BRITISH ACOUSTICAL SOCIETY

"SPRING MEETING" at Chelsea College, London, S.W.3. on
Wednesday 25th April / Friday 27th April, 1973.

SPEECH AND HEARING: Session 'C' Speech Properties and Recognition.

Paper No:

73SHC2

# On the Adaptation in Speech Recognition

B.M.LOBANOV

Radiotechnical Institute of Minsk, USSR

(Invited paper)

It is well known that the speech signal is extremely variable, depending on the characteristics of the acoustical tract, vocal peculiarities of the speaker and his physiological and pcychological state. Due to the slow changing nature of these intefering factors an adaptation procedure of speech recognition is made possible.

In the present paper an attempt of working out an adaptation procedure model with the object of automatic recognition on the basis of the available data on adaptation and the authors previous investigations is made.

Let $\{\Phi_1, \dots \Phi_j, \dots \Phi_N\}$ be a set of speech patterns intended for recognition; $\{S_1, \dots S_K, \dots S_M\}$ a set of the speech signal parametres and $\{\Pi_1, \dots \Pi_i, \dots \Pi_p\}$ a set of intefering factors conditioned by the influence of the individual characteristics of the electroacoustical tract and the speaker. Automatic recognition of a given set of speech patterns is only possible if the chosen speech signal parametres meets definite recuirements. The minimal requirement is that the distributions of the speech signal parametres for each of the speech patterns should not intersect (above a preliminary fixed level) at any fixed intefering effects.

A set of speech signal parametres may be considered optimal if they are fully invariable in respect of any intefering effects (labeled a priori adapted parametres). It means that the speech pattern distribution centers vary so insignificantly that it is possible to

apply the same decision rule at different values of intefering factors. In this case adaptation procedure is not necessary. Unfortunately, an invariant system of speech parametres (or features) can only be found but for a limited set of speech units.

Radically different is a system of speech signal parametres with the distribution centres varying considerably in different a priori unknown directions at variations of characteristics of intefering factors (labeled fully unadapted parametres). Adaptation is quite necessary then and a complete readjustment of all elements of the recognition system is required with the help of displaying all the speech patterns intended for identification.

Between these two is a set of parametres having the same exact or statistical linear regularities in the distribution centres displacement for all the speech patterns (labeled linear unadapted parametres). If an intefering effect produces a linear transformation of the space of parametres reduced to expansion (or compression) and shifting along the coordinate axes, then we have equantion $S^i_{kj} = a^i_k S^o_{kj} + b^i_k$ ,(1) where $S^i_{kj}$ is the value of the $k^{th}$ parametre for the $j^{th}$ speech pattern at the $i^{th}$ state of the intefering factors; $S^o_{kj}$ is the value of the $k^{th}$ parametre for $j^{th}$ the speech pattern if intefering factors are not present; and $a^i_k$ and $b^i_k$ are expansion and shifting coefficients. In this case adaptation process may be accomplished after displaying of only two realizations of preliminary known different patterns $\Phi_\alpha$ and $\Phi_\beta$. The unknown $a^i_k$ and $b^i_k$ are determined from the system of two linear equations

$$\begin{cases} S^i_{k\alpha} = a^i_k S^o_{k\alpha} + b^i_k \\ S^i_{k\beta} = a^i_k S^o_{k\beta} + b^i_k \end{cases}$$

and then adapted values of the parametres are found according to the formula

$$S^A_k = \frac{S^i_k - b^i_k}{a^i_k} = \frac{S^i_k (S^o_{k\beta} - S^o_{k\alpha}) - (S^i_{k\alpha} S^o_{k\beta} - S^i_{k\beta} S^o_{k\alpha})}{S^i_{k\beta} - S^i_{k\alpha}}. \quad (2)$$

The unknown coefficients $a^i_k$ and $b^i_k$ can also be found directly during the process of recognition that is by means of self-training. For this purpose it is sufficient to define some extremal [1] or mean-statistical values [2] of parametres. Really, substituting in formula (2)

$S^i_{\kappa max}$ and $S^i_{\kappa min}$ for $S^i_{\kappa\beta}$ and $S^i_{\kappa\alpha}$, assuming that $S^o_{\kappa max}=1$ and $S^o_{\kappa min}=0$ we come to the formula $S^A_\kappa = \frac{S^i_\kappa - S^i_{\kappa min}}{S^i_{\kappa max} - S^i_{\kappa min}}$, suggested by Gerstman [1]. Considering, further, in formula (1), $S^o_{\kappa j}$ to be a random value with $M_j = 0$ and $\sigma_j = 1$ and finding for $S^i_{\kappa j}$ the mean and root mean square values we can write

$$M_j(S^i_{\kappa j}) = M_j(a^i_\kappa S^i_{\kappa j} + \beta^i_\kappa) = \beta^i_\kappa$$

$$\sigma_j(S^i_{\kappa j}) = \sigma_j(a^i_\kappa S^i_{\kappa j} + \beta^i_\kappa) = a^i_\kappa$$

that is $S^A_\kappa$ according to formula (2) can be expressed by the formula $S^A_\kappa = \frac{S^i_\kappa - M_j(S^i_{\kappa j})}{\sigma_j(S^i_{\kappa j})}$, suggested by the author [2].

Information of the expansion and shifting coefficients can also be obtained through using their statistical correlations with the other parametres that are not directly involved in the process of recognition. It has been shown [3] that current values of highest formants and fundamental frequences can be used in adaptation process.

We have noted, besides, that on all occasions $a^i_\kappa$ and $\beta^i_\kappa$ have to be determined only for those speech pattarns for which they are identical. In general the set of speech patterns may be subdivided into a number of groups [4], for which $a^i_\kappa$ and $\beta^i_\kappa$ are determined separately.

Different types of adaptation should be applied in speech recognition depending on the parameter system employed. Let us consider, for instance, two parameter systems: a) signal-parameters of the parallel spectral analyzer and b) formant frequencies of the articulatory tract. The first parameter system may be considered "a priori adapted" at buss and hiss sounds identification. But for the majority of the phonetic elements the system of signal parametres is either "linear unadapted" (for interfearing factors due to the changing of the frequency characteristics of the electroacoustical tract and spectral characteristics of the excitation sourses) or "fully unadapted" (for interfearing factors due to the changing of the average sizes of the speaker's articulatory tract). The second system is better since it is "a priori adapted" for the first type of interfearing factors and only "linear unadapted"for the second type of interfearing factors.

In temporal aspect we could offer the following organisation of the adaptation procedure. Atbthe first stage (a proiri adaptation) the average characteristics of the most probable type of the speaker and the electro acoustical tract must be taken. At the second stage (momentary adaptation) the average characteristics are corrected by taken into account the information about the carrdent values of some speech parametres (the fundamental frequency, the frequency of the third formant, etc.). At the final stage (integral adaptation) they are finally corrected by calculating the statistical characteristics of the speech parametres on the chosen time intervals.

The author is indepted to L.Karnevskaya and L.Leladze for the help in preparation of the English version of the present paper.

REFERENCES

1. L.J.Gerstman. Classification on Self-Normalized Vowels. IEEE Trans. on Audio and Electroacoust., v.AU-16, No1, 1968.

2. B.M.Lobanov. Classification of Russian Vowels Spoken by Different Speakers. J.A.S.A., v.49, No2 (p.2), 1971.

3. H.Fyjisaki, T.Kawashima. The Roles of Pitch and Higher Formants in the Perception of Vowels. IEEE Trans. on Audio and Electroacoust., v.AU-16, No1, 1968.

4. B.M.Lobanov. On the Classification of Russian Fricatives in the CV-Syllables for Different Speakers. J.A.S.A., v.49, No4 (p.2), 1971.