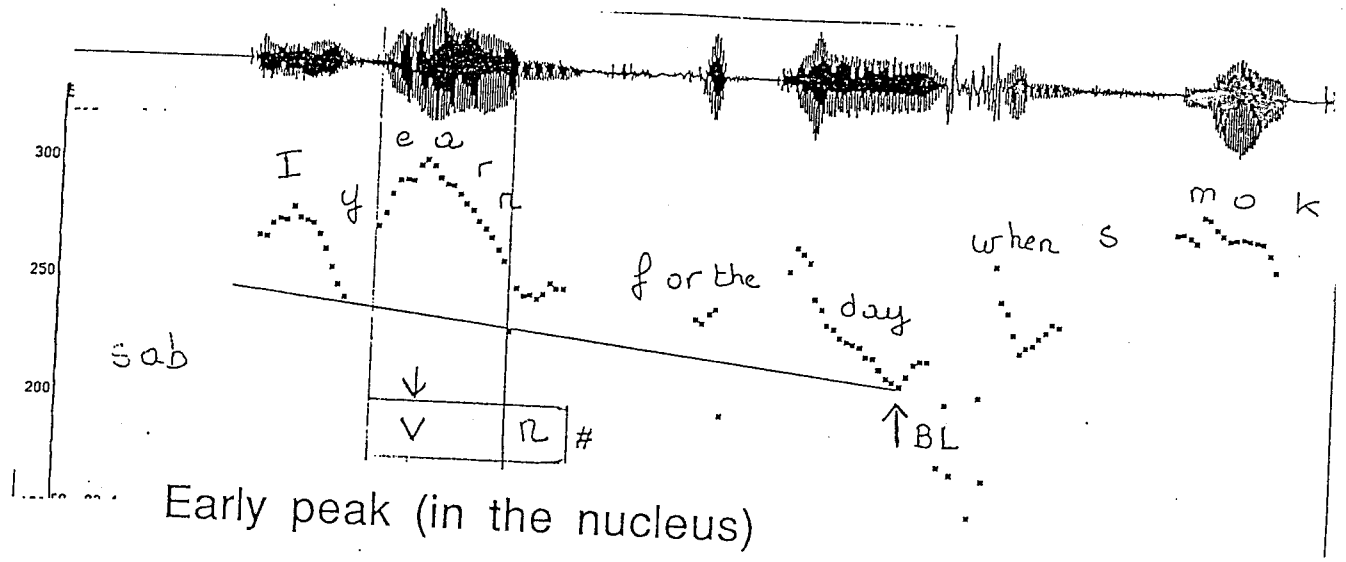


WITHOUT AND WITH AN EXTRA PEAK

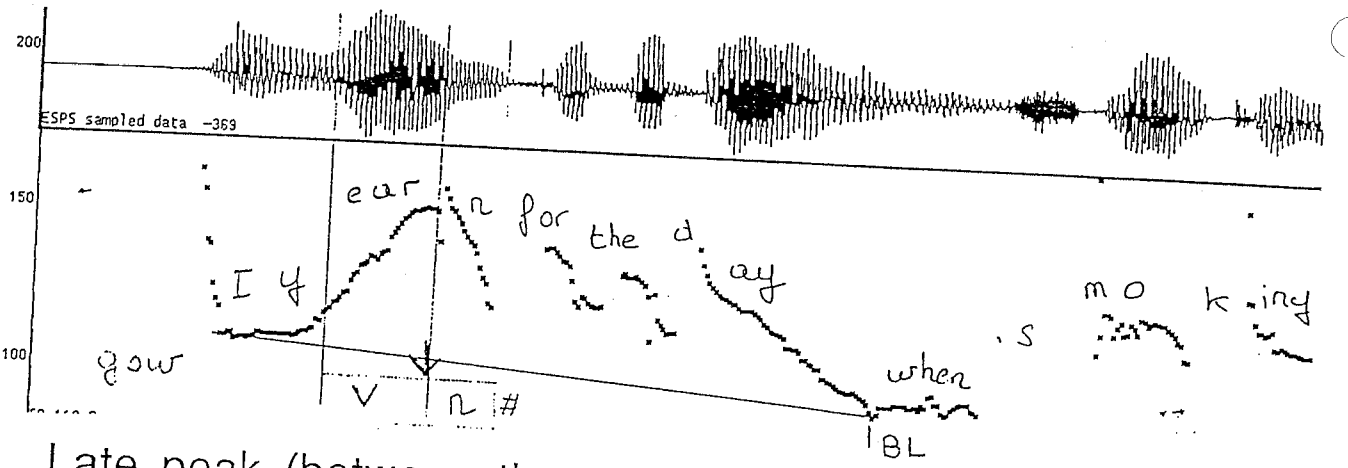
Speaker mlp

Fig.III.14

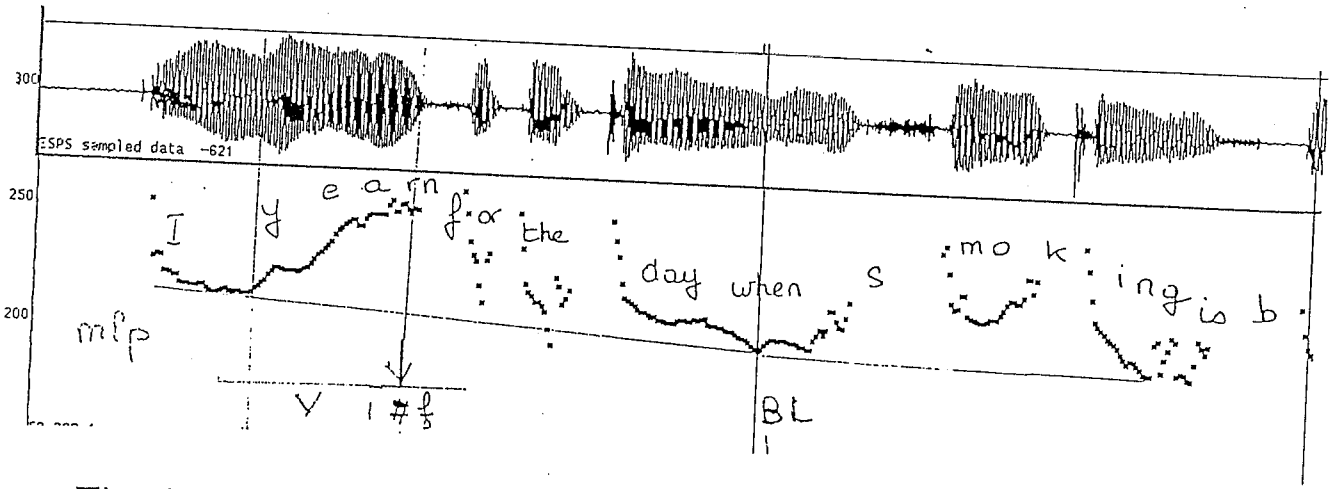
VARIATIONS IN PEAK LOCATION (A)



Early peak (in the nucleus)



Late peak (between the nucleus and the coda)



Final Peak (at the end of the coda)

(I yearn for the day) when smoking ...

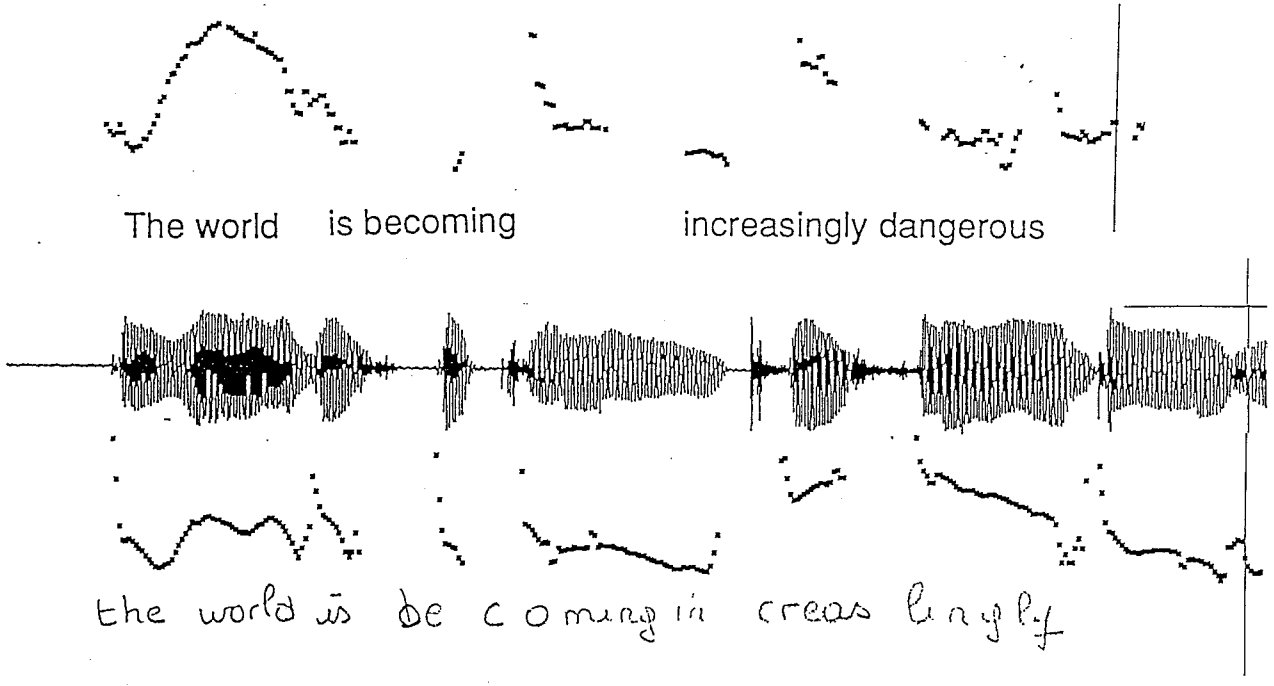
Fig.III.15

Fig.III.16

IOPTIONAL DESACENTUATION OF A LEXICAL WORD

General case

Desacentuation



FOCUSING

gaw
047 mlp

INTERSPEAKER VARIATIONS
DIFFERENT CHOICE OF ACCENTUED WORD

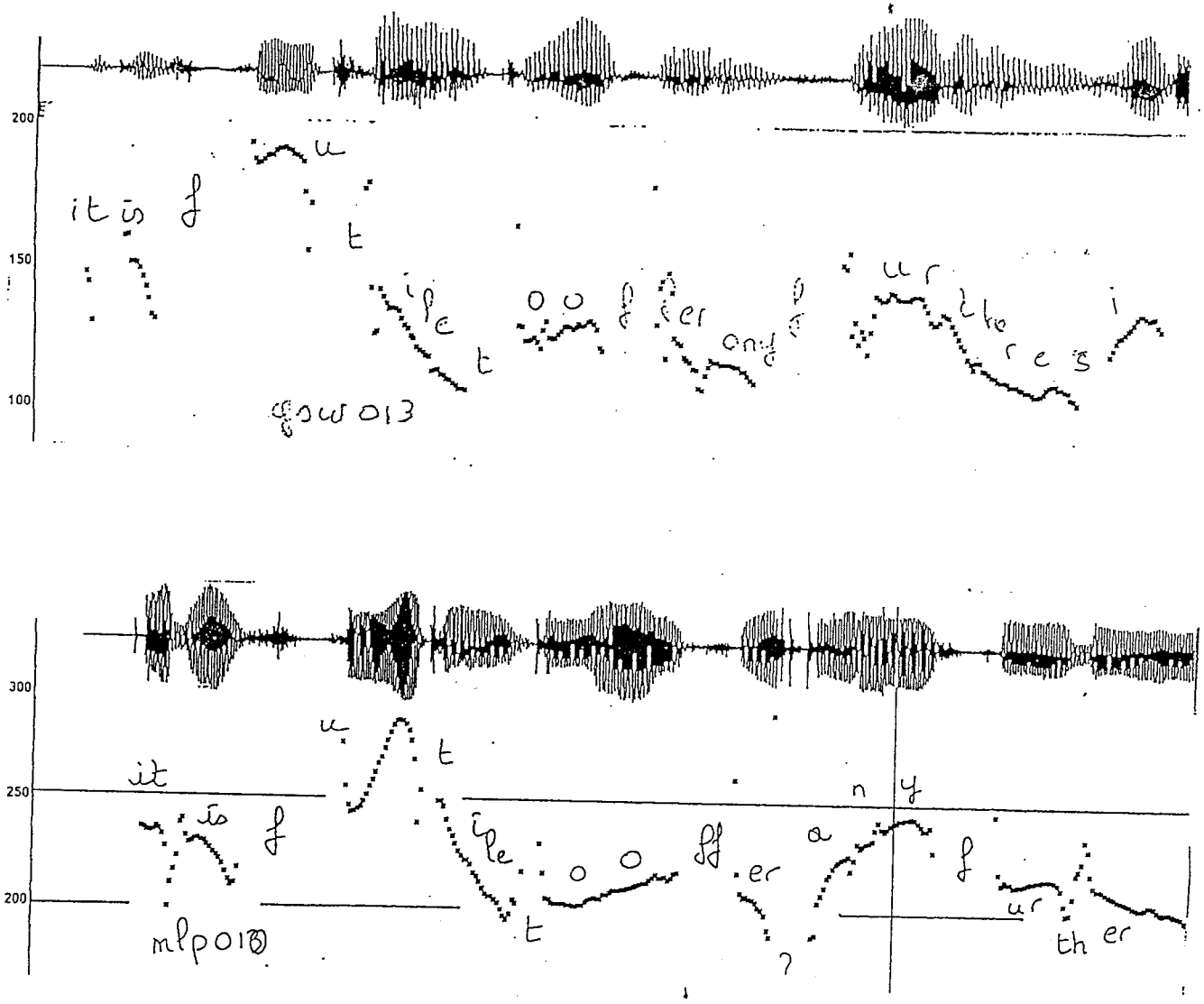
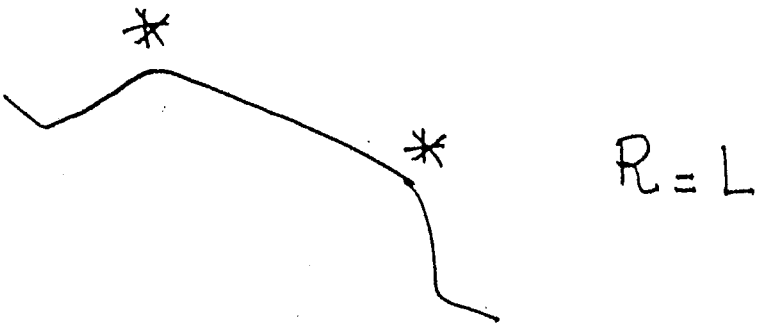
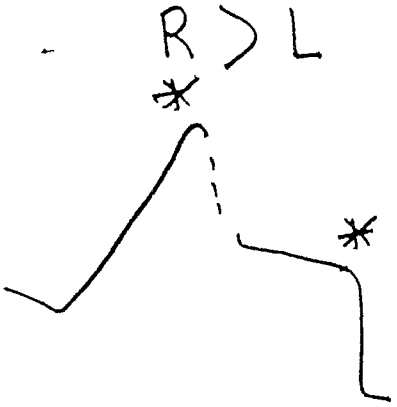
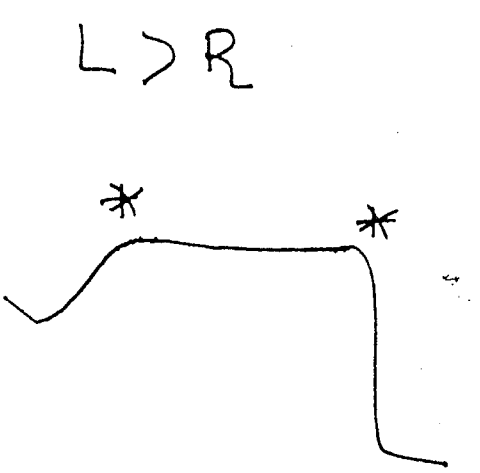


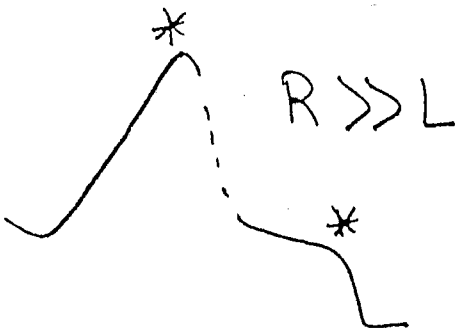
Fig.III.17

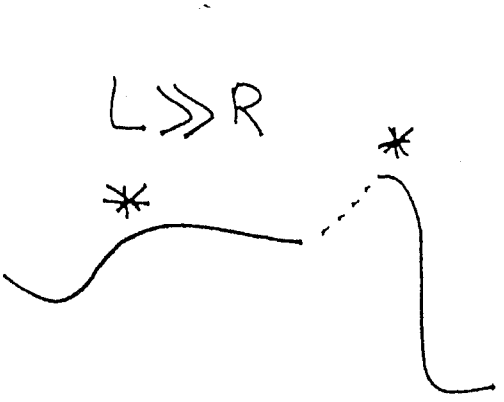
SPEAKER TENDENCIES







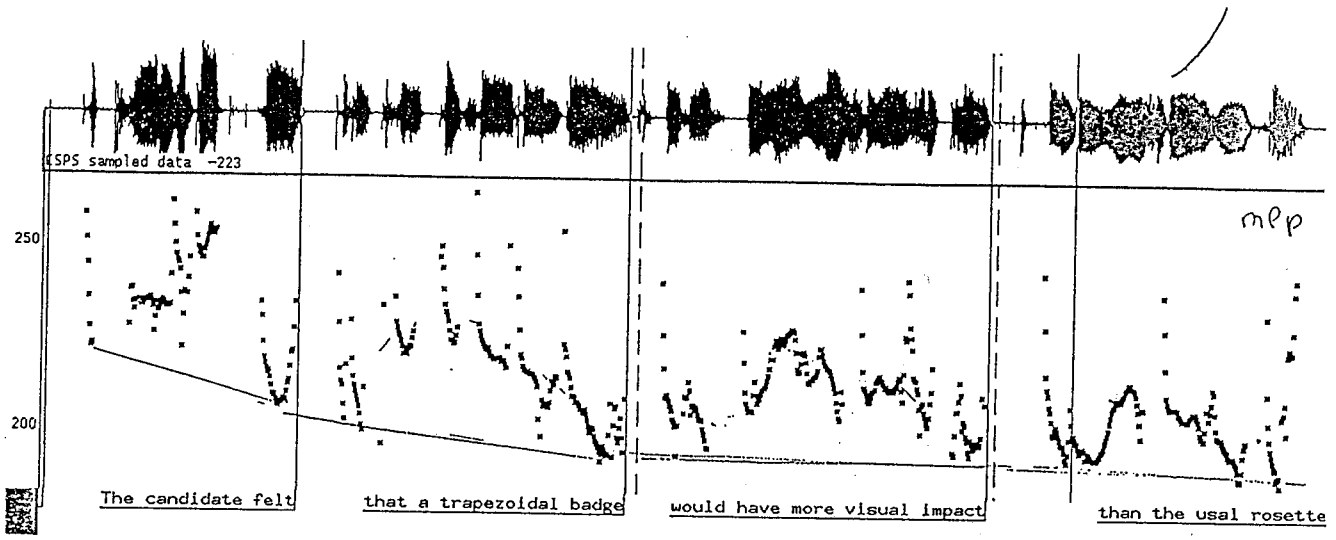


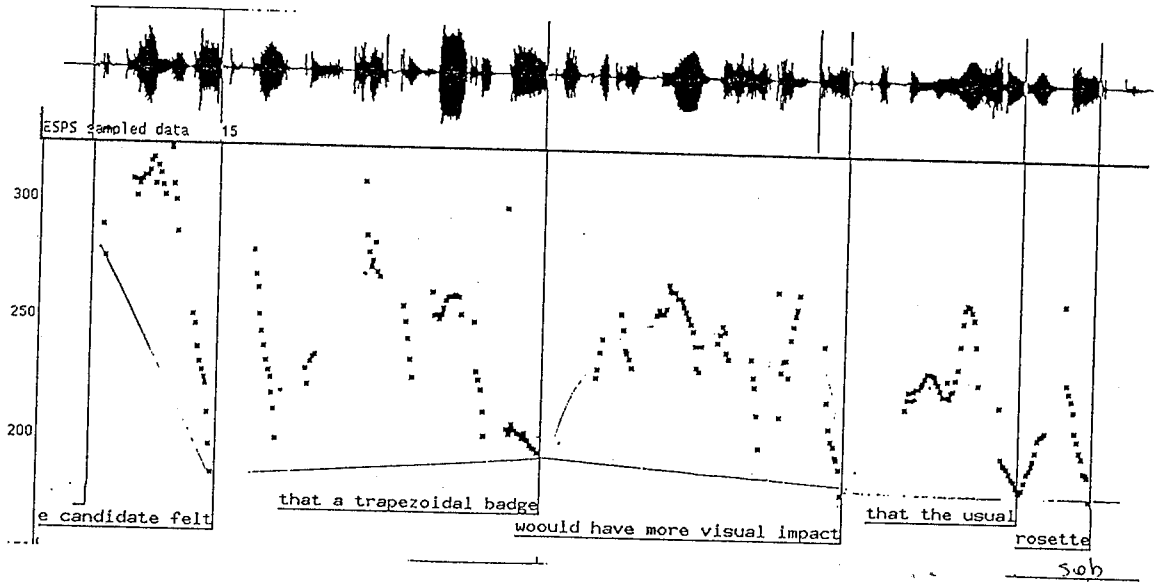


P4 TENDENCY

P2 TENDENCY

Fig.III.18


ESPS sampled data -223
250
200
mpp
The candidate felt
that a trapezoidal badge
would have more visual impact
than the usal rosette


ESPS sampled data 15
300
250
200
e candidate felt
that a trapezoidal badge
woould have more visual impact
that the usual
rosette
sob

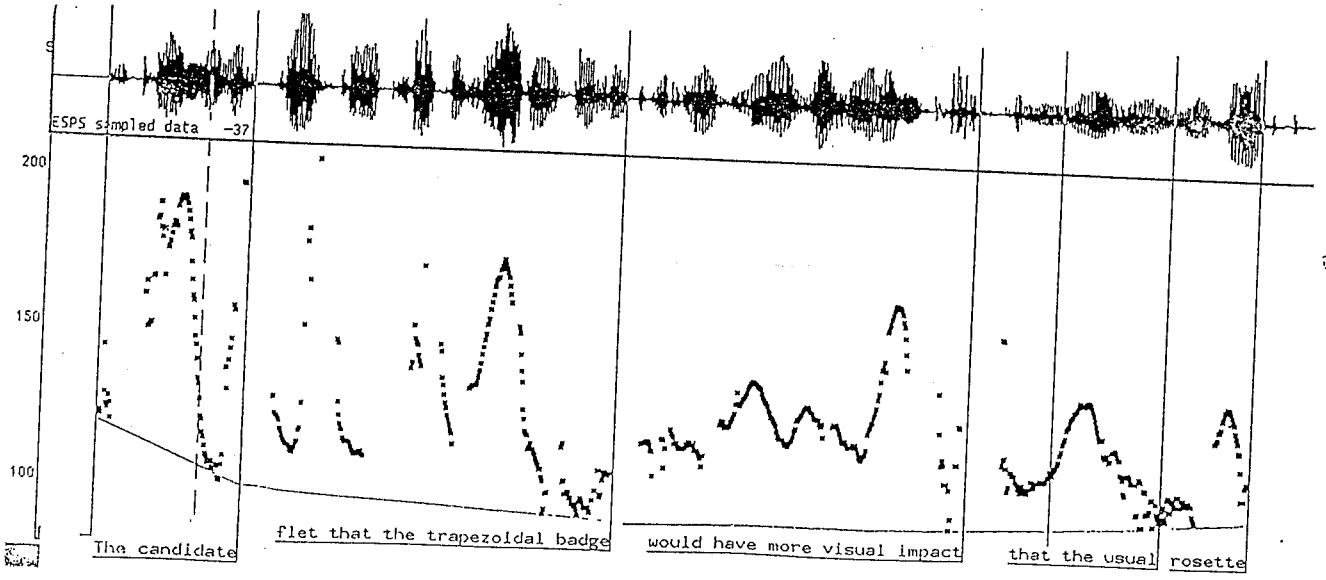
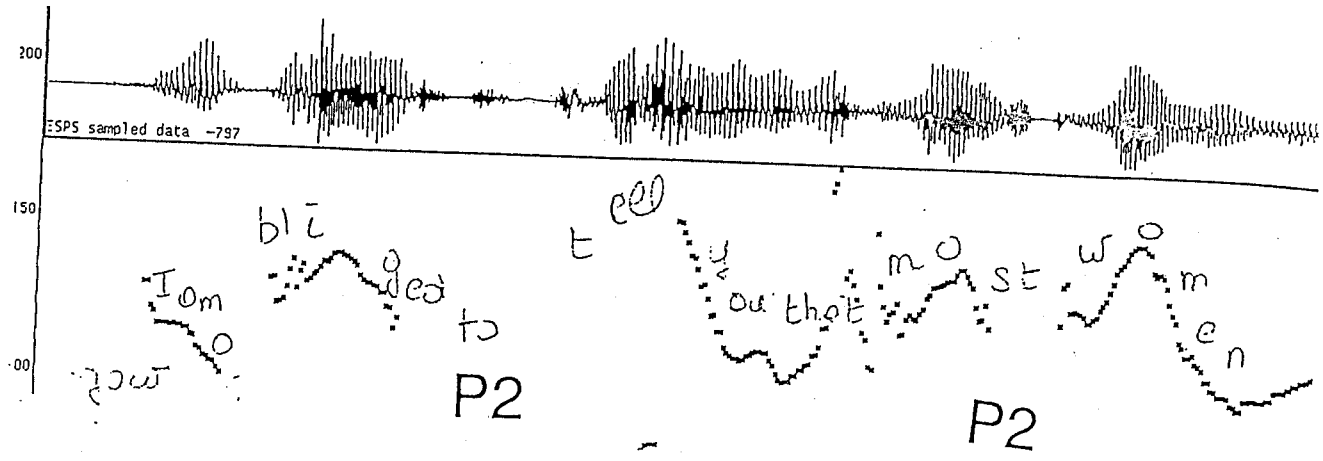
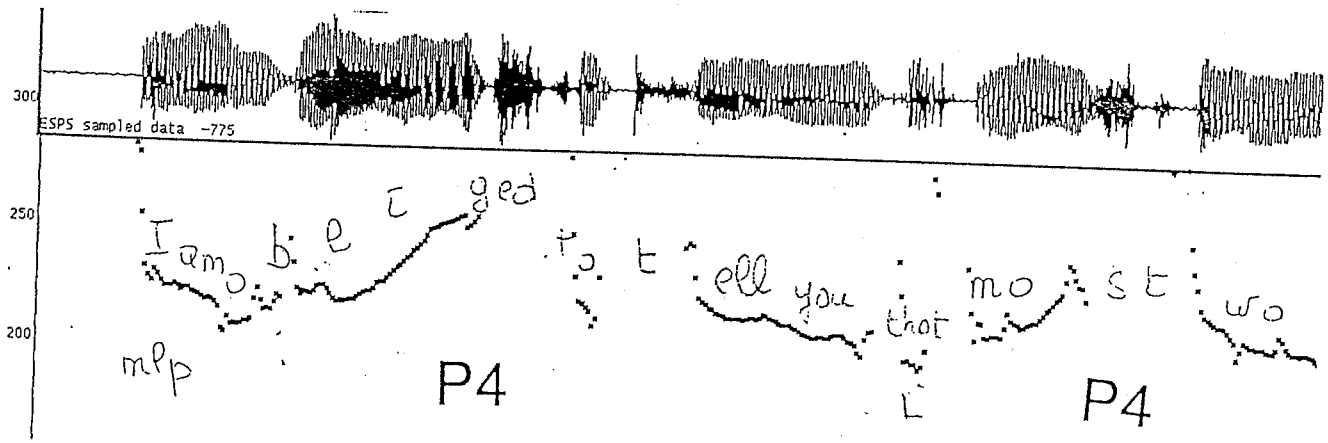

ESPS sampled data -37
200
150
100
The candidate
flet that the trapezoidal badge
would have more visual impact
that the usual
rosette
qsar

Fig.III.19

INTERSPEAKER VARIATIONS: SPEAKER PREFERRED PATTERN



Spk 1: P2 tendency
(BL+ Rise + Plateau + Rise on the plateau + Lowering+ BL)

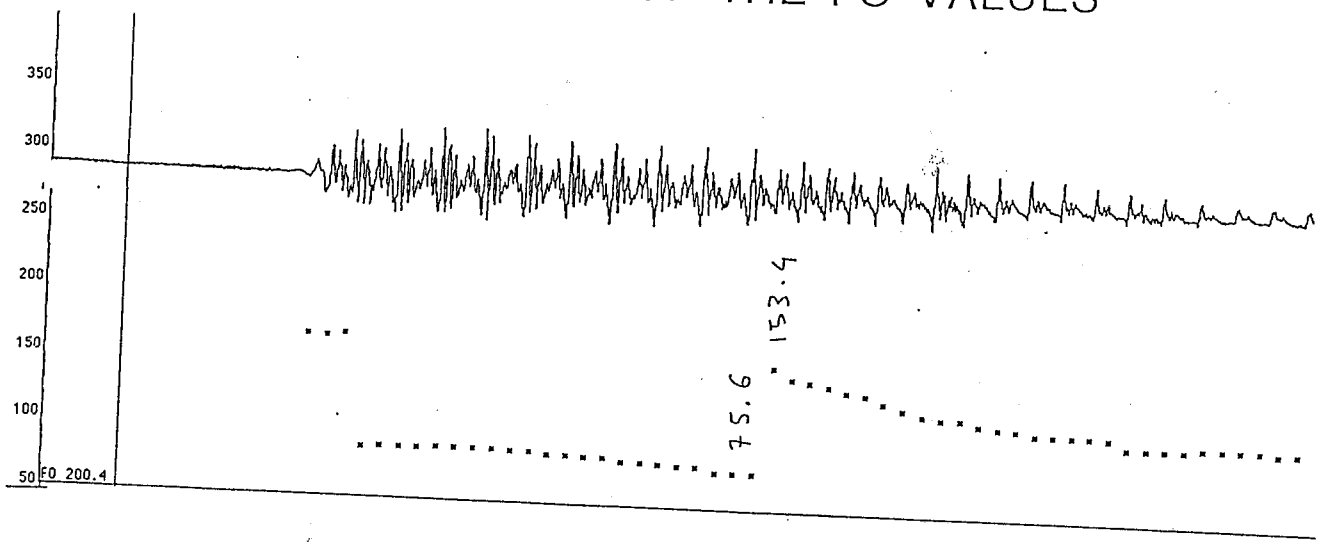


Spk 2: P4 tendency
(BL + Rise + Fall + BL)

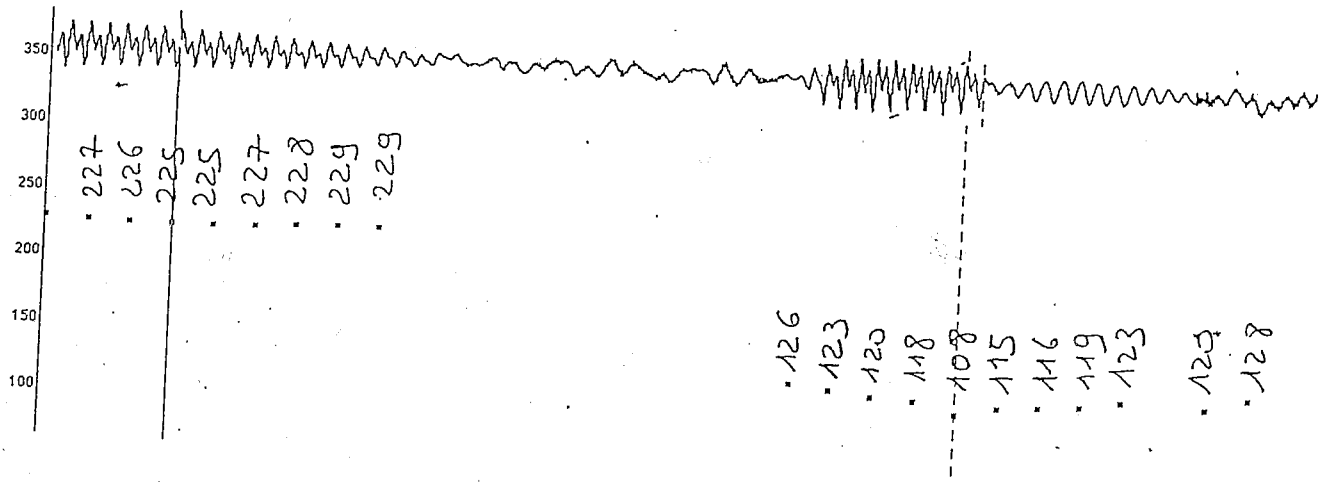
Fig.III.20

PITCH DETECTION ERRORS:

HALVING OF THE FO VALUES

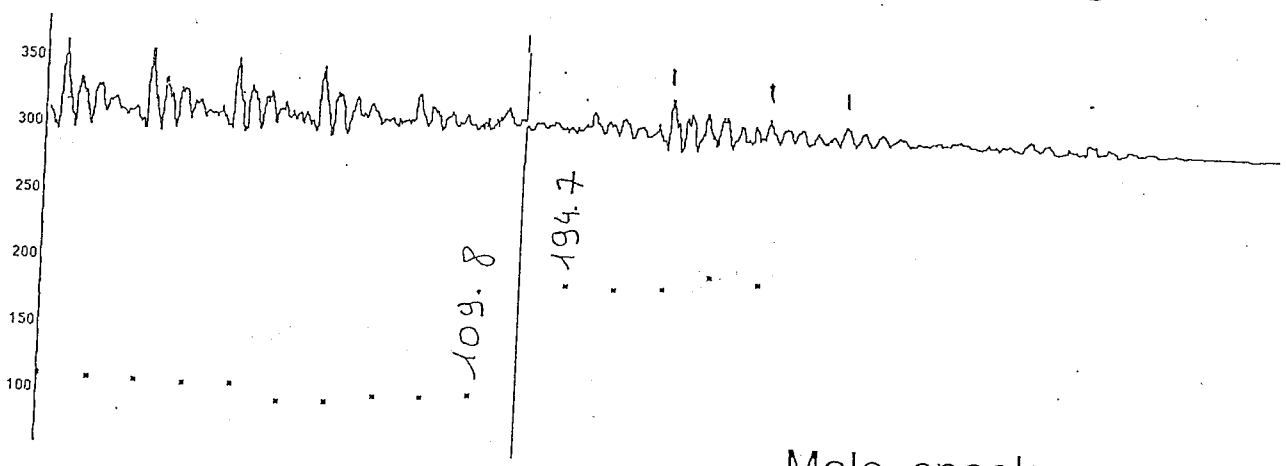


a) Male speaker gws



B) Female speaker sab

DOUBLING OF THE FO VALUES



Male speaker gsw

Fig.III.21