A STEP TOWARD GENERAL MODEL FOR SYMBOLIC DESCRIPTION OF THE SPEECH SIGNAL

Dr Goran S. Jovanović
Institute for Applied Mathematics and Electronics
Mathematical Institute of Serbian Academy of Arts and Sciences
Kneza Miloša 37, 11000 Beograd, SR Yugoslavia
FAX: +381 11 186105, E-mail: jovanovicg@buef31.etf.bg.ac.yu

ABSTRACT

The paper presents an improved and extended version of previously defined general model for symbolic description of the speech signal. In the first part of the paper we formally define symbolic description segments that correspond to the lower speech coding levels (word and subword speech signal segments). In the second part of the paper we perform an analysis of practical applicability of the proposed model. Experimental evidence confirmed that one way to develop automatic procedure for symbolic description of the speech signal is by the use of *IFC*-guided speech signal processing, which provides specific focusing structural analysis. We believe that presented experimental results are inspiring from the standpoint of new research projects, especially in the field of automatic speech recognition and efficient speech coding.

1. INTRODUCTION

Significant obstacle for efficient scientific communication in speech processing community is absence of widely accepted models for symbolic description of the speech signal. This is probably one of the reasons why inspite of enormous research efforts, for example, in the field of automatic speech recognition (ASR) and understanding (ASU) we still cannot be satisfied with performances of the systems existing in practice. Similar statement holds in the field of (very)low bitrate speech coding, too.

In this paper we define symbolic description models that are applicable for different languages. They can be observed as an extension of the general model described in [1,4]. According to [1]. speech segment G^{J+1}, corresponding to hierarchical coding or processing level (J+1), can be presented as a sequence of speech/nospeech pairs belonging to lower coding level,

$$G^{J+1} = \underset{i}{CAT}(CAT(G_i^J, \overline{G}_i^J)), \qquad (1)$$

where $i = \overline{0, K^J}; K^J$ – number of speech J-level segments that are used in acoustical realization of the speech coding element assigned to level (J+1); $G_0^J = \lambda$, $\overline{G}_i^J \in \overline{\boldsymbol{G}}^J \cup \{\lambda\}$; λ - empty sequence; $\overline{\boldsymbol{G}}^J$ - the set of possible nospeech segment realizations. '$\overline{G}$' is general designation for nospeech segment (voice excitation absent), which according to language production rules at acoustic level does not belong to the observed speech code element For convenience and due to lack of space we shall restrict our consideration to the case when the speech segment corresponds to a word of the given language, i.e. $G = G^J = R$. In addition, in this paper we present experimental evidence confirming the possibility to generate symbolic description by the use of so-called *IFC*-guided speech signal analysis (*IFC* function was originally defined in [2]).

2. MODELS FOR SPEECH SIGNAL SYMBOLIC DESCRIPTION AT DIFFERENT LEVELS

In the case of arbitrarily segmented (X-segmented) word R its symbolic description can be defined by (2)

$$R = [*\_\overline{G}^R\_]\ X_1\_\overline{G}_1^X\_X_2\_\overline{G}_2^X\_\cdots \cdots\_X_i\_\overline{G}_i^X\_\cdots\_X_{K_X}[\_\overline{G}^R\_*] \qquad (2)$$

Content in the brackets designates the corresponding (left or right) context of acoustical realization of the word. Symbol '*' is general substitute; 'X' is the speech segment of the given processing level; $X_i \subset \boldsymbol{X}$, $\boldsymbol{X}$ - the set of all possible acoustic realizations of X; $\overline{G}_i^X \subset \{\overline{\boldsymbol{G}}^X, \lambda\}$, where $\overline{\boldsymbol{G}}^X$ corresponds to all possible acoustic realizations of nospeech segments within acoustical appearance of word R in the case of its X-segmentation.

In the case of phonemic symbolic description (2) has the form designated by (3)

$$R = R^F = [*\_\overline{G}^R\_]\ F_1\_\overline{G}_1^F\_F_2\_\overline{G}_2^F\_\cdots ...\_F_i\_\overline{G}_i^F\_...\_F_{K_F}[\_\overline{G}^R\_*] \qquad (3)$$

where $F_i \in \boldsymbol{F}$, $\boldsymbol{F}$-the set of possible acoustic realizations of phonemes; $\overline{G}_i^F$ elements, also, 'cover' the case of spelling speech production mode.

Each segment corresponding to phonemic speech element can be represented by (4)

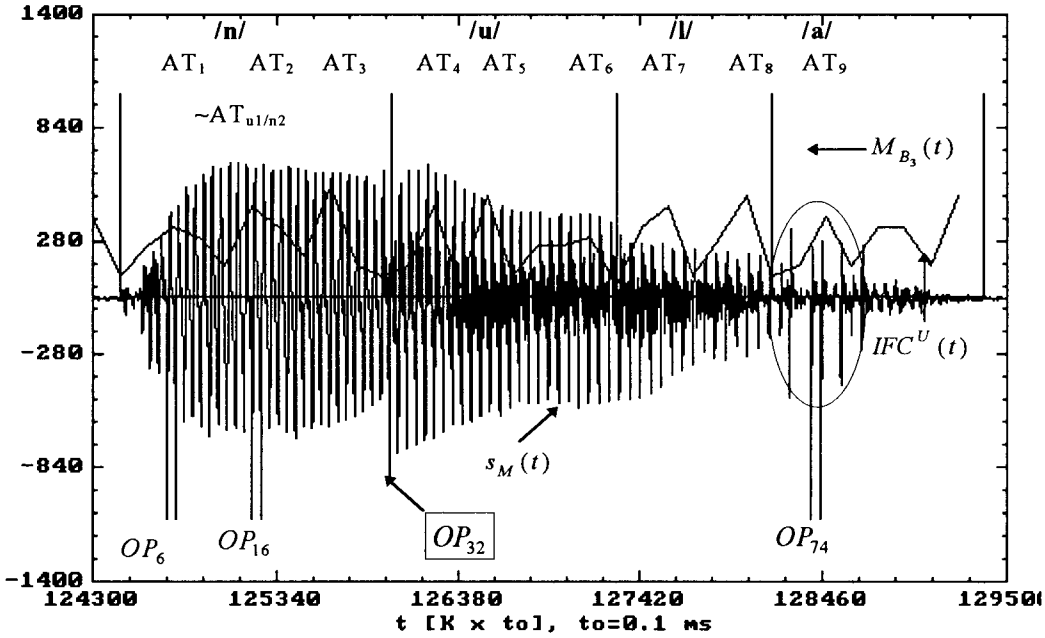

FIGURE 1: $s_M(t)$ is original speech signal which corresponds to the word */n_u_l_a/*. Local maxima of $IFC^U(t)$ function indicate subphonemic acoustic targets. $M_{B3}(t)$ determines phonemic segments obtained by an independent expert analysis. Pitch period segments OP_6, OP_{16} and OP_{74} have been used to generate synthetic phoneme-like segments in order to check if they can be used as representatives of the corresponding phonemic acoustic targets.

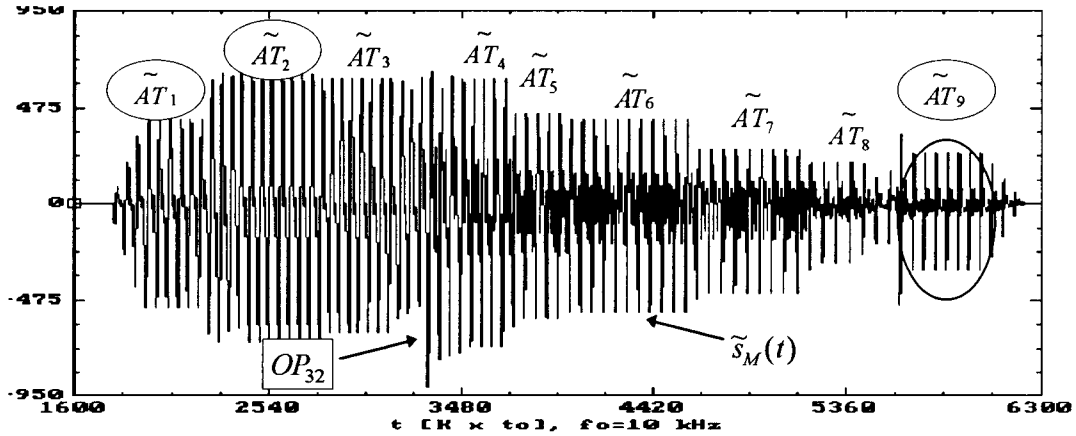

FIGURE 2: Synthetic signal $\tilde{s}_M(t)$ corresponds to signal $s_M(t)$ in Fig. 1. Encircled is the signal portion corresponding to diplophonia phenomenon in Fig. 1.

TABLE 1: The results of psychoacoustical tests that have been conducted.

SUBSTITUTION OF SEG | AT_x					
Phoneme (/ x /)	**/ n /**	**/ u /**	**/ l /**	**/ a /**	**All**
MOS_{AS}	**4.63**	**4.60**	**4.53**	**4.07**	**3.78**
DIF_{AS} (S-10)	**-0.20**	**-0.28**	**-0.38**	**-1.15**	**-1.97**
Number of AT per /x/	**1**	**2**	**5**	**1**	**9**
Perceptual effect rank order	**4**	**3**	**2**	**1**	**0**
Perceptual recognizability of synthetic segments [%]	**30(1)**	**70(2)**		**100**	

NOTE: Recognizability of synthetic segments as /n/ or /m/ was 60 percent.

$$F = [*\_\overline{G}^F\_]\ P^F\_AT^F\_Q^F\ [\_\overline{G}^F\_*] \qquad (4)$$

where P^F / Q^F designate transitional parts of the phonemic segment. Phonemic acoustic target AT^F should contain invariant coding features of the given poneme. Because of speech production process properties it is reasonable to describe this segment as follows.

$$AT^F = [*\_P^F\_]\ AT_1^F\_q_1\_\overline{G}_1^{AT}\_p_2\_AT_2^F\_q_2\_\overline{G}_2^{AT}\_\cdots$$
$$\cdots p_i\_AT_i^F\_q_i\_\overline{G}_i^{AT}\_\cdots\_p_{K_{AT}}\_AT_{K_{AT}}^F\ [\_Q^F\_*] \quad (5)$$

Here AT_i^F is subphonemic acoustic target. In standard speech production (when $\overline{G}_i^{AT} \to \lambda$) $q_i\_p_{i+1}$ correspond to fluctuations or changes in (sub)phonemic target's production. In the case of voiced speech segments subphonemic acoustic targets can be described by speech signal portions representing pitch periods (abbreviation OP):

$$AT_i = OP_1^i\_OP_2^i\_\cdots\_OP_j^i\_\cdots\_OP_{K_{OP}^i}^i$$
$$= \underset{j}{CAT}(\{OP_j^i\}),\quad j = \overline{1, K_{OP}^i} \qquad (6)$$

Previously defined speech signal symbolic description models stimulate research efforts oriented to define suitable methodology for their practical (and automatic) implementation. Encouraging results have been obtained by the use of so-called *IFC*-guided speech signal analysis [2, 3].

3. IFC-LOCALIZATION OF SUBPHONEMIC ACOUSTIC TARGETS

The first step to implement symbolic description (5) is localization of possible subphonemic acoustic targets. An original solution based on *IFC*-function (mnemonic abbreviation in Serbian for 'indication of focused targets') was formulated in [3]. It was briefly discussed in [4], and here it is illustrated in Fig. 1. *IFC*-function was designed to indicate speech signal portions where the best hits of the acoustic phonemic/subphonemic targets had occurred. Local maxima of the function is expected to determine the events of supraglottal acoustical system stabilization in speech production. Consequently, local minima of *IFC*-function correspond to transitional speech signal portions. The version of *IFC*-function depicted in Fig. 1, was obtained by the use of 'uniform LPC-parametrization' (analysis frame size and time shifts were 15 ms). This version of *IFC*-function is designated by superscript 'U'. The input signal $s_M(t)$ contains speech component corresponding to word $/n\_u\_l\_a/$. In this case by the use of (3) and (4) we obtain

$$R|<0>= [*\_\overline{G}^R\_]\ P^n\_AT^n\_Q^n\_P^u\_AT^u\_Q^u\_$$
$$\_P^l\_AT^l\_Q^l\_P^a\_AT^a\_Q^a\ [\_\overline{G}^R\_*] \quad (7)$$

Without any *apriori* information concerning real locations of phonemic and subphonemic acoustic targets realistic symbolic description on the basis of *IFC*-indications is

$$R|<0>= [*\_P^n\_]\ AT_1^n\_q_1\_p_2\_AT_2\_q_2\_\cdots$$
$$\cdots\_p_i\_AT_i\_q_i\_\cdots\_p_9\_AT_9^a\ [\_Q^a\_*] \quad (8)$$

For illustration purposes in Fig. 1 by function $M_{B_3}(t)$ is designated preliminary phonemic segmentation obtained by an independent expert audio-visual analysis. It differs from also preliminary segmentation that follows from the results of psychoacoustic (PA) tests that have been conducted previously [4]. This disagreement is designated by ambiguous labeling of subphonemic segment AT_2 .

Verification of linguistic relevancy of *IFC*-indications of AT_i segments was very complex and it included several PA-tests. All tests participated 20 listeners. Some of the results that have been obtained are presented in Table 1. The subjects of subjective judgments were perceptual effects of original $AT = AT^F$ substitutions by the corresponding synthetic $\tilde{AT}$ segments. AT_i segments' labeling regarding these results is obvious from the following set of relations:

$$AT^n \supset \{AT_1\},\ AT^u \supset \{AT_2, AT_3\},\ AT^a \supset \{AT_9\}. \quad (9)$$

Synthetic segments were produced by the use of symbolic description (10)

$$\tilde{AT}^F = [*\_P^F\_]\ \tilde{AT}_1^F\_\tilde{AT}_2^F\_\cdots$$
$$\cdots\_\tilde{AT}_i^F\_\cdots\_\tilde{AT}_{K_{AT}}^F\ [\_Q^F\_*] \quad (10)$$

Subphonemic synthetic segments were obtained by concatenation of the 'best' pitch periods $OP^*|AT_x$) (corresponding to *IFC*-function local maxima),

$$\tilde{AT}_i^F = CAT^{(K_{OP}^i)}(OP^*|AT_i^F) \qquad (11)$$

K_{OP}^i is repetition number for $OP^*|AT_i^F$ which was determined by some of synthetic segment duration restrictions

$$\ell(\tilde{AT}_i) \approx \ell(p_i\_AT_i\_q_i)$$

$$\ell(\tilde{AT}_i) \approx \ell(AT_i \_ q_i)$$

$$\ell(\tilde{AT}_i) \approx \ell(p_i \_ AT_i)$$

Generated synthetic speech segments are illustrated in Fig. 2.

Additionally, Table 1 contains the results concerning the case of simultaneous substitutions of all segments in the original signal. All measurements included replying of signal pairs (original and synthetic utterances with $\overline{G}^R \sim 0.5\ s$) in both presentation orders. So, in Table-1 $MOS_{A,B}$ is the average mean opinion score, and $DIF_{A,B}$ the average differential score obtained by the use the modified A/B comparison test with 10-point scale for speech quality numerical evaluation (numerical value 10 corresponds to the best speech signal quality).

The analysis of $MOS_{A,B}$ and $DIF_{A,B}$ values indicate that degradations in synthetic utterances are acceptable, even in the worst case -- when all phonemic targets were substituted by synthetic surrogates. Surprisingly, the worst scores were obtained for the case of phoneme /a/, but this was clarified by another set of experiments.

The purpose of the second set of experiments was to check whether in the case of sustained phonemes in $s_M(t)$ (/a/, /n/ and /u/) selected pitch periods $OP^*|AT_i^F$ can be representatives of the corresponding phonemes. Synthetic phoneme-like utterances were produced by the use of (12)

$$\tilde{SEG}^F = CAT^{(K_{OP})}(OP^*|AT_i^F) \qquad (12)$$

and restrictive condition regarding K_{OP} value

$$\ell(\tilde{SEG}^F) \sim 1.2\ s$$

In Fig. 2 are encircled AT_i designations which uniquely determine pitch periods used in these experiments. Listeners were asked to make subjective phonemic assignment of perceived impressions of segments generated by (12). The results are presented in the last row of Table 1. Analysis of these results points out that the best score was obtained for phoneme /a/ what is in virtual contradiction concerning the result from previous set of PA-tests. Explanation is contained in the encircled part of signal $s_M(t)$: diplophonia phenomenon is clearly observable! However, bearing in mind recognition score obtained for /a/ (100%) the corresponding scores for AT_1 and AT_2 indicate that additional research effort is necessary. A preliminary analysis has shown that research should be directed toward more precise *IFC*-analysis of the speech signal, i.e. pitch synchronous parametrization of the speech signal and, possibly, complex OP^*-representation (with more neighboring pitch periods) of AT_i^F-segments instead of 'simple' representation that has been applied in this research.

4. CONCLUSIONS

In this paper we defined one general model for symbolic description of the speech signal at different coding and processing levels. It is open for extensions toward higher as well as lower symbolic description levels, in agreement with the existing or new knowledge concerning complex speech coding process at different hierarchical levels. We demonstrated that the model works at low speech processing levels by the use of *IFC*-guided processing of the speech signal. Presented experimental results are inspiring from the standpoint of new research projects - as they point out some interesting questions. Among them of the greatest practical importance is what performance gain in different speech processing systems can be obtained by speech signal parametrization that is focused on AT_i segments.

We also believe that proposed symbolic description model will improve scientific communication in speech processing community.

5. ACKNOWLEDGMENTS

This research has been supported by the Department for Science and Technology of Republic of Serbia, Grant No. 04M02c, through Mathematical Institute of Serbian Academy of Arts and Sciences, Belgrade.

6. REFERENCES

[1] Jovanović G.S., *"Hierarchical speech signal model for expert knowledge based ASR/ASU systems"*, Proc. ISITA-94, Vol. 2, pp. 1215-1220, Sydney, Australia, 1994.

[2] Jovanović G.S., *"Identification of stationary acoustic targets in the speech signal"* (in Serbian), Proc. SYM-OP-IS'94, pp. 855-858, Kotor, Yugoslavia, 1994

[3] Jovanović G.S., *"A model for generation of symbolic description of the speech signal at phonemic and subphonemic level"*, Proc ISITA-96. pp. 270-273, Victoria, B.C., Canada, 1996.

[4] Jovanović G.S., *"ASR_L speech recognition system based on expert defined focusing structural analysis"*, Proc. 1995 IEEE Workshop on Nonlinear Signal and Image Processing, Vol. I, pp. 404-407, Neos Marmaras, Halkidiki, Greece; 1995.

[5] Markel J.D, Gray A.H., "**Linear prediction of Speech**", Springer-Verlag, Berlin Heidelberg New York, 1976.

[6] Flanagan J.L., "Speech analysis, synthesis, and perception", Springer-Verlag, Berlin·Heidelberg·New York, 1972.