A CROSSLINGUISTIC DESCRIPTION OF INTONATION CONTOURS OF A MULTILANGUAGE TEXT-TO-SPEECH SYSTEM

G. Olaszy

Phonetics Laboratory, Linguistics Institute
of the Hungarian Academy of Sciences, Hungary

ABSTRACT

Building elements to realise intonation contours in the MULTIVOX multilingual text-to-speech system are discussed. The description concerns intonation patterns on the word, phrase, and sentence levels from the point of view of Hungarian, German, Finnish, Italian and Esperanto. The crosslinguistic features of the patterns will be shown as well.

1. INTRODUCTION

In text-to-speech synthesis the robot-like sound can be improved towards a more natural, human-like voice quality – among other things – by superimposing intonation patterns. The newest directions in text-to-speech synthesis point in many cases towards a multilingual approach combined into one modular system [2]. The Multivox system is a general, text-to-speech system developed in Hungary [4] for multilingual synthesis. The system works in Hungarian, German, Italian, Esperanto and Finnish. New languages can be adapted easily to the basic system. Dutch and Spanish are under development. The synthesis hardware is the PCF8200 formant synthesizer. In MULTIVOX a modular representation of intonation has been implemented.

2. ELEMENTS FOR INTONATION AND STRESS

In devising acceptable intonation for unrestricted text we must formulate a set of rules which result in natural sounding pitch contours for utterances that may have never been spoken [9]. In the MULTIVOX system the following elements of pitch movements and timing correction are used as modular units in intonation and stress generation.

1. Starting (S) point of the pitch contour
2. Direction of the pitch movement: rise (R); fall (F)
3. Degree: high (H), medium (M), low (L)
4. Steepness (St) of movement in time
5. Jump down (Jd)+(level) or(S) or (E)
6. Jump up (Ju)+(level) or(S) or (E)
7. No change (N)(ms)
8. End point (E) of the pitch contour
9. Lengthening (L) of the stressed vowel

The degree parameter can be adjusted to all units. Examples:RM means rising to medium level; Ju(SM) means jump up to a medium starting point.
The physical values concerning these three degrees are shown in Table 1.

Table 1.

Unit	Degree			
	High	Medium	Low	
S/E	125	110	95	Hz
R/F	25	15	5	%
N	–	–	–	
St	2	0,5	0,25	Hz/ms
L	3x	2x	1,5x	times

These data are used for a male voice generation.

3. PITCH AND TIMING IN WORD STRESS

Two questions were taken into consideration in the formation of word stress, the relation of pitch variation and the duration of the vowel in question.
(i) Whether pitch change cooccurs with lengthening or not? This and the place of the accent is shown in Table 2.
Table 2. expresses that in Italian and in Esperanto the pitch contours and the lengthening of the vowel in question have to be treated together. In the other languages these two parameters are treated separately.

Table 2.

Language	Pitch change	Lengthening	Accented syllable
Hungarian	+	–	initial
German	+	–	any
Italian	+	+	any
Esperanto	+	+	penultimate
Finnish	+	–	initial

(ii) Vowel duration influences the form of the pitch pattern. Our experience is that the same pitch contour cannot be used automatically in the case of a short and a lengthened vowel. Slight changes characterise the pattern for long vowels. A stress pitch contour for these cases looks like this:
for a short vowel (V):
SM(RM)(StH)+(FM)(StM)EM
for a long vowel (VV):
SM(RM)(StM)+N(x)+(FM)(StM)EM
The value of (x) is language dependent. For Hungarian and German it is cca. 30 ms, for Italian and Esperanto in closed syllables cca. 30 ms, in open syllables cca. 60 ms, in Finnish cca. 60 ms.

3.1. Word stress categories:

Rule 1: Stress on the first syllable. Languages: Hungarian, Finnish, German
Rule 2: Stress on the last syllable. Languages: Italian, German
Rule 3: Stress on the penultimate syllable. Languages: Italian, German, Esperanto.
Rule 4: Stress on other syllables. Languages: Italian, German.
Rule 5: Unstress the sequence. Languages: all.
These 5 types of rules serve for word stress realisation in the mentioned five languages.

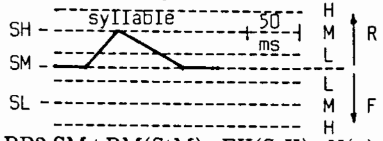
3.2. Algorithms for stress assignment

As Table 2. shows, Hungarian, Finnish and Esperanto can be treated as fixed stress languages, German and Italian are free stressed ones. For fixed stress languages the stressed syllable in the word can be determined by the rules 1 and 3. If the stress is signalled by diacritics – like in Italian –, rule 2 will be used.
For free stress languages the algorithms for finding the stressed syllable in the word are based in many cases on a large morpheme inventory (10.000–50.000 entries) and a morpheme analyser algorithm. Such solutions are known for English [1] for German [3] and for Italian [7], too.
The MULTIVOX system was designed to work with a relatively small memory (max. 100 kbyte) and in real time on a PC. Therefore no morpheme inventory and no morpheme analysis is used at all. To assign the proper place of the stress in the word (for Italian and for German) the "letter sequence" method (LSM) [5] and some other special algorithms were developed. The output of LSM is a sound level representation of the written text where the final duration of vowels is already set correctly in 95% (incorporating the necessary lengthenings coming from stress or from other linguistic rules).
In the Italian version of MULTIVOX the stress algorithm searches the syllable to be stressed on the basis of vowel durations. The stress will be superimposed where a vowel is lengthened in the word. This solution is an indirect approach to stress determination.
A more complicated solution appears in the German version, where the place of stress was assigned by the following rules.
D1.There is only one stress in one word.
D2.Stressed prefix suffix has priority against other rules (*ankommen*, *Komponist*, *studieren*).
D3.An unstressed prefix is followed by a stressed syllable (*bekommen*, *gesagt*).
D4.In two syllable words the long vowel (if there is any) is stressed (*fahren*, *sehen*, *primär*), else the first (*Silbe*, *Tausend*). This last rule is based on empirical observations.
Using these rules for finding the place of stress in German words a correct pitch superimposing is performed in 95% of the cases. The evaluation of these rules were done by listening to 1600 German sentences [8] and 50 text files (one A4 page each) gathered from books and newspapers. A weaker point of the German word stress assignment is the case of compound words. Here only rules D2 and D3 can assign a place of the stress for pitch patterns. Incidentally, the correct timing structure (without a pitch pattern superimposed) gives the feeling of correct stressing in most cases.

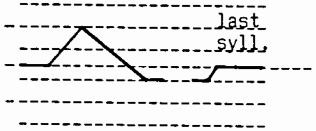
3.3. Pitch patterns for word stress

The following types of pitch patterns (PP) are used to create the frequency component of stress:

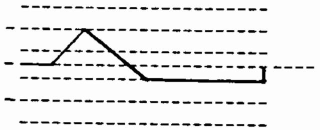
PP1.SM+RM(StM)+FH(StH)+EM
Hungarian: first syllable,
Italian: every stress except final,
German: stressed suffix
Esperanto: every stress.



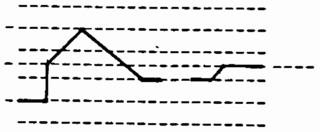
PP2.SM+RM(StM)+FH(StH)+N(x)+ +RL(StM)+EM
German: first syllable in more-than-two-syllable words.



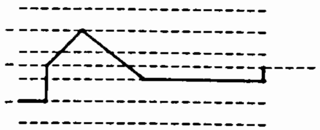
PP3.SM+RM(StM)+FH(StH)+N(x)+ +Ju(EM)
German: first syllable in two-syllable words,
Finnish: every stress.



PP4.SL+Ju(SM)+ PP2
German: unstressed prefix in more-than-two-syllable words.



PP5.SL+Ju(SM)+ PP3
German: unstressed prefix in two-syllable words.



The question of unstressing is just as important as stress if we want to get closer to the natural variation among stressed, unstressed, and neutral parts in human speech. Unstressing in MULTIVOX is generated by reducing the pitch value to SL during the sequence (word, prefix, suffix, etc.). This method is used for every language in the system. In sum, concerning word stress generation three types of cases are used: stressed, unstressed and neutral sequences. All these patterns remain present in higher level intonation patterns, i.e. in phrase and in sentence intonation.

4. PHRASE LEVEL INTONATION

The detection of phrase boundaries is performed in general on the basis of parsing [1], [3]. The MULTIVOX system is irregular with respect to this solution, too. A simple phrase boundary detection was designed and realised, similar to the solution proposed by O'Shaughnessy [6] for English. Function words and some other special words are used to detect boundaries [5]. This solution is done for all the languages in the system. Exceptions are Esperanto and German, where additional rules also help to improve phrase detection. For Esperanto noun and verb phrases can be detected because of the regularity of the language. In German the nouns are detected by searching capital letters as initials in words. For phrase intonation the same pattern is used in all languages i.e. the pitch is slightly rised continuously in the last two syllables of the phrase e.g. RL(StM). The pitch is set back (JdM) during the phrase pause which is 200–300 ms between the phrases.

5. SENTENCE INTONATION

In sentence intonation a serious problem is to find such rules that make the monotonous sounding more natural, so that listening to long texts should not be uncomfortable [2].

Two types of sentence intonations are generated automatically in the MULTIVOX system: one for declarative sentences and one for questions. For declaratives the general theoretical pattern is a linear falling one. This pattern is used for all the languages except Italian, where a rising–falling pattern is superimposed. To achieve variability in long texts (sentence by sentence) the following simple rules were built into declarative intonation: the starting pitch value and the steepness of the declination is changed as a function of sentence length (Table 3.)

Table 3.

Sentence length		Start pitch	Steepness
very short	(300ms)	120 Hz	10Hz/100ms
short	(600ms)	118	3
medium	(1 s)	116	2
normal	(3 s)	114	0.5
long	(8 s)	112	0.2
very long up to	(15 s)	110	0.1

In addition, the last word of the sentence is set to a lower pitch value for creating the feeling that the sentence has ended. At phrase boundaries the pitch is set higher (1-2 Hz/boundary) in the long and very long categories. This gives the feeling that a new phrase has begun. With these simple rules a relatively diversified sounding has been reached in reading long texts.

In questions, different types of pitch patterns have to be superimposed depending on the kind of question, like question with Q word/ without Q word; one-syllable question.

5.1. Question with Q word

A general pattern is used for all the languages in the system. A high peak is set on the Q word i.e.

RH(StH)+ FH(StM)+ FL(StM)

and afterwards a falling pattern is superimposed (similar to the declarative sentence but with less steepness). It is important to set the end of the falling part of the peak lower than the starting point was. The place of pitch change depends on the Q word and on the language (first, second etc. syllable). Markers sign the subgroups of Q words and the peak is placed where the marker points.

5.2. Questions without Q word

A general pattern for all the languages – except Hungarian – is as follows: The beginning is Jd(M) and the end is like in the phrase pattern. It is important to set a lower starting point than in the declarative sentences. In Hungarian the end pattern is a peak i.e.

RH(StH)+FH(StM)

on the penultimate syllable.

5.3. One-syllable questions

The pattern is the same for all the languages for one-syllable questions. This is a rising one i.e.

SL+RL(StL)+RL(StM)+RH(StH).

This pattern expresses a gradually increasing pitch value in the question.

6. CONCLUSIONS

An attempt at multilingual intonation synthesis with a limited number and sort of pitch patterns was described. Our findings are that the patterns shown above are enough to realise the most characteristic pitch contours of many languages. The practical working of the above patterns was tested in the MULTIVOX system. The results are tolerably good.

7. REFERENCES

[1] ALLEN, J.–HUNNICUTT, M.S.–KLATT, D. (1987), *"From text to speech. The MITalk system"*, Cambridge.

[2] COLLIER, R. (1990), "Multilingual intonation synthesis: principles and applications", *Proc. of the ESCA Workshop on Speech Synthesis*, Autrans, 273–76.

[3] KOHLER, J.K. (1990), "Improving the prosody in German text-to-speech output", *Proc. of the ESCA Workshop on Speech Synthesis*, Autrans, France, 83–87.

[4] OLASZY, G. (1989), "Speech synthesis in Hungary from the beginnings up to 1989", *Proc. of the Speech Research '89 Conference*, Budapest, 289–92.

[5] OLASZY, G. (1991), "Timing algorithms in the MULTIVOX text-to-speech system", In: *Temporal patterns of speech.* Ed. Gósy, M. Budapest.

[6] O'SHAUGHNESSY, D.D. (1989), "Parsing with a small dictionary for applications such as text to speech", *Computational Linguistics* Vol.15 num. 2., 97–108.

[7] SALAZA, P.L. (1990), "Phonetic transcription rules for text-to-speech synthesis of Italian", *Phonetica* 47, 66–83.

[8] SOTSCHEK, J. (1984), "Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die deutsche Sprache", *Fortschritte der Akustik, DAGA'84*, 873–876.

[9] TERKEN, J.M.B.–COLLIER, R. (1989), "Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch", *Proc. of Eurospeech'89*, Edinburgh, 357–359.