Voxton, Russon: Systems Generating Impersonal Hungarian and Russian Speech by Rule

K. Bolla
Budapest, Hungary

1. The structure and use of Voxton

The physical constituents of the sound substance of speech are organised in two ways in linguistically relevant structures. One results in the segmental and the other in the suprasegmental structure of speech. A speech synthesizing system generating impersonal speech by rule, named Voxton, consists of three main parts : a) a collection of sound sections, i.e. the data base realizing the basic units of the acoustic structure; b) the operational computer programs; and c) a code which organizes the phonetic elements and structures of the sound stream.

A sound section can be defined as a homogeneous segment of the acoustic structure of the stream of speech sounds which can be isolated by considering changes in the acoustic constituents. The number of acoustic parameters determining the structure of a sound section is between 1 and 23. Four types of elements can be differentiated according to the acoustic quality of the sound section: pauses, elements with voiced structure, elements with noise structure and elements with mixed structure. The data of 550 sound sections are included in the Voxton speech synthesizing system. Of these elements 27 are used to synthesize the vowels and 89 are used to synthesize the consonants. Sound sequences are realized with the help of 'transitional sections', which are largest in number: 414. Not only do transitions have a role in forming the acoustic structure, they are also important in speech perception. Voxton can be used to deal with this question in depth. Temporal variation, i.e. the phonetic realization of long–short oppositions, is achieved by doubling one of the sound sections making up the relevant speech sound. According to its position in the word, each speech sound can be synthesized in three variants, with different qualities corresponding to word-initial, word-medial and word-final positions.

The phonetic code of Voxton consists of the identifiers of the sound sections and the 'call signals' of the speech sounds. The identifiers indicate a) which speech sound the section belongs to; b) its status in the structure of the speech sound; and c) its position in the sound sequence. The computer automatically selects the sound sections from the data base and combines them. The sound transitions are also built in automatically according to the structural code. The sound sequence called by the phonetic symbols corresponds to the segmental structure of Hungarian speech. The speech sounds in the sound sequence are realized on a monotone with their characteristic quantity and intensity. The appropriate suprasegmental structure is built up in a separate step. This can be carried out quickly after the data characterizing the intonation pattern, the dynamic structure, the tempo and the rhythm are fed in. It is possible to change the intonation pattern of a single segmental structure as needed; an infinite number of suprasegmental variants can be made (see Fig. 1). Each synthesized sentence or clause is given an identification, which makes repetitions, storage, repeated use and the production of longer texts more effective.

2. A brief description of Russon

Russon is a synthesizing system suitable for artificial production of the phonetic form of Russian speech. Its minimalized and optimalized data base contains the data of 265 sound sections from which 87 different speech sounds (35 vowels and 52 consonants) and moreover 4 pauses of different duration can be produced. The 87 speech sounds also include the positional variants of the sounds. Russon can be used in two variants: in a phonetic and a phonematic speech generating system. In the first case we use the phonetic characteristics of speech sounds in structuring the text, while in the second one, we describe the text or sequence of sounds to be synthesized with the phonemes of the Russian language. E.g.:

- phonetically: [SA“DY CV’IETUT V’IESNO’J’] # 00
- phonematically: /SADI’ CV’ETU’T V’ES+NO’J’+/ # 13

According to the phonetic code and phonotactic rules built into Russon the following steps take place: a) the building up of speech sounds from the proper sound sections, b) the selectiôn of realizations/allophones of the phonemes used, c) the distinction of stressed and unstressed positions, d) the recognition of phonetic positions of vowels (apart from the word-stress, the modifications arising from the word-initial, word-medial and word-final positions; furthermore the patalized and pharyngealized consonantal surroundings are to be implemented, e) the assimilation of consonants according to their voiced/unvoiced quality, f) the assimilative palatalization of consonants, g) the alternation (lengthening) of duration of sounds, h) the recognition of focus in intonation in the phases and phrases of speech, i) transitions of the selected intonational constructions in the sound sequence. An F_0 matrix is used for the automatic synthesis of the intonation of Russian speech, (see Fig. 2).

3. The Common Features of the Two Speech Synthezising systems

We mention only the most important aspects. Both are built on the same hardware (a PDP 11/34 computer and an OVE III speech synthesizer).Its software consists of an RT-11 operating system and programmes written in

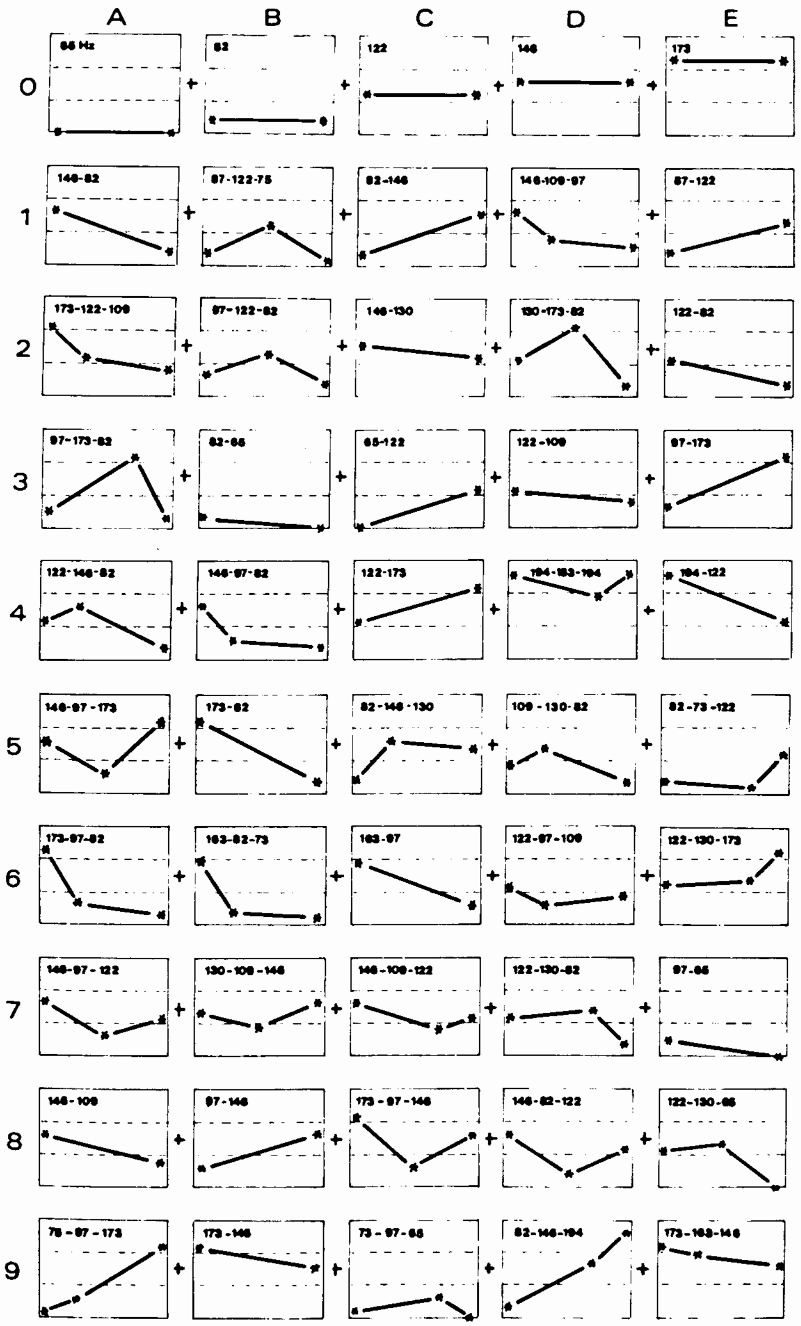

Fig. 1. F_0 patterns in Voxton used for automatic synthesis of Hungarian speech intonation.

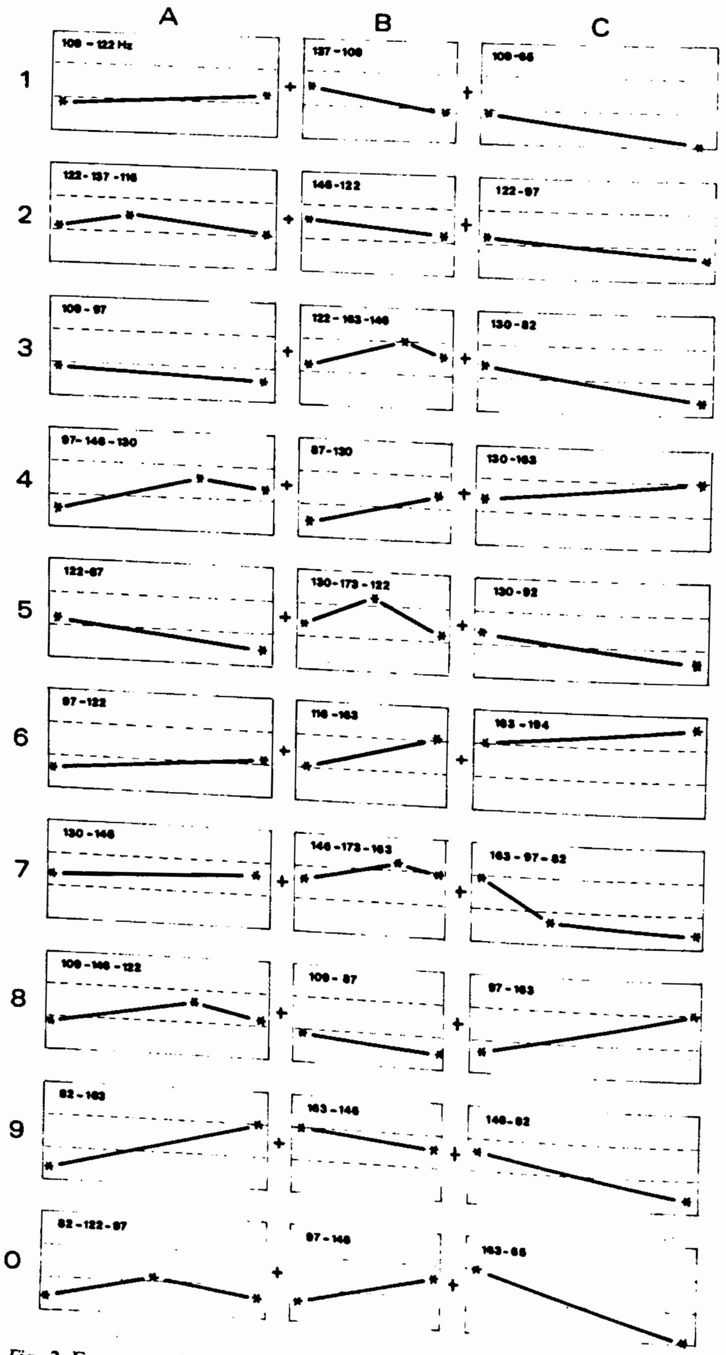

Fig. 2. F_0 patterns in Russon used for automatic synthesis of Russian speech intonation.

FORTRAN IV. The mode of producing speech based on rules is that of formant-synthesis. It is suitable for producing any kind of Hungarian or Russian text. The length of the texts which can be synthesized at one time is 5 s. The prescribed succession of sounds starts speaking with a delay of 30 s. The real speed of the speech is between 0.1 sounds/s to 25 sounds/s, but it can be altered between 6 sounds/s to 20 sounds/s. Voxton and Russon are phonetic systems i.e. their constructions and functions folow the phonetic–phonological systems of the Hungarian and Russian languages. There are three levels represented in its construction: a) the physical–structural level of the acoustic characteristics of speech, b) the so-called empiric phonetic level, c) the abstract phonological system-level of the language. These parts, which are well separable, easy to survey and stand in close connection with each other, constitute our phonetic synthesizing system as a structurally and functionally arranged whole. Any component of the acoustic structure can easily be changed within wide limits. The sound elements with their specific duration, pitch and intensity take part in the building up of the segmental structure. The formation of the suprasegmental structure of sounds is possible in three different ways: a) we manually give the F_0, A_0, Ac, T and tempo values by TON, ERO, IDO and IRA commands, b) by the automatic building of the intonational model chosen from the F_0 matrix and c) by the automatic transition of intonational structures from the patterns of intonation. The intelligibility of the speech produced by Voxton and Russon can be said to be good.

Reference

Bolla, K. (1982). Voxton: A system generating impersonal Hungarian speech by rule. *HPP* **10**.