The use of prosodic parameters in automatic speech recognition

Jacqueline Vaissière

THE USE OF PROSODIC PARAMETERS IN AUTOMATIC SPEECH RECOGNITION

J. Vaissière
Centre National d'Etudes des Télécommunications
22301 Lannion, France

KEYWORDS.

prosody/ suprasegmental/ automatic speech recognition/ stress/ boundaries/ junctures/ segmentation/ phrase/ sentence type

ABSTRACT.

The present communication concerns the use of prosodic parameters in automatic speech recognition (ASR), i.e. the feasibility of automatically extracting prosodic information from a set of acoustic measurements done on the signal, and the incidence of integrating such information on the performance of ASR. Prosodic parameters include pauses and contrasts in pitch, duration and intensity between successive segments (mainly the vocalic parts). This notion is also extended to number of syllables and to ratios of voiced to unvoiced portions of the words. Part one introduces the various aspects of prosody (linguistic and non linguistic) and the main problems to be solved in automatically extracting linguistic messages conveyed by prosodic features. Part two deals with word level and lexical search: it presents work done (1) on the feasibility of word stress detection (primary stress, estimation of its magnitude, and evaluation of the complete word stress pattern) and (2) on the estimation of the amount of lexical constraints imposed by stress information in lexical search, completed by other suprasegmental information (number of syllables, word boundaries, ratios between voiced and unvoiced portion in the word, etc.). Part three deals with phrase and sentence levels and syntactic constraints provided by the automatic detection of word, phrase and sentence boundaries. Part four relates a number of miscellaneous uses at the phonemic level: phonetic segmentation, identification of the voicing feature of consonants, and estimation of the "segmental quality" of the underlying segments.

It is observed that prosodic parameters have been exploited rather poorly compared to segmental aspects of speech. Integration of prosodic knowledge with segmental knowledge in an ASR system is a difficult problem: to know when and where to integrate the prosodic knowledge into the system and how to combine the evidence and scores obtained from different sources. The exact contribution of the use of prosody in ASR is still to be estimated in an ASR system flexible enough to efficiently test such an integration.

NATO ASI Series, Vol. F46
Recent Advances in Speech Understanding and Dialog Systems
Edited by H. Niemann et al.

0. INTRODUCTION

In the early seventies there were a number of papers directed to specialists of automatic speech recognition (ASR), written by Wayne Lea [LEA 73a, 73b], pointing out the many possible uses of the prosodic parameters. As early as 1960 Lieberman showed that it is possible to automatically determine the stress pattern of bisyllabic words uttered in isolation with only 1 percent error [LIE 60]. As Lea suggested, prosodic parameters can be used at all levels in the decoding process of an ASR system: for example, at the **acoustic-phonetic** level to separate voiced from unvoiced stops, at the **lexical level** to detect word stress position, at the **syntactic level** to locate the principal phonological boundaries in sentences, and at the **pragmatic level** by locating the emphasized portion(s) of the discourse. Lea even suggested a prosodically guided speech understanding system and discussed some of the alternative strategies for using prosodics to aid speech recognition [LEA 75b].

Presently, despite the growing number of articles devoted to that subject in recent years, the effective use of prosody in ASR is rather limited. In particular, a number of papers have calculated the theoretical advantages of adding prosodic information to ASR systems, to restrict the lexical search and speed up the process of word retrieval in very-large-lexicon ASR. Studies have shown that automatic detection of word stress and main junctures in sentences is feasible with a very low error rate. The results obtained in terms of improvement of performance in ASR remain in most cases inconclusive for the following reasons: in some studies the process of extracting the basic prosodic parameters is not completely automatic, and automation is a necessary prerequisite for integration into an ASR system, at least in bottom-up systems; in others, the decision algorithms are manually run. More importantly, the value of integrating prosodically delivered information into strictly spectral information in isolated words and in continuous speech has not yet been tested in an ASR system flexible enough to combine different types of information and to run systematic tests.

However, it seems obvious that the use of all the information that can be automatically extracted from speech signals (including prosodic information) is necessary to achieve progress, particularly to ease the heavy constraints imposed by the existing recognizers: limited vocabulary size, restricted syntax, necessity to adapt the system to every new speaker, or obligation of the speaker to insert a pause between every word in the sentence.

As solicited by the organizer of this course, the purpose of this contribution is to provide a general survey of the use of prosodic knowledge in ASR. The first part includes introductory remarks on prosody which are relevant to the use of prosody in ASR. It aims at meeting the needs of readers wishing to quickly up to date about the state-of-the-art in this domain. The following parts provide a survey of the work already done in the field in English, French, Swedish, Italian and Japanese. Part two mainly concerns **word level**, specifically the detection of stress and the use of prosodic filtering in lexical search. Word level is the level at which most work has been done, particularly in the field of isolated-word systems. Part three deals with the detection of **phonological boundaries** in the continuum and part four summarizes some other uses at the phonemic level.

1. PART ONE: INTRODUCTORY REMARKS ON PROSODY

1.1 The four basic aspects of prosody

In a spoken sentence, the successive sounds (vowels and consonants) vary in pitch, duration and intensity. Such variations are conditioned by a number of factors which are reviewed succintly in the following paragraph.

a) Phonetically conditioned aspects intrinsic values and coarticulation

Part of the observed variations in pitch, duration and intensity depend on sounds in sequence and are often called " phonetically-conditioned aspects of prosodic parameters". These variations are conditioned by differences in the physiological mechanism involved in the production of **each** individual sound, on one hand, and by temporal, **coarticulatory** constraints due to the overlapping of articulatory gestures corresponding to phonemes in sequence, on the other hand. Let us examplified briefly the two types of phonetically-conditioned variations: intrinsic characteristics of the phonemes and contextual modifications.

1) Intrinsic characteristics of the phonemes have been rather well investigated. In particular, duration, fundamental frequency (Fo) and intensity of the vowels are known to be correlated with tongue height. High vowels (such as /i/), for example, have an intrinsically higher pitch, a shorter duration and an inherently lower intensity than low vowels (such as /a/). Nasal vowels in French are intrinsically longer than oral vowels. Tense vowels in English are intrinsically long and lax vowels short. Duration of a vowel is therefore correlated with the degree of vowel openness in French and English, the feature oral/nasal in French and the feature tense/lax in English.

2) The immediate context also has an influence. Vowels in unvoiced consonantal context (such as /p-p/) have a relatively higher pitch and are shorter in duration than the same vowels in voiced context (such as /b-b/). The influence of voicing on the duration of the preceding vowel is increased when both the vowel and the following consonant belong to the same syllable. Duration of a vowel therefore depends on syllable boundary location and the voice-voiceless feature of the following consonant [See Lehiste 's book on suprasegmentals for an excellent review of phonetically-

conditioned aspects of prosodic parameters, LEH 70). Such aspects are mainly speaker-independent and, to a great extent, independent of the language spoken. They may indirectly contribute to the identification of the underlying sounds, but do not have a linguistic function per se. In automatic decoding of prosodic information, it would be advantageous to normalize the prosodic parameters for such variations (see details later), but this can only followed identification of the underlying segments.

b) Linguistic aspects of prosody: syntax and rhythm

The part of variations in pitch, duration and intensity which is not conditioned by purely articulatory constraints is the carrier of linguistic messages. Those linguistically-motivated aspects of prosody are of utmost interest for automatic decoding of sentences. They carry information concerning **individual** linguistic units at various levels (on phonemes, but also on syllables, on word stress, accent and word boundaries, on phrase and sentence types and boundaries), on one hand, and concerning the acoustic **structuring** of each unit into larger units (phonemes into a syllable, syllables into a prosodic word, prosodic words into a sentence), on the other hand. These linguistically conditioned aspects of prosodic parameters and the prosodic structuring of each spoken sentence which are, for a given language, speaker-independent also seem language-independent to a certain extent (see VAI 83 for further references).

The basic problem is the following: prosodic parameters are at the same time governed by the **syntactic-semantic** organisation of the sentence (KLA 76), and also by **rhythmic** principles (LEH 80). There may be a substantial difference between the units as defined by syntactic-semantic constraints and the rhythmic units (FOW 77). As a consequence, both influences may become conflicting. Syntactic influence results in **varying** duration and Fo to contrast the units in sequence to mark boundaries. Rhythmic influences results in a tendency for the syllables and the stresses to **succeed at regular temporal intervals**. The segments duration are compressed or expanded to preserve the global duration of the successive syllables as invariant (isosyllabicity) and to equalize the time-interval between two stressed syllables (isochrony). At the same time, they are shortened or lengthened to mark the syntactic-semantic organisation. The coexistence of the three types of influence, (phonetically-conditioned aspects, rhythmic tendencies and syntactic-semantic organisation) is the main source of difficulties for interpreting the cause of a striking lengthening or shortening phonemenon (at the level of a syllable, a vowel or a consonant).

Rhythmic tendencies could be exploited in a predicting way, fo gaining information, for example, on syllable boundaries (Gerar Bailly, personal communication). The present use of prosody in AS mainly concerns its syntactic aspects, and exploits the constrasts in Fo and duration in a down-up fashion. One of the most interesting functions of prosody from an ASR point of view, is the grouping of semantically related words such as a noun and an adjective preceding or following it into a single so-called prosodic word and to express different degrees of junctures between successive prosodic words. Segmentation and structuring by prosody do not allow recovery of the complete syntactic structure of sentences in every day conversation. It may happen (i.e., when isolated sentences are pronounced in a neutral but careful way) that the prosodic structuring exactly maps out the syntactic structure, but in most cases There is not a one-to-one correspondance. In a sentence like "The dog likes the small cat", or "le chien aime le petit chat" (this tendency is language- independent), the main (and only) boundary may be placed either after the subject noun phrase, or the verb. Or there may be an equivalent boundary after the subject noun phrase and the verb. The adjective and the following noun are lileky to be grouped into a single prosodic group and no boundary is expected between these two lexical words.

As a result, prosodic information decoded from the signal should be compatible with lexical hypotheses made by the lexical search and with hypotheses made on the sequence of words by the syntactic module. Because it is necessary to provide for possible inter- and intra-speaker variability, algorithms for compatibility checking at the lexical and syntactic levels (in a verification process) are easier to conceive than algorithms for prediction (in a hypothetisation process).

c) Non-linguistic aspects: speaker's feeling about what he says

A sentence may also be uttered in a non-neutral, but more or less **marked** manner. The prosodic variations are then knowingly controlled by the speaker to eventually communicate knowledge about his attitude towards what he is saying: doubt, irony, involvement, conviction, etc... or to direct his listener's attention toward the **more important** words of his discourse. There is a continuum between completely neutral and marked manners of uttering a sentence. The acoustic correlates of "doubt" or "irony" marking are not yet well investigated, and, no need to say, the sentences to be prosodically decoded should be uttered in a rather neutral way (see the work done on automatic classification of Halliday's tones superimposed to the words "yes", "no", "mmm" and "well" , AIN 87). The system has to be however flexible enough to handle a certain amount of **emphasis** (such as the assignment of emphatic stress or **focal accent** to particular word(s) in the sentence). This aspect of prosody is particularly important in a **dialogue** where the speaker tends to emphasize key words to clarify a question or an answer.

d) Paralinguistic aspects: physiological differences and dialects

Furthermore, deviations from commonly observed patterns may contribute to inform the listener(s) **about the speaker** himself: pathological accent, foreign accent, emotional state, physical condition, etc... (Note that these are generally not under the control of the speaker).

The last two aspects (non-linguistic and paralinguistic) of prosody, which are speaker-dependent, may obscure the linguistically conditioned aspects. None of the actual achievements have tackled the problem of automatic adaptation to the "abnormal" particularities of a speaker. The existing systems expect speakers to differ very little from each other.

1.2 Interactions on prosodic parameters and normalisations

a) A specific function is carried on by a combination of all cues

As stated before, the **relative** contribution of phonetically-conditioned variations and of each of the many functions of prosody to the determination of the observed quantitative values of the three physical variables (Fo, duration and energy) is not easily determined. One specific function (such as stress marking or juncture marking) is generally not defined by a single prosodic parameter but by a combination of all prosodic cues: duration, intensity and Fo (and eventually by the insertion of a pause). Furthermore, the exact contribution of each parameter varies as a **function of the context**. Let us illustrate the complexity of the phenomena.

In a stressed language (such as English or Italian), the **phonologically stressed syllable** of a word tends to have a relatively longer duration, a higher (often rising) pitch and a greater intensity than the surrounding unstressed syllables. These are only tendencies. The criterion of a **longer duration** may fail, and the cases of failure are predictable.

First, the criterion of a longer duration may fail. In cases where the stressed vowel is an intrinsically short vowel (such as /I/ in English) surrounded by intrinsically long vowels, and/or it belongs to a syllable with a larger number of phonemes than the surrounding syllables (the larger the number of units at one linguistic level, the shorter the length of each unit), and/or it is located next to a syllable which has been lengthened because it is a word-final or a phrase-final syllable, etc..., the stressed vowel may be shorter than the surrounding vowels.

Second, the criterion of a **higher intensity** may also fail when intrinsically low intensity vowels (such as /i/ and /I/) are surrounded by inherently high intensity vowels (such as /a/). The criterion of a **integrated intensity** over the vowel fails when the vowel is too short.

Third, the criterion of a **higher Fo** may fail. A pitch movement is often present on each stressed syllable. Depending on the position of the word in the phrase, Fo may be mainly rising or falling during the stressed syllable (When two words are regrouped into a single prosodic word, there is a (language-independent) tendency for Fo to rise during the first word and to fall during the second; see illustration later). An Fo rise, which is often concommitent to a stressed syllable, may also be found on an unstressed syllable (such as the realisation of a continuation rise on the final vowel before a non-final pause). There is also a natural tendency for Fo values to rise at the beginning of sentence, then to decline (the so-called "declination line") and for the pitch range to diminish throughout the sentence: the relative height of the stressed syllables and the amplitude of the pitch movement are to be interpreted according to the position of the syllables in the sentence. The largest Fo movements are expected near the beginning; if not, a word located elsewhere than at the sentence beginning is emphasized. The lowest Fo values are expected at the end of the utterance; if not, the sentence is marked (it might be an interrogative sentence).

There are three consequences: first, a combination of at least three parameters (pitch, duration and intensity) is desirable to achieve more reliable decisions, even for isolated words; second, contextual rules are necessary to achieve reliable decisions in continuous speech; third, a prosodic event (such as a vowel lengthening or an Fo rise) will often receive more than one interpretation.

b) Normalisation by articulatory and perceptual considerations

As expected, some algorithms attempt to substract the effect of **phonetically-conditioned** variations. Such normalisation requires identification of the sounds and possibly the position of syllable boundaries.

Ideally, compensation includes:

1. the intrinsic characteristics of the phonemes and the influence of the surrounding phonemes.
2. the number of phonemes in the syllable, of syllables in the word, and of words in the sentence.
3. the speech rate;
4. the correction for prepausal lengthening.

Present systems only include partial compensation of the phonetically-conditioned variations, i.e. of the ones which are the easiest to integrate. It is not clear whether or not such partial normalisation is better than no compensation at all (there is no known comparative studies). Indeed, complete compensations can be carried out only in a verification process. Details on compensation in present systems are described in Part II.

It has often been said (LEA 80c) that the raw data for prosodic decoding should not only be corrected for the phonetically-conditioned aspects but should also take into account knowledge about **speech perception**. It is evident that such a "normalisation" by perceptual considerations should lead to a more integrated view of the relative contribution of the three basic prosodic parameters, duration, Fo and intensity. However, knowledge on how duration, intensity and Fo are perceived and processed in continuous speech is not advanced enough to be applicable in an ASR system.

1.3 The role of prosody in speech decoding

a) Is prosody used by human listeners?

It is often argued that ASR systems should parallel, to a certain extent, the way humans perceive speech to be successful. One may question the real use of prosodic parameters by human listeners.
Do listeners use prosodic cues at all? How do they use them and when? And how far do the prosodic cues contribute to the understanding of the whole message?

Despite the apparent complex manifestation of the organisation of prosodic parameters, listeners seem to have no difficulties in decoding prosodically-carried information, at least in controlled experiments. The effective use of prosodic parameters in every day conversation and its exact contribution to the complete decoding of sentences are rather difficult to test. Despite the fact that it is easy to invent pairs of sentences which can only be made unambiguous by a single prosodic event (the position of a juncture or of a word stress), the cases where prosody cues are vital for comprehension of the message are rather rare in every day conversation. Because of that, prosodic aspects are often said to be **redundant** with spectral aspects, making the assumption that spectral aspects are sufficient for uniquely decoding the sentence. The assumption of redundancy may be questionned. It has been shown that listeners pay attention to prosodic continuity at the expense of semantic continuity (DAR 75). Nevertheless overall intelligibility systematically declines with an increased degree of time-compression; sentences heard in normal intonation are significantly more capable of withstanding the debilitating effects of compression than those heard in anamalous intonation (WIN 75). Understanding speech synthesis when there is no control of the prosodic parameters requires a particular effort from the listener. Deciphering a spectrogram without using temporal cues and Fo contour is much harder than using all cues available (personal experiment). It is equally reasonable to assume that not only the spectral aspects but also the temporal aspects and Fo are parallel inputs to the ears of the listener who simultaneously decodes them, and that a simultaneous treatment of both inputs might be absolutely necessary for automatic recognition of continuous speech. Such view is generally accepted in the psycho-linguistic literature.

b) The particular role of word stress

The exact part played by word stress in continuous speech processed by listeners is not clearly established. Nevertheless, number of studies have investigated the potential use of stressed syllable detection in an ASR system. The arguments are the following:

- 1) the position of word stress is **strictly necessary** to the differentiation of words uttered in isolation. Some words such as noun/verb pair (PERmit-perMIT) are distinguished from each other almost entirely on the basic of stress position (in English or Italian, for example). [In languages with fixed word stress position like French, syllables found as "stressed" by prosodic considerations, i.e. a longer duration and an higher Fo, may not only correspond to a phonological stress on the last word syllable, but also to emphasis at word onset].

- 2) in English stressed syllables seem to represent **islands of reliability** where the acoustic cues are relatively more robust. Experiments in spectrogram reading by experts and analysis of the phonetic confusions made by automatic acoustic-phonetic decoders confirm the fact that stressed vowels are more easily recognized than unstressed vowels (LEA 75a). Easier recognition is often explained by the fact that in word stress the syllables are lengthened. [Such an observation is compared to the notion of acoustic dominancy in French. In French, the "dominant" consonants, i.e. consonants which are located in portions of the sentence where Fo is rising, have characteristics closer to the norm than the same consonants in other contexts (VAI 86). Note that phoneme dominancy is an acoustic notion and it is determined by the actual position of phonemes in the prosodic structuring of the sentence, while stress is a phonological notion].

- 3) a number of theorical studies have demonstrated the usefulness of stress detection for a **multi-level access** to the lexicon. It has been shown that stress pattern information can reduce the lexical search in English and Italian. For **English**, dividing phonemes in stressed syllables into broad phonetic categories (manner of articulation) while "wild-carding' the unstressed syllables is almost as restrictive as representing the vocabulary by six broad phonetic classes (HUT 84; see also CAR 87). Stress information make thus restrictions that are potentially very useful for large-vocabulary, isolated-word speech recognition. Using a 15,000 word lexicon, and assuming a three level convention for the description of each syllable (stressed, unstressed and reduced) and a single stressed syllable per word, Aull showed that knowledge of only the number of syllables of the word yields an expected class size equal to approximately 41 percent of the size of the lexicon. When the stress pattern is known (i.e. the

correct number of syllables and the correct assignment of stress for each syllable), the expected class size is reduced to 19 percent. If the number of syllables is known as well as the location of the stressed syllable in the word, then the expected class size is 22 percent [AUL 84]. The charateristics of lexical stress and their possible use in automatic speech recognition of the Italian language have been investigated at GSELT. A theoretical analysis of the constraints imposed by stress in the strategy of a large lexicon was conducted on the basis of a 12,000 word vocabulary: knowledge of the number of syllables and the stress location potentially allows reduction of the cohort size to 4.3 percent of the entire lexicon [PIE 87].

c) The basic function of alterning Fo rises and falls

Perhaps because a Fo rise is mainly realized by tensing the vocal folds and Fo fall by relaxing them, an upward change of Fo between two successive vowels seems to be associated with the notion of beginning and a downward movement with the notion of end in a number of languages (see VAI 83 for references; see also AIN 86, LIN 83). As a first result, a decrease in Fo usually occurs at the end of each major syntactic constituent and an increase in Fo occurs near the beginning of the following constituent (the so-called **fall-rise** pattern used as a boundary marker [disjuncture phenomena] (COO 77, LEA 72, LEA 80). As a second result, rise and fall often appear as a pair, and thus the schematized Fo pattern exhibits a so-called **hat-pattern** (MAE 76 for English) used for word grouping (juncture phenomena). In English, Fo movements are limited to the stressed syllables; in French, to word boundaries.

As noted before, there is more than one way to utter a given sentence. There are four basic ways to structure prosodically the following declarative:

[1]

[2]

[3]

[4]

(the pattern are schematized according to Maeda's method; see MAE 76 for further examples). The number of rises and falls depends on speech rate. As noted before, the number of large Fo movements decreases as rate of speech increases. [1] is more likely to correspond to slow speech, and [4] to rapid speech. Due to rhythmic constraints, there is also a tendency for the long words to correspond to a complete pattern, and for short words to be regrouped into a single pattern: a boundary (marked by a fall-rise

pattern) between the subject noun phrase and the verb is more likely to happen when the subject noun phrase is long (in terms of number of syllables). As a consequence, for a given sentence and a given rate of speech, a certain type of prosodic patterning is more likely to occur than the others. It is however necessary to integrate the possibility of differences between speakers into the system.

d) Speaker-dependent variations

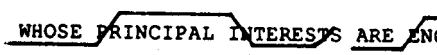
The same basic principles govern the determination of the prosodic parameters for all speakers, but the relative contribution of phonetically-conditioned variations, stress, rhythm, syntax, semantic, style and rate is speaker-dependent. Each speaker tends to be consistent, but there are regular differences between speakers. The study of speaker-dependent regularities is very important for ASR: the tuning of the system to the particular habits of the speaker is a prerequisite for a full extraction of the information.

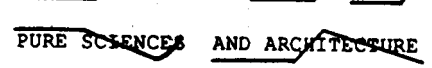
Let us give some typical examples. I have compared the prosodic "habits" of two native speakers of English, JA and KNS, while reading long texts. The following represents the schematized Fo contours (schematisation is done using Maeda's method, MAE 76) for the first sentence of the first text (unpublished data). Only the Fo movements from the "plateau" to the "base-line" and vice-versa are schematized.

Schematized patterns for speaker JA

[5]

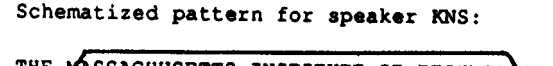
[6]





[8]

Schematized pattern for speaker KNS:



[9]

[10]

[11]

[12]

There are at least four points for which the two speakers differ.
1) The **Fo falls** for JA are always rapid (that is superimposed to a single syllable) and concommitent to a stressed syllable (primary or secondary) (excepted at sentence end, where the fall is superimposed to the last syllables)(see Fo falls in [5] to [8]). In contrast, for KNS, the falls are rapid or gradual (see "coeducational" in [10], "principal" in [11]). Rapid falls are superimposed either to a stressed syllable or to a function word (a regularity observed in the remaining parts of the texts).

- 2) For JA, there may be a fall in a stressed vowel, directly followed by a Fo rise marking continuation in the next (unstressed) syllable (see "sciences", [3] and [4]). For KNS, when a continuation word is used on the word's last syllable, the fall on the preceding stressed vowel is always suppressed (see "sciences" in [10] and [11]).

- 3) For JA, there was no lexical word without at least an Fo movement superimposed on it, unless than it is uttered entirely in the upper Fo register (see "institute" [5]). For KNS, the lack of Fo movement on a lexical word mark its dependency with the surrounding lexical word (see the following paragraph).

- 4) For JA, adjective and following nouns are always regrouped into a single prosody word (by a rise during the first word and a fall during the last): "independant institution" in [6] , "principal interest" in [7], "engineering sciences" in [7], "pure sciences" in [8], etc... For KNS, adjective and following noun may be regrouped in the same way as used by the first speaker ("independent institution"), but in most cases, there is a rise in the first word, followed by a gradual fall, and there may be no Fo movement on the last word.

Such speaker differences are very important to grasp by rules. Such rules are a prerequisite for a full use of prosodic information in ASR. There is a need of more work in that direction (there is no known systematic work on modelling intra- and extra-speaker differences).

e) **The use of pauses as boundary marker**

The acoustic analysis of speech shows that speakers insert a large number of pauses while talking. **Respiratory pauses** represent only a part of all pauses; there are also **hesitation pauses** , but the majority of pauses (including the respiratory pauses) are located at **grammatical junctures**. The use of pauses as major boundary markers between and within sentences seem to be similar across those languages for which there is available data.

f) **The use of lengthening as boundary marker**

In a number of languages (English, German, French, Spanish, Italian, Russian and Swedish) there is a tendancy to lengthen the final elements in the linguistic structure, particularly the last vowel before a pause, as well as the final elements in words and phrases (see VAI 83 for references). It seems that lengthening, like Fo fall, is associated mainly with the notion of termination.

1.4 Conclusion

The preceding remarks on prosody are far from of being sufficient to introduce the complexity and richness of information carried by prosodic parameters. Articles concerning descriptive analysis of prosodic parameters corresponding to different speech styles, to different speakers and to different dialects, automatic generation of prosodic parameters for automatic synthesis-by-rules program, the rapidly developping area of non-linear phonology, psycho-acoustic and psycho-linguistic studies on the perception of pitch, duration and intensity in continuous speech; the articles on automatic segmentation of speech and on fundamental frequency detection, etc..., may be of utmost importance to those interested in adding prosodic parameters to their system. The proceedings of a seminar devoted to the role of prosody in ASR may also be of some interest [DIC 82, in French).

In modelling prosody for integration into an ASR system, it is essential to keep in mind the following points:

- (1) the speaker is expected to speak in a more or less **neutral way**
- (2) there is **more than one** neutral way to utter a single sentence although the number of ways is still limited;
- (3) the acoustic evidence of secondary marked junctures decreases as the **speech rate** increases: roughly speaking, the amount of prosodic information is inversely proportional to the number of syllables per second;
- (4) **vocalic** portions of the signal play a prominent part and are the main carriers of temporal and Fo contrasts; however, contrasting **syllable** duration and relative lengthening of **consonants** within a syllable are carriers of relevant information (word-initial marking for example),
- (5) **the binary division** between long and short vowels, or long and short syllables or stressed and unstressed syllables, often used to describe the perception of prosodic parameters is not entirely adequate in an ASR system; in continuous speech there is **continuous** variation in values,
- (6) **Adaptation to the speaker's** particular patterns is very useful to fully extract the information contained in continuous speech; Without adaptation, less information can be extracted and finally

- (7) there are some speaker differences in the amount of prosodic information that can be decoded from the signal (at least manually using present knowledge): sentences spoken by some speakers are more easily deciphered prosodically than sentences spoken by others, independently of the speech rate;

The following sections describe specific work done in the field. There are a large number of papers more or less related to the use of prosody in ASR in French and in English. Only those works where prosodic features are automatically extracted from speech signals are cited. As for the other languages (Swedish, Japanese, Italian), all papers known to the autor are cited (See also Noeth'es paper in the present volume, NOE 88).

2. PART TWO: STRESS DETECTION, STRESS MAGNITUDE AND PROSODIC FILTERING

A number of studies have been devoted to the role of stress in ASR systems. Despite differences, the systems follow approximately the same procedure. Thus the problem of detecting stressed syllables in isolated words or continuous speech can be divided into five steps:

- 1) segmentation of the speech wave into syllable-like units,
- 2) extraction of the prosodic features,
- 3) normalisation of the features,
- 4) detection of stressed syllables and
- 5) testing the effectiveness of the delivered information in a complete ASR system.

2.1 Syllabic nuclei detection

As noted before, most programs start by detecting the syllabic nucleus to locate the vocalic portions of the signal. A typical program of that kind, written by Mermelstein, can detect over 92 percent of the syllabic nuclei. The program uses a spectrally-weighted "loudness" function (the intensity of the speech wave is in the frequency range 500 to 4000 Hz) to compensate for different intrinsic energies in various vowels, and an interactive application of a "hull function" to first detect large energy dips then subdivide long segments between those big dips, by locating smaller dips within the chunks (MER 75). Segmentation can also be carried out by using the phonetic lattice produced by the acoustic-phonetic analyzer as in the French system, KEAL, used in my own studies (see the article by Mercier & al in the present book, MER 88) or the output of a broad classifier (such as in Aull's work). Interestingly enough, the segmentation errors are very similar, independently of the system used and of the language concerned. Segmentation is mainly based on abrupt discontinuity. Algorithms fail when segmental parameters are changing slowly over time. They fails often when two vowels or more are in sequence (vocalic linking) and when there is a problem with the separation of vowels with adjacent sonorants (/l/ and /r/). The problem is particularly acute when the vowels have a low first formant.

2.2 Features detection

Once the frontiers of the syllabic nuclei have been detected, the next step consists of calculating a certain number of parameters for each region (up to five, depending on the system). The calculated parameters are the following:

- 1) the **duration** of the sonorant portion of the segment or the duration between syllable boundaries defined by the onset of the sonorants of consecutive syllables [WAI 86];
- 2) the **energy integral** (average of energy level over the vowel);
- 3) the **fundamental frequency**, generally the pitch maximum [AUL 84; WAI 86]
- 4) a measure of **spectral change** for English (AUL, 84; WAI 86);

2.3 Normalisation of the features

A number of normalisations of the raw parameters can be made. Some of the systems include:

- 1) compensation for the **intrinsic** characteristic of the vowels: The average energy level is multiplied by a correction factor depending on the intrinsic intensity of the vowel when the vowel is identified (carried out by the acoustic-phonetic decoder or found in a dictionary when the word is known); in case the detailed identity of the vowel is not known, duration and pitch of the vowels are modified depending on an estimation of the first formant of the vowel (NIS 82; MEL 82) and the evidence of nasality (MEL 82). The values for the correction factors are often taken from LEH 60 for English and from ROS 67 for French (the tendencies are language-independent).

- 2) compensation for the **lengthening-before-voicing** phenomenon: Automatic detection of the voicing feature of the consonant following the vowel is relatively feasible, but a unique correction factor is hardly adequate in continuous speech. The magnitude of the effect varies as a function of the tense-lax feature of the vowel, the nature of the following consonant (stop or fricative), the position of the word and syllable boundaries and the position of pauses (CRY 85).

3) correction for **prepausal lengthening.**

Such a correction is currently used in isolated words where the position of the pause is known (note that in this case, the correction includes not only prepausal lengthening but also the combined effect of word final lengthening); Dumuchel for example used a fixed factor (DUM 86), while in Aull's algorithm, the correction factor is adjusted depending on the gross category of the final context (AUL 84).

2.4 Algorithms for stress detection

2.4.1 Stress detection in isolated words

a) The largest number of studies concerns stress detection for **English**. Medress & al (cited in LEA 80) showed that while 72 percent of stressed syllables in isolated words could be detected from higher peak energy levels, 70 percent and 68 percent could be detected from syllabic nucleus durations and Fo peaks, respectively. Cheung et al (CHE 77) presents an automatic method which estimates the magnitude of syllabic stress on a continuous scale in continuous speech using a composite of three acoustic parameters: F0, intensity and duration. Results showed that perceptual stress correlates highest with F0(r=0.683), then intensity (r=0.495) and lastly duration (r=0.306) and that a combination of the three cues provides a close approximation of the perceptual data (r= 0.876). In Aull's training data (350 words, 7 speakers), the most stressed syllable has the longest duration 65 percent of the time, the maximum amount of energy 84 percent of the time, and the maximum in Fo 76 percent of the time. Aull made two interesting observations: first, by reducing the duration of the final syllable to account for prepausal lengthening the maximally stressed syllable has the longest duration 90 percent of the time; second, the measurements were nearly as effective in separating stressed syllables from unstressed syllables if sonorants adjacent to the vowels were also included. The latter observation is important in automatic stress determination since it is often difficult to separate sonorants from adjacent vowels automatically (AUL 84). In Aull's final system for recognizing lexical stress patterns from speech signals, duration (compensated for prepausal lengthening), logarithms of the average energy measure in two frequency bands (400-5000Hz; 1200-3300Hz), maximum Fo on the syllable, and a measure of spectral change are computed for each sonorant syllable. The assignment of stress is made on a relative comparison of the syllables. In her study of 1,600 isolated words, Aull has shown that 98 percent of the primary stresses can be detected. For the remaining 2 percent that are mislabelled, nearly 40 percent of the labelling errors is due to the front-end processor. In 30 percent of the labeling errors, the stressed syllable is marked as an alternative choice for stress. The correct number of syllables and the correct position of the stressed syllable were found in 90 percent of the cases. The entire stress pattern (number of syllables, position of the stressed syllable and correct labeling of the remaining syllables as unstressed or reduced) was estimated in 87 percent of the cases (AUL 84). Dumuchel (DUM 86) examined the contribution of stress information in a HMM-based large vocabulary recognition system. The duration of the vowel part (normalized for lengthening-before-voicing phenomena and for prepausal lengthening) and the average energy level of the vowel (normalized for intrinsic energy) were used for estimating the probability of the correctness of the estimated stress pattern. After an initial training phase,

tests on a new word list yielded 95 percent correct detection of the syllable carrying the primary stress. During word recognition, the likelihood of each word derived from acoustic data is modified by the probability that the required lexical stress pattern is supported by observed data. The rank of the correct word in the word hypothesis list improves by an average of 0.3 word positions when using the stress information. Excluding the two thirds of the list where the correct word was already ranked first, the improvement amounts to an average of 0.9 word positions. Dumuchel's experiments showed a (minor) improvement in the average word position in an HMM system by using stress information (DUM 86).

b) Some work has also be conducted for **Italian** An investigation into several thousands of words spoken by different speakers was made in order to extract the statistical properties of the main stress correlates in Italian. The results suggest that in Italian duration has the most importance for stress determination (after correction of prepausal lengthening, duration has a stress relevance of about 91 percent. PIE 87; see also KOR 87). By using durations of vowels, average log-energy of vowels, maximum Fo within vowels, and average spectral change (Euclidian distance between consecutive cepstral patterns), over 96 percent of stressed syllables are well detected in isolated words (PIE 87).

2.4.2 Stress detection in continuous speech

The particular problems in stress detection for continuous sentences are the following: first, the word boundaries are not known, and it is not possible to compensate for word-final lengthening; second, the grammatical, rhythmic and pragmatic aspects of prosody interact with word prosody; third, the words are not equally emphasized and the lexically stressed syllable may be not heard as "stressed" by listeners.

a) Lea and Waibel have proposed algorithms for stress detection in continuous speech for **English**.

Lea (LEA 73a) has shown that 89 percent of the syllables heard by listeners as stressed in continuous speech can be automatically detected. The procedure locates stressed syllables from an "archetype algorithm" that uses increases in Fo and energy integrals (with 21 percent false stress assignments). If durations of nuclei alone are used, 84 percent are located with 32 percent false alarms; by using Fo rise only, 77 percent are detected with 24 percent false alarms.

Waibel adopted a pattern-recognition approach to optimally combine intensity, duration, pitch and spectral changes into one minimum-error stressed syllable classifier. When a forced decision is imposed by setting a threshold at stress probability 0.5, error

rates of 7.79% to 14.85% percent missed stresses were obtained Waibel concludes that amplitude integrals are the strongest predic tors of English stress in continuous speech (WAI 86).

b) The automatic detectability of syllables heard by listeners as stressed seems to be similar for English and for French. In a limited study, Martin (MAR 77) has shown that in **French** , ranking in continuous speech of the syllabic nuclei into decreasing integrated intensity is effective in 81 percent of cases in detecting syllables heard as stressed by a panel of listeners.

c) In applying the previously described GSELT's system to 51 **Italian** sentences, over 95 percent of the stressed syllables are detected but also about 16 percent of false stress (PIE 86).

d) A joint research project shared by the Phonetic Institutes at the Universities of Lund and Stockholm, "Prosodic Parsing for **Swedish** Recognition" has just started. The algorithms that are beginning to emerge are intrinsically more complex that the algorithms developed for stressed languages like English and Italian, simply because Swedish prosody seems to be of a more complex nature. In addition to the basic distinction between stressed and unstressed syllables, the primary stressed syllable is characterized by having one of two tonal accents: acute and grave. The Fo representation of initial juncture bears a strong resemblance to that of an acute focal accent, while the representation of a final juncture resembles that of a grave word accent. House et al (HOU 87), however show a promising 81 percent correct detection rate for grave accents.

2.5 Lexical filtering by suprasegmental patterns and prosodic cues

There are very few known studies on lexical filtering estimating the real contribution of the introduction of prosodic information at the word level in an existing ASR.

a) In a study by Waibel (WAI 87) on **English** , the suprasegmental features exploited were temporal cues (syllable ratios, ratios of unvoiced segment durations to syllable durations, voiced segment durations), intensity profiles and stress likelihoods. Using a multispeaker continuous-speech data base for evaluation, each suprasegmental feature is shown to hypothesize the correct word substantially better than chance. All suprasegmental features were then combined and compared with a speaker-independent acoustic-phonetic word hypothesizer. After applying the suprasegmental information, the correct word ranked on average 25th out of 252 words. The acoustic-phonetic knowledge alone yielded an average rank of 40 (out of 252) without the addition of suprasegmental information. After suprasegmental and phonetic KSs were combined the average rank was reduced to 15 out of 252. The results indicate that suprasegmental information indeed adds complementary information that substantially

improves word hypothesization in speaker-independent continuous speech recognition.

b) One of my own studies (VAI 76, 82) has shown that about 60 percent of the lexical words in **French** are marked by an Fo rise at their beginning and 70 percent by an Fo fall at the end. Ninety-six cent of Fo falls occur on word-final syllables and never on word-initial syllables. Ninety-six percent of Fo rises occur at word-boundary syllables, either the first syllable in the word (word-initial rise) or the last syllable (a manifestation of continuation rise on the word-final syllable). Taking into account the fact that an Fo rise can only happen on the last or the first syllable in a word, the system was able to eliminate 33 percent of the word possibilities (consisting of more than two syllables) proposed by the lexical module from spectral knowledge only, leading to an appreciable reduction of the syntactic load (VIV 77). Only rather long words were suppressed, but no correct word was eliminated.

3. PART THREE: DETECTION OF SYNTACTIC BOUNDARIES

As seen before, prosodic knowledge provides a form of constraint on each word. Juncture phonemena (also called boundary phenomena) carry additional constraints on word sequences. While stress detection concerns both isolated-words and continuous speech systems, word- and phrase- boundary detection relates to continuous speech only. The basic prosodic features used for boundary detection are pauses, Fo typical movements and lengthening.

3.1 Pauses

O'Malley and his colleagues (OMA 73) have lead a study of the relationship between syntax and the position of pauses in spoken **algebraic** expressions. It was found that subjects were very consistent in their placement of pauses when reading algebraic expressions slowly. Furthermore, there was an almost perfect correlation between measured silence and perceived juncture. Rules were developed for inserting parentheses based on the location and measured duration of silence intervals in an utterance. Listeners were asked to insert parentheses given the spoken form, and the consistency of their answers was measured by a chi-square test. For those cases where there was listener agreement in a single answer, the rules were tested and found to agree with the listeners form 91 to 95 percent of the time.

In a study, Lea showed that 95 percent of the clauses and sentence boundaries are marked by a pause of 350 milliseconds or more (LEA 72).

3.2 Fo rises and falls

In one of his studies on cues to constituent structure and sentence boundaries in English, Lea (LEA 72) showed that 94 percent of the syntactic boundaries of 500 sentences were marked by a fall-rise

pattern in the pitch contour. He also noticed that sentence boundaries were always accompanied by fall-rise Fo contours. The program searches for substantial decreases (7 percent) in Fo followed by substantial increases , and marks a boundary at the last of the lowest Fo values in the valley. A computer program (LEA 73b) correctly detected over 80 percent of all syntactically predicted boundaries.

3.3 Pauses and Fo movements

It is interesting to combine the information given by the position of pauses and the Fo contour.

Komatsu and his colleagues (KOM 86) recently presented an algorithm for detecting boundaries between grammatical units and for formulating structural hypotheses in conversational speech (the task of a PBX telephone operator) using Fo contour and the location of pauses. Pauses and Fo histogram are first used to detect sentence boundaries. Then, within a sentence, Fo contour is analyzed in detail by means of a piecewise linear approximation with a sequence of linear lines. A number of solutions is proposed for each sentence. The preliminary results are encouraging. Their analysis relies on Fujisaki's and co-workers' findings on Fo contours in Japanese. According to Fujisaki (FUJ 87), the Fo contour is composed of accent components and phrase accents. There already exists a fairly detailed account of the interaction between both components, but there is no known attempt as yet to automatically detect word accents in Japanese.

3.4 Lengthening

Lea (LEA 75a) estimated how far phrase boundaries can be detected from a measure of " phrase-final lengthening phenomena". It was found that 91 percent of all phrase boundaries heard by listeners in spoken sentences can be detected by finding vowels and sonorant consonants that are at least 20 percent above the median lengths of the same vowel or sonorant in all its speech occurences. The duration of interstress intervals tends to be short only when a word boundary intervenes but increasingly longer for syntactically predicted boundaries between phrases, clauses and sentences. He also noted that the phonetic error rate is more readily predictable from the interstress interval than from other measures.

Phrase-final lengthening affects both the duration of the successive vowels and syllables. In a previous study, I compared the use of syllable duration (automatically detected) with the use of the vocalic part alone. There was no clear advantage of one method over the other. In 97 percent of the cases, syllables (or vowels) detected as long by the program corresponded to first or last syllable of a lexical word and to monosyllabic grammatical words (VAI 77).

3.5 Fo movements and lengthening

Combining Fo and duration information is necessary to improve recovering of prosodic information. By using both duration and Fo in each vowel and a set of heuristic rules for detecting the main boundaries in sentences uttered without internal pauses in continuous French, a program was able to detect a main boundary in 86 percent of sentences (5 percent false alarms, most of them due to segmentation errors) (VAI 80).

Main boundary detection is probably the easiest to obtain automatically. To extract further information from duration and Fo contour and to determine clause, phrase and word boundaries , more complicated rules are necessary. PROSEIDON is an expert system developped for chunking the continuum into prosodic words [A prosodic word can contain one or several lexical words, preceded or not by grammatical words], for giving information about sentence type , position of the main boundary, left and right word boundaries and dominant and dominated syllables for a given sentence whenever possible. Connected to the acoustic-phonetic decoder (providing segmentation of the continuum into phoneme-size and syllable-size segments), the prosodic module uses both down-up and bottom-up informations. In particular, the number of syllables and a measurement of speech rate determines the number of boundaries the module is expected to find (about one boundary for four syllables). The main boundary is first detected, and the continuum is chunked into two parts. Extra boundaries in the each part are detected then, taking into account the number and the position of the segment relatively to the main boundary and the surrounding pauses. Figure 1 illustrates the input data to the phonetic module. Visualisation of the parameters computed directly from the signal is important for the expert who elaborates the rules. A "prosodic transcription" of the sentence to be interpreted is done automatically from (1) its temporal aspects (derived from calculation on relative length of the successive vocalic nuclei and syllables) and (2) its melodic aspects (derived from calculation on Fo relative height between vocalic nuclei and on Fo contour in lengthened vocalic parts). The principle on which the program is based is to interpret concommitently the rhythmic (temporal) aspects and its melodic (Fo) aspects. In a preliminary experiment, the program was able to give an average of 2.8 pieces of prosodically extracted information per sentence and to segment the continuum into chunks containing an average of 1.3 lexical words with an error rate of 3 percent (VAI 84). For example, for the sentence displayed in Fig.1, the running of the prosodic program PROSEIDON delivered the following information: (1) declarative sentence, (2) main boundary (//) on the fifth syllable (in fact, the offset of the subject noun phrase) and secondary boundary on the eigth syllable (the offset of the verb). The sentence was divided into three prosodic words (no error). The main weakness of the prosodic module is that it works in a top-down fashion starting calculation when the end of the

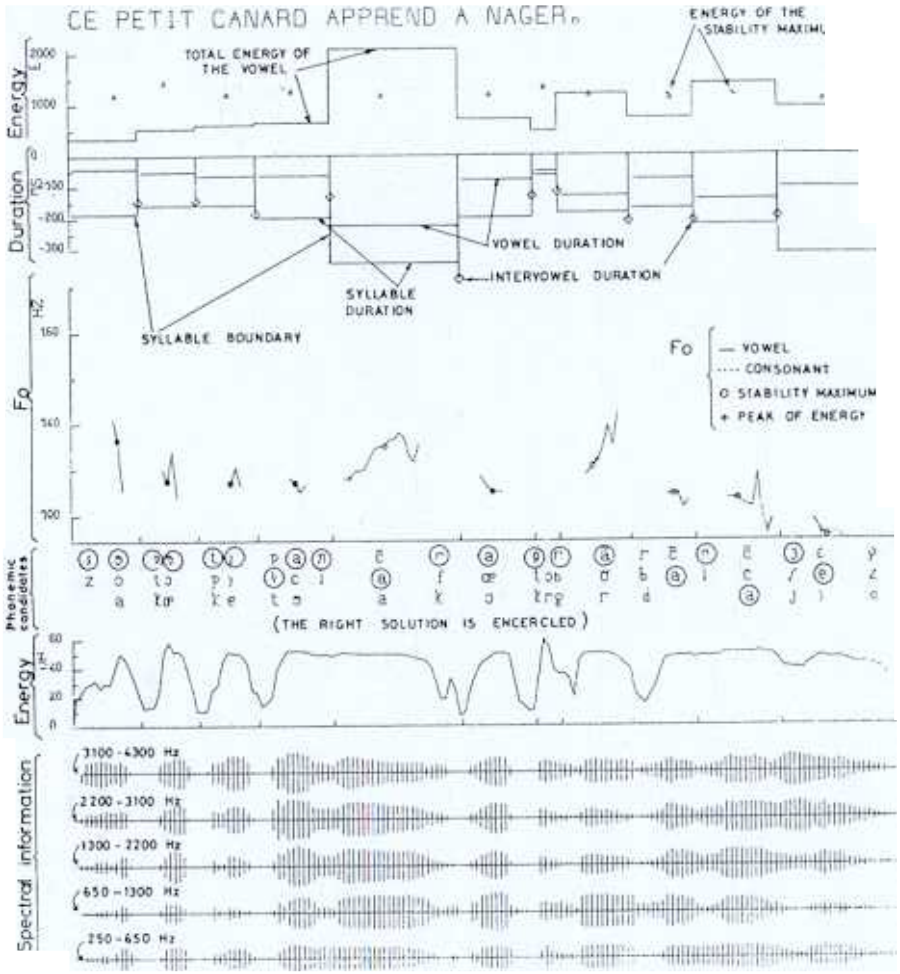


Figure 1: Visualisation of the prosodic parameters

This figure shows one possible output for the sentence :" Ce petit canard apprend à nager." (The little duck learns how to swim). It displays, from down to top,
1) SPECTRAL INFORMATION: energies in five frequency bands (bottom),
2) ENERGY,
3) PHONEMIC CANDIDATES: the three first candidates proposed by the phonemic analyzer,
4) MELODIC INFORMATION: Fo contour (+ indicates the selected value at the most stable part and the empty circle the selected value at the maximum of energy of each vowel),
5) RHYTMIC INFORMATION: the computed duration of the successive vowels and syllables, (vertical lines indicate syllable boundaries),
6) INTENSITY INFORMATION: the integrated energy over the vowel, and the intensity at the most stable part of the vowel.

sentence is reached. I am presently working on a new version of this program going strictly from left to right in order to be useful in a real time application by chunking incoming speech into logical sense-groups.

3.6 Integration of the information about boundaries in an ASR

Allmost no work is done yet in that direction. An attempt has been made to use Fo-detected phrase boundaries to aid the syntactic parser in the BBN HWIM system (LEA 80). This procedure involved marking first all the words in the grammar that were expected to be immediately preceded by Fo valleys (i.e., main verbs, first stressed syllables in noun phrases, adverbs,etc.). Only about one sixth of the words generated by the grammar was expected to be preceded by Fo-detected boundaries. If a word was expected to be preceded by a boundary and a boundary was acoustically detected, then the priority or "score" for that word was increased substantially; if the expected boundary was not detected, the word's score is decreased. Unfortunately, while BBN implemented Lea's ideas in a version of their HWIM system, their project terminated before any results were available. There is no attempt as yet to communicate the syntactic information delivered by PROSEIDON to KEAL's syntactic module.

As a consequence, it is not possible to conclude on the efficiency of integrating information on prosodic boundaries in automatic recognition of continuous speech.

4. PART FOUR: AID TO SEGMENTATING AND ACOUSTIC-PHONETIC DECODING

4.1 Segmentation

a) It is obvious that the error rate obtained by prosodic modules using duration as a parameter is very sensitive to segmentation errors. On the other hand, Vaissière has proposed how the **Fo contour** over the successive acoustic segments detected by the acoustic-phonetic analyzer of a ASR system can be used to estimate some cases of phoneme spreading and merging. For example, when a vowel is matched to a large Fo rise (generally a manifestation of the so-called continuation rise), it tends to be spread into two (or more) vocalic segments by the segmenter. In the verification process, if two contiguous but automatically segmented sonorant segments are superimposed on a continuously rising Fo contour, the two segments should correspond to a single phoneme (VAI 82b).

b) Bush (BUS 86) has examined the influence of **durational** constraints on recognition accuracy in an acoustic phonetically-based speaker-independent connected digit recognizer. The constraints are expressed using a set of finite-state pronunciation networks, together with specifications of minimum and maximum allowable durations for network primitives. In particular, a set of networks incorporates additional paths representing prepausal lengthening for the digits "oh" and "eight". The recognizers were tested on a corpus of 1232 5-digit and 7-digit strings with and without a priori

knowledge of the string length. Recognition accuracies ranged from 33.9 percent to 94.6 percent and from 91.6 percent to 96.8 percent, for unknown and known string lengths, respectively, depending on the particular durational constraints incorporated in the network models.

4.2 Voiced-voiceless distinction

Fo and timing of the events may be used to distinguish between voiced and unvoiced consonants.

a) In a language like French, vibration of vocal cords during the entire or major portion of the consonant is the major cue distinguishing pairs of consonants like /p,b/, /t,d/, /k,g/. The **presence of Fo** and the concommitent presence of low frequency energy are generally used in knowledge-based ASR systems for French (such as in the Keal system).

b) In a language like English, the presence of Fo is a weak cue for voicing (Fujimura, 1961). The **shape of Fo at the vowel onset**, just after the release, is a cue that has been proposed by Lea (Lea, 1973) to distinguish between unvoiced and voiced consonants in ASR systems. The **timing** of the segmental events plays a preponderant part in distinction: **voice onset time, closure duration and vowel duration**. How far such cues are used in the existing knowledge-based ASR system for English is unknown.

4.3 Obstruent-non obstruent distinction

The **Fo shape during the voiced consonant** is very useful in the visual decoding of utterances for which the spectrographic representation is supplemented by Fo contour, at least for French (personal experience). During the occlusion period of the vocal tract for the realisation of obstruent consonants such as for stops and fricatives, there is a concommitent drop in transglottal pressure which lowers Fo (FAN 59). In other words, Fo is not expected to rise, nor to stay level during the realisation of the obstruent consonants. Moreover, the stop and fricative consonants, at least in French and particularly when the consonants are in a dominant position, are typically accompanied by a typical Fo fall-rise pattern which does not exist (or is much less marked) for non-obstruent consonants. Such a characteristic is not yet used in present ASRs, but is potentially useful for distinguishing between /m/ and /b/ or /v/, or /n/ and /d/.

4.4 Dominant-dominated phonemes

A given phoneme is ideally realized with a number of acoustic cues. In actual speech, a number of cues may not be clearly present. Even a phonologically-stressed syllable may be acoustically destressed.

Ongoing research at CNET by Roland Vives and myself has shown the absolute necessity of prosodic consideration before performing

detailed verification on phonetic hypotheses to dramatically reduce the number of fatal errors (i.e. wrong suppression of the right solution). After word matching by the lexical analyzer, verification of the presence or absence of such acoustic cues corresponding to each matched phoneme is carried out depending not only on the position of the phoneme in the hypothesized word, but also on the position of the detected phoneme in the actual prosodic structuring of the sentence. For example, an unvoiced stop in a dominant position (the decision of dominancy is based on **Fo and duration** considerations) has to be realized with the absence of low-frequency energy for a period of time. In a dominated position, continued presence of low frequency energy is tolerated. The quantitative contribution of such prosodic considerations has not been systematically estimated since the work is far from being completed. The application of a first series of verification rules using a combination of segmental, suprasegmental and lexical knowledges leads to a suppressing of 45 percent of the lexical hypotheses. Three percent of right candidates are suppressed due to errors in segmentation and in Fo detection. Detailed verification of the acoustic attributes of phonemes without prosodic consideration leads to numerous fatal errors, i.e. the suppression of right candidates.

5. CONCLUDING REMARKS

It is not an easy task to come to a conclusion on the use of prosodic parameters in ASR. It is easy however to agree that very little has actually been done, that most studies are in fact preliminary, and that most of the work is still ahead. While prosody is potentially more useful in continuous speech recognition than isolated word recognition, less work has been done in the former direction. The feasibility (a small percentage of errors) and usefulness of stress detection of isolated words in ASR has already been rather well investigated. In constrast, there is a need for more research and experimentation in the domain of continuous speech and in dialogues. It has been shown that about eight out of ten syllables heard as stressed by listeners and most phonological boundaries are automatically detectable. Such information has not been however fully used. The state-of-the-art of the current ASR systems does not allow easy integration of prosodic information. Prosody needs a flexible host: it should be used both in a top-down and in a down-up fashion (for hypothetisation and verification), at all levels (acoustic-phonetic decoding, lexicon, syntax and pragmatics) and it requires feasibility of exchanging information between the different components of the recognizers.

The small amount of work done on prosody in ASR is by no means an indication that prosody is not important. Prosody is undoubtedly a necessary component for successful recognition of continuous speech. Prosodic knowldege can assist us in solving uncertainties arising from acoustic-phonetic errors and ambiguities. If properly

exploited, such constraints could suggest promising hypotheses to pursue as well as eliminate unlikely interpretations from consideration. In more concrete terms, first, prosodic knowldege enables us to determine whether each word candidate has a prosodic profile compatible with the input signals. In particular, a prosodically stressed syllable should correspond to a "stressable" syllable as indicated by the lexicon. Second, it allows testing to find whether a particular sequence of hypothesized words can occur within a prosodically correct sentence. Word- and phrase- boundaries corresponding to hypothesized word sequence have to be compatible with the juncture and disjuncture phenomena detected from prosodic analysis. Two prosodically-regrouped words should correspond to two syntactically closed words. Conversely, detected major and minor boundaries should correspond to syntactic breaks. Third, prosodic knowldege provides a basis for predicting additional but unhypothesized fragment of sentences: a short, very low Fo frequency syllable often corresponds to a function word. Prediction of the presence of a word category is very useful in certain cases because it is often very difficult to hypothesize function words in a down-up fashion, from segmental information only. Fourth, the relationship between "segmental" quality of phonemes and their position in the acoustic structuring of sentences seem to be very crucial in attributing the acoustic evidence to other sources of knowledge. Automatic identification of phonemes in short, low Fo syllables should be given less weight. As described in this communication, there have been already some work in the former directions. Two other directions should be seriously investigated, concerning dialogue and quality control of incoming speech. First, sentence type (interrogative, declarative, ...) and emphasis phenomena detected within the sentence should be plausible and appropriate in the context of an ongoing dialogue. Second, prosody is the only way to control overall "quality" of incoming speech. Given the state-of-the-art of the current recognizers a sentence has to be uttered in a particular way to be recognized automatically. The proposals to use prosody in "guiding" the manner of speaking based on estimated speech rate, range and distribution of Fo movements, and the use of natural tendencies to emphasize important portions of the message (see VAI 86, pp 218) have not yet been tested.

It is often said that prosody is complex, too complex for straightforward integration into an ASR system. Complex systems are indeed required for full use of prosodic information. Lea's and my own experiments have clearly shown that it is not easy to integrate prosodic information into an already existing system such as HWIM or KEAL. It is necessary therefore to build an architecture flexible enough to test "on-line" integration of information arriving in parallel from different knowledge sources, particularly from prosodic aspects of the sentence.

Acknowledgements

I wish to thank Shinji Maeda, and Luc Mathan for kindly accepting to review a version of this paper, and particularly Gerard Bailly for his valuable suggestions.

REFERENCES:

(AIN 86) Ainsworth, W.A., (1986), "Pitch cahnge as cue to syllabification" J. of Phonetics, 14, 257-264.

(AIN 87) Ainsworth, W.A., & Lindsay, D., (1987), "Identification and discrimination of Halliday's primary tones", Proc. of Institute of Acoustic of Keele.

(AUL 84) Aull, A.M. (1984), "Lexical stress and its application to Large Vocabulary Speech Recognition", Master's Thesis, MIT. (See also Proc. ICASSP-85, 1449-1552).

(BUS 86) Bush, M.A., (1986), "Durational Contraints for Network-based Connected Digital Recognition", Proceedings of Montreal Symposium on Speech Recognition, McGill University, July 21-22, 89-90.

(CAR 82) Carbonell, N., Haton J.P., Lonchamp, F, & Pierrel, J.M., (1982), "Elaboration Experimentale d'Indices Prosodiques pour la Reconnaissance; Application à l'Analyse Syntaxico-sémantique dans le Système Myrtille II", in (DIC 82).

(CAR 87) Carter, D. Boguraev & Briscoe, T., (1987), "Lexical Stress and Phonetic Information: Which Segments are most Informative", Proc. European Conference on Speech Technology, Edinburgh, Vol. 2 234-238.

(CHE 87) Cheung, Y. J. & al, (1977), "Computer Recognition of Linguistic Stress Patterns in Connected Speech", Correspondence, IEEE Trans. on ASSP, 252-258.

(COO 77) Cooper, W.E. & Sorensen, J.M., (1977), "Fundamental frequency contours at syntactic boundaries", JASA, Vol. 62, No. 3, 683-692.

(CRY 86) Crystal, T.H. & House, A.S., (1986), "On the Availibity of Durational Cues", Proceedings of Montreal Symposium on Speech Recognition, McGill University, July 21-22, 71-72.

(DIC 82) Di Cristo, A., Haton, J.P., Rossi, M. & Vaissière, J., editors, PROSODIE ET RECONNAISSANCE AUTOMATIQUE DE LA PAROLE, GALF Groupe de la Communication Parlée, 289 pp.

(DAR 75) Darwin, C.J., (1975), "On the dynamic use of prosody in speech perception", In Cohen & Nooteboom, S. (eds), STRUCTURE AND PROCESS IN SPEECH PERCEPTION. Heidelberg: Springer.

(DUM 86) Dumouchel, P. & Lenning, M., (1986), "Using stress Information in Large Vocabulary Speech Recognition System", Proceedings of Montreal Symposium on Speech Recognition, McGill University, July 21-22, 75-76.

(FAN 59) Fant, G., (1959), "Acoustic Description and Classification of Phonetics Unito", Ericsson Technics, No. 1.

(FAR 86) Fartenati, E. & Kori, S., (1986), "Effects of syllable and word structure on segmental durations in sopken Italian", Speech Communication 5, 17-34.

(FOW 77) Fowler, C.A., (1977), TIMING CONTROL IN SPEECH PRODUCTION, Ph.D. Thesis, University of Connecticut.

(FUJ 87) Fujisaki, H. & Kawai, H., (1987) "Realization of Word Accent in Connected Speech in Japanese", Eleventh International Congress of Phonetic Sciences, August 1-7, Tallinn, Estonia, U.S.S.R. Se 29.2.

(HOU 87) House, D., Bruce, G., Lacerda, F. and Lindblom, B., (1987), "Automatic Prosodic Analysis for Swedish Speech Recognition", Proc. European Conference on Speech Technology, Edinburgh, Vol. 2, 215-218.

(HUT 84) Huttenlocher, D., (1984), ACOUSTIC-PHONETIC AND LEXICAL CONSTRAINTS IN WORD RECOGNITION USING PARTIAL PHONETIC INFORMATION, S.M. Thesis, M.I.T., Cambridge, Mass.

(KLA 76) Klatt, D.H., (1976), "Linguistic uses of segmental duration in English: acoustic and perceptual evidence", JASA 59, 1208-1221.

(KOM 86) Komatsu A. &al, (1986), "Prosodic Aids to Structural Analysis of Conversational Speech" ICASSP 86, 2283-2285.

(KOR 87) Kori, S., Farnetani, E. & Cosi P., (1987), "A Perspective on Relevance and Application of Prosodic Information to Automatic Speech Recognition in Italian", Proc. European Conference on Speech Technology, Edinburgh, September 1987, Vol. 2, 211-214.

(LEA 72) Lea, W., (1972), "Use of Syntactic Segmentation and Stressed Syll-

able Location in Phonemic Recognition", 84th Meeting of the Acoustical Society of America, Miami, Florida.
(LEA 73a) Lea, W, (1973), "Segmental and Suprasegmental Influences on Fundamental Frequency Contours", in CONSONANT TYPES AND TONE, Southern California Occasional Papers in Linguistics No. 1, L. Hyman editor.
(LEA 73b) Lea W.A., (1973), "An Approach to Syntactic Recognition Without Phonemics", IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 3, 249-258.
(LEA 75a) Lea, W., (1975), "Isochrony and Disjuncture as Aids to Syntactic and Phonological Analysis", JASA, Vol. 57, 533.
(LEA 75b) Lea, W., Medress, M.F., and Skinner, T.E., (1975), "A prosodically Guided Speech Understanding Strategy", IEEE Trans. Vol. ASSP-23, 30-38.
(LEA 80a) Lea, W., (1980), "Prosodic Aids to Speech Recognition", in (LEA 80a), 166-205.
(LEA 80b) Lea, W., (1980), "Speech Recognition: What is needed now?", in (LEA 80a), 562-569.
(LEA 80c) W. Lea (editor), TRENDS IN SPEECH RECOGNITION, Prentice Hall Inc Englewoods Cliffs, New Jersey,
(LEH 70) Lehiste, I., (1970), SUPRASEGMENTALS, The MIT Press, Cambrdige, Mass. and London England.
(LEH 80) Lehiste, I., (1980), "Phonetic manifestation of syntactic structure in English", Ann. Bull. RILP, 14, 1-27.
(LIE 60) Lieberman, Ph., (1960), "Some Acoustic Correlates of Word Stress in American English", JASA, Vol. 32, No. 4.
(LIN 83) Lindsay, D., (1983), "A method of describing pitch phenomena" in INVESTIGATIONS OF THE SPEECH PROCESS (ed. Winkler), Bochum: Studienverlag Brockmayer, 189-210.
(MAE 76) Maeda, S., (1976), A CHARACTERIZATION OF AMERICAN ENGLISH INTONATION, PhD., M.I.T., Dpt of Electro-Engineering.
(MAR 79) Martin. Ph., (1979), "Automatic Location of Stressed Syllable in French", in Current Issues in Linguistic Theory, Vol. 9, p 1091-1094.
(MED 71) Medress, M.F., Skinner, T.E. and Anderson, D.E., (1971), "Acoustic correlates of word stress", presented to the 82nd Meeting of the Acoustical Society of America, Denver, Paper K3.
(MEL 82) Meloni, H. & Guizol, J., (1982), "Utilisation des Paramètres Prosodiques dans un Sysème de Reconnaissance Automatique de la Parole Continue", in (DIC 82), 93-120.
(MER 88) Mercier, G., Cozannet, A. & Vaissière, J., (1988), "Recognition of speaker-dependent continuous speech with Keal-Nevezh", in RECENT ADVANCES IN SPEECH UNDERSTANDING SYSTEMS, NATO ASIS, Springer Verlag.
(MER 75) Mermelstein, P., (1975), "Automatic Segmentation of Speech into Syllabic Units", JASA, Vol. 58, 880-883.
(NIS 81) Nishinuma, Y., Barber, S. & Hirst D.J., (1981), "Estimation de la Durée intrinsèque des Voyelles", XIIth Journées d'études sur la Parole, Montréal, 419-428.
(NOE 88) Noeth, E., (1988), "Prosodic features in German Speech" in RECENT ADVANCES IN SPEECH UNDERSTANDING SYSTEMS, NATO ASIS, Springer Verlag.
(OMA 73) O'Malley, M.H., Kloker, D.R., & Dara-Abrams B., (1973), "Recovering Parentheses from Spoken Algebraii Expressions", IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, 217-220.
(PER 82) Perennou, G. & Caelen, G., (1982), "Utilisation de la Prosodie pour la Reconnaissance de la Parole Dictée", in /DIC 82/, 25-57.
(PIE 86) Pieraccini, R., (1986), "Lexical Stress and Speech Recognition", 112th Meeting of the Acoustical Society of America, Anaheim, 8-12 December.
(VAI 76) Vaissière, J., (1976), "Automatic Procedure for Segmenting Continuous Speech into Prosodic Words, in French", Recherches Acoustiques, Centre National d'Etudes des Télécommunications, Vol. 2, 193-208 [in French].
(VAI 77) Vaissière, J., (1977), "Premiers Essais d'Utilisation de la Durée pour la Segmentation en Mots dans un Système de Reconnaissance", VIIIth Journées d'Etudes sur la Parole, Aix-en-Provence, 345-352.
(VAI 82a) Vaissière, J., (1982), "A suprasegmental Component in a French Speech Recognition System: Reducing the Number of Lexical Hypotheses and Detecting the Main Boundary", Recherches Acoustiques, Centre National d'Etudes des Télécommunications, Vol. VII, 109-125.
(VAI 82b) Vaissière, J., (1982), "Utilisation des Paramètres Suprasegmentaux en Reconnaissance Automatique de la Parole comme Aide à la Segmentation en Phonèmes", in (DIC 82), 123-140.

(VAI 83) Vaissière, J., (1983), Language-Independent Prosodic Features", in PROSODY: MODELS AND MEASUREMENTS, (A.Cutler & R. Ladd, editors), Tokyo: Springer-Verlag, pp. 53-66
(VAI 84) Vaissiere, J., (1984), "PROSEIDON: Automatic Detection of Prosody Cues in Continuous Speech", Proceedings of the XIIIth Journées d"Etudes sur la Parole", Bruxelles, 28-30 May, 189-190.
(VAI 86) Vaissière, J., (1986), "Speech Recognition: a Tutorial", in COMPUTER SPEECH PROCESSING, F. Fallside and W.A. Woods, editors, Prentice Hall international, 191-236.
(VIV 77) Vives, R., Le Corre, C., Mercier, G. & Vaissière, J., (1977), "Use of Prosodic Markers in the Automatic Recognition of Continuous Speech", Proceedings of the VIIIth Journées d'Etudes sur la parole, GALF, Aix-en-Provence, 353-363.
(WAI 86) Waibel, A., (1986), "Recognition of Lexical Stress in a Continuous Speech Understanding System. A Pattern Recognition Approach", ICASSP 86, 2287-2290.
(WAI 87) Waibel, A., (1987), "Prosodic Knowledge Sources for Word Hypothesization in a Continuous Speech Recognition System", ICASSP-87, 856-859.
(WIN 75) Wingfield, A., (1975), "Acoustic Redundancy and the Perception of Time-compressed Speech", J. Speech Hearing Res., 18, 96-104.