SPEAKER AND CHANNEL-NORMALIZED SET OF FORMANT PARAMETERS FOR TELEPHONE SPEECH RECOGNITION

B. Lobanov, T. Levkovskaya, I. Kheidorov

Institute of Engineering Cybernetics Nac. Ac. of Sc. Belarus.
Surganov St. 6, Minsk, 220012, BELARUS
e-mail: lobanov@newman.basnet.minsk.by

ABSTRACT

The speech parameters, most commonly used nowadays, are Cepstral coefficients derived from FFT or LPC Spectrum. An alternative approach that can potentially provide maximum speaker and channel independence is estimation of articulatory based features such as formant frequencies, amplitudes and voicing degree. A present report describes a new method and algorithm of robust estimation of F1(t), F2(t), F3(t), A1(t),A2(t), A3(t), V(t) from telephone speech signal, and also the procedures of their normalization against speaker and channel variability. The results obtained from the experiments confirm the efficiency of the suggested set of formant parameters in a view of speech signal speaker – and channel variability resistance. According to the experiments it gives significant improvement in the recognition performance as compared with cepstral parameters use.

1. INTRODUCTION

There are many causes of acoustic variation in speech connected with speaker and channel conditions such as speaker voice quality, vocal tract size, pitch and channel characteristics, distortions, noise. The problems that we address are, firstly, what kind of speech parameters should be chosen, and, secondly, what type of their normalization should be utilized in order to obtain maximum speaker and channel independence in telephone speech recognition. The importance of the problem solution is extremely high for almost all the applications of speech technology in the public telephone network.

The speech parameters, most commonly used nowadays, are Cepstral coefficients derived from FFT or LPC Spectrum. An alternative approach that can potentially provide maximum speaker and channel independence is estimation of articulatory based features such as formant frequencies and amplitudes, voicing degree and some other features that provide information on place and manner of articulation for different sounds [1].

It is well known that current values of formant frequencies F_1, F_2, F_3 and their dynamics provide information on place of articulation while relative values of formant amplitudes A_1, A_2, A_3, their dynamics, and voicing degree V are responsible for manner of articulation. It is also known that this set of speech parameters provides (i) maximum phonetically important information by mean of (ii) a minimum information stream. Also, the formant frequencies are potentially (iii) robust against the spectrum distortions, and (iiii) might be easily speaker-normalized by rescaling of the formant space [2]. The only problem that is still open, is how to provide their most robust estimation from the real speech and especially from telephone speech signal.

Most of the formant estimation methods rely on the peaks of the LPC [3], analysis by synthesis with FFT spectra [4], and peak picking on cepstrally smoothed spectra [5]. These methods are not very robust against possible missing or spurious peaks. So, using formants for recognition can sometimes cause problems, and they have not yet been widely adopted. A successful attempt to use formant frequencies in speech recognition was described in [6]. It was suggested instead of peak picking spectra to compare it with about 150 typical spectral cross sections with labeling of the lowest three formants provided by a human expert.

A present report describes an algorithm of robust estimation of $F_1(t)$, $F_2(t)$, $F_3(t)$, $A_1(t)$, $A_2(t)$, $A_3(t)$, $V(t)$ from telephone speech signal, and also the procedures of their normalization against speaker and channel variability for telephone speech recognition.

2. METHOD AND ALGORITHM FOR FORMANT ANALYSIS

2.1. Estimation of the formant frequencies

In [7] a robust method for measurement of formants was suggested. The method is based on estimation of spectral center-of-gravity for time-domaine trajectories using mixture spline models and it requires a large amount of calculations. We suggest a new method that is not relied on the peaks of the spectra either, but is much simpler for calculation. The method is based on an iterative estimation of the partial centers-of-gravity for each spectral cross section that finally corresponds to the current formant frequencies $F_1(t)$, $F_2(t)$, $F_3(t)$.

Estimation of the formant frequencies is carried out by measuring the centers-of-gravity of the different regions of mel-spectrum:

$$CG = \sum_{b1}^{b2} S(j)*j \Big/ \sum_{b1}^{b2} S(j), \qquad (1)$$

where CG is center-of gravity, $S(j)$ is a value of j-th spectral component, b_1 is the left and b_2 is the right boundaries of the region. First, the center-of-gravity of full spectrum CG_1 is calculated by using $b_1 = 1$ and $b_2 = J$, where J is a number of spectral components. Then the calculation of CG_2 with $b_1 = CG_1$ and $b_2 = J$ is carried out. Finally, the first order estimation of the three formant frequencies is obtained as:

$$\begin{array}{llll} F_{11} = CG_3 & \text{with} \quad b_{11} = 1 & \text{and} & b_{21} = CG_1 \\ F_{21} = CG_4 & \text{with} \quad b_{11} = CG_1 & \text{and} & b_{21} = CG_2 \\ F_{31} = CG_5 & \text{with} \quad b_{11} = CG_2 & \text{and} & b_{21} = J \end{array} \qquad (2)$$

After the first order estimation of the formant frequencies is reached, an iterative procedure of their more accurate definition is carried out. The second order estimation of the n-th formant frequency F_{n2} is conducted by changing left b_{11} and right b_{21} boundaries to the new b_{12} and b_{22} according to the equation:

$$b_{12} = \begin{cases} b_{11} & \text{if} \quad |F_{n1} - b_{11}| < |F_{n1} - b_{21}| \\ 2*F_{n1} - b_{21} & \text{if} \quad |F_{n1} - b_{11}| > |F_{n1} - b_{21}| \end{cases} \qquad (3)$$

$$b_{22} = \begin{cases} b_{21} & \text{if} \quad |F_{n1} - b_{21}| < |F_{n1} - b_{11}| \\ 2*F_{n1} - b_{11} & \text{if} \quad |F_{n1} - b_{21}| > |F_{n1} - b_{11}| \end{cases} \qquad (4)$$

The iteration process for each n-th formant is continued m-times until the value $\delta = |F_{nm} - b_{1m}| - |F_{nm} - b_{2m}|$ becomes smaller than 1.

In Fig.1a (male voice) and in Fig.2a (female) the first order estimation of the formant frequencies' trajectories $F_1(t)$, $F_2(t)$, $F_3(t)$ are shown. In Fig.1b and Fig 2b – the same after m-th iterations.

2.2. The full set of the formant parameters and their normalization

The full set of the formant parameters includes $F_1(t)$, $F_2(t)$, $F_3(t)$, $A_1(t)$, $A_2(t)$, $A_3(t)$ and also $V(t)$ - voicing degree. Additionally, the first derivatives of these parameters are involved in recognition process. The current decision about the voicing degree V is based on the analysis of autocorrelation function $C(k)$ of the speech signal. The voicing degree is determined as:

$$V = 1 - \max_{\tau} \; C(\tau)/C(0), \qquad (5)$$

where time interval τ = (5 – 20) mc.

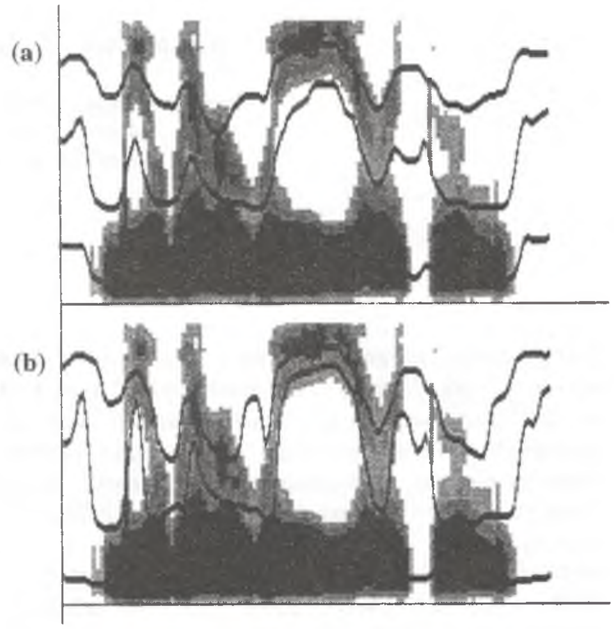

Fig 1. Frequencies' trajectories (*female voice*) for the phrase "*We were away a year ago*":
a) the first order estimation of the formant frequencies' trajectories;
b) the formant frequencies' trajectories after *m*-th iterations.

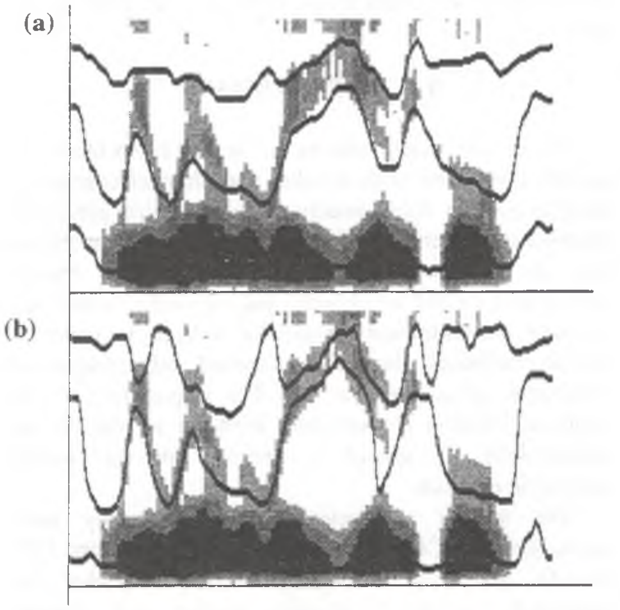

Fig 2. Frequencies' trajectories (*male voice*) for the phrase "*We were away a year ago*":
a) the first order estimation of the formant frequencies' trajectories;
b) the formant frequencies' trajectories after *m*-th iterations.

The values of V are normalized and supposed to be close to 1 for vowel-like sounds and close to 0 for fricatives. The voicing degree parameter $V(t)$ is also used for linear interpolation of the formant frequencies' trajectories at the time intervals where $V(t)<0.5$.

In Fig.3a the formant frequencies' trajectories $F_1(t)$, $F_2(t)$, $F_3(t)$ for male voice are shown with no interpolation, and in Fig.3b – the same with interpolation of trajectories. In Fig. 4a-b there are formant frequencies' trajectories for female voice.

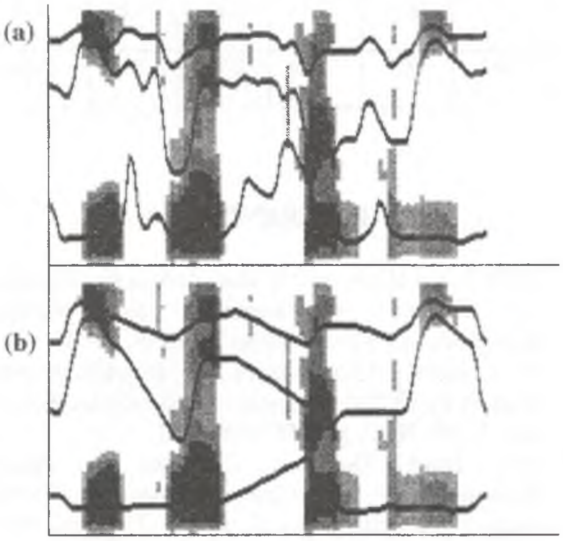

Fig.3. The formant frequencies' trajectories $F_1(t)$, $F_2(t)$, $F_3(t)$ *(female voice)* for the phrase *"East-West company"*:

a) **no interpolation; b) with interpolation**

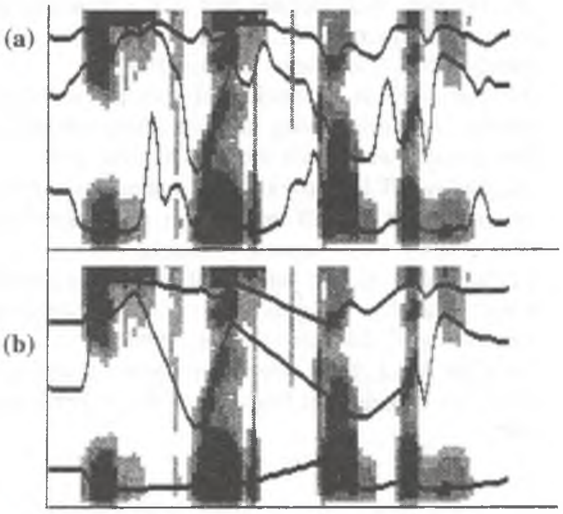

Fig.4. The formant frequencies' trajectories $F_1(t)$, $F_2(t)$, $F_3(t)$ *(male voice)* for the phrase *"East-West company"*:

a) **no interpolation; b) with interpolation**

Formant amplitudes $A_1(t)$, $A_2(t)$, $A_3(t)$ are simply derived as average values of spectral components in the neighborhood of non-interpolated $F_1(t)$, $F_2(t)$, $F_3(t)$, or more exactly in the corresponding regions $b_1 - b_2$ (3,4).

The normalized values of $F_1(t)$, $F_2(t)$, $F_3(t)$, $A_1(t)$, $A_2(t)$, $A_3(t)$ are obtained by using the formula:

$$P_j^N = (P_j - P_j^{\min})/(P_j^{\max} - P_j^{\min}), \qquad (6)$$

where $P_j^{\min}$ and $P_j^{\max}$ are minimal and maximal values of each j-th parameter over the utterance.

3. EXPERIMENTS and RESULTS

The aim of the experiments was to compare recognition results using suggested set of formant parameters with those obtained using more conventional mel-cepstrum coefficients (with subtraction mean value over the utterance). In order to assess directly the efficiency of the formant parameters, the same total number of features was used for both representations: 7 formant parameters and their 7 first derivatives, and 7 mel-cepstrum coefficients and their 7 first derivatives. Both sets of parameters derived from the same 24 filter bank with triangle bandpass mel-frequency responses by using 256-points FFT spectrum that was evaluated every 8 ms within 32 ms Hamming window at 8 kHz speech signal sampling frequency. The recognition system employed in the experiment was the same we used in [8], and it is based on the originally modified DTW algorithm. The test was carried out within a practically important task of spoken name recognition [9]. In order to emphasize the strength or weakness of two employing sets of speech parameters (in a view of speaker-and channel independence) we utilized an extremely simple training procedure: in every experiment only one sample of the name were used for training.

The efficiency of the suggested speaker - and channel-normalized set of formant parameters was tested by means of organizing two series of the experiments. In the first series a Russian names database was used. Ten speakers (8 male, 2 female) pronounced through telephone (PBX-system) their first, last and full names 10 times each and then they were recorded at 8 kHz speech signal sampling frequency. So, for all types of names the database includes 3,000 samples (100 samples per each type of personal name). This first part of the database was employed in testing. Another part of the database includes only one sample per each name spoken (a) by each of ten speakers using (a) electret PC microphone, or (b) PBX telephone for speech signal recording. This second part of the database was employed in the procedure of one sample training, and it provides 10*3*2 different recognition experiments for each type of parameters by using first part of the database. The results of the testing (average error rate for first, last and full name recognition) are shown in table 1.

Table 1. Average error rate (in %) for first, last and full name recognition (M1...M8 – male, F1, F2 – female)

Trained by the speaker		**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**F1**	**F2**	**Average**
Microphone training	Formant parameters	3.4	5.1	4.0	5.1	5.3	4.1	3.8	9.3	6.8	8.4	**5.5**
	Cepstral parameters	22.4	17.8	16.3	13.4	17.0	21.8	23.4	18.7	28.8	34.2	**21.4**
Telephone training	Formant parameters	2.1	2.9	3.1	2.2	3.5	4.1	2.6	5.2	3.4	4.5	**3.4**
	Cepstral parameters	10.1	7.4	9.7	8.7	9.9	10.1	18.1	13.6	23.9	26.1	**13.8**

In the second series of the experiments an American last names database taken from LDC telephone speech corpora [10] was used. This database contains thousands of names collected via toll-free telephone number throughout the United States. So, each spoken name is unique in a view of a speaker voice (male and female) as well as of telephone line condition. Five hundred last names pronounced one time each by the US-native male speaker through PC microphone were used for training. Then, using LDC-corpora the same five hundred last names were introduced for recognition within 50 portions each included 10 randomly chosen names. The average error rate was 24% for the formant and 53% for the cepstral parameters.

5. CONCLUSION

The results obtained from the experiments confirm the efficiency of the suggested set of formant parameters in a view of speaker – and channel variability resistance. According to the experiments the suggested set of formant parameters gives significant improvement in the recognition performance as compared with cepstral parameters using. Notice one more that in all the experiments that were carried out an extremely simple, one sample training procedure was used.

ACKNOWLEDGMENTS

This research was partly funded by NovCom NV as a part of COGNITEL project (see: www.novcom.com).

REFERENCES

1. M.Philips, J.Glass, and V.Zue. Automatic discovery of acoustic measurements for phonetic classification. In Proceedings ICASSP, 1992.
2. B. Lobanov. Classification of Russian Vowels Spoken by Different speakers. Journal Ac. Soc. of Am. V. 49, No 2, pp. 606-608, 1971.
3. M.J. Hunt. Dellayed Decisions in Speech Recognition – The Case of Formants. Pattern Recognition Letters, Vol. 6, pp. 121-137, July 1987.
4. L. Welling and H. Ney. A Model for Efficient Formant Estimation. Proc. ICASSP, pp. 797-800, Atlanta, 1996.
5. Y. Laprie and M.-O. Berger. Active Model for Regularizing Formant Trajectories. Proc. ICSLP, pp. 815-818, Banf, 1992.
6. J.N. Holmes, W.J. Holmes and P.N. Garner. Using Formant Frequencies in Speech Recognition. Proc. EuroSpeech'97, Rhodes – Greece, 1997.
7. D.X.Sun. Robust estimation of spectral center-of-gravity trajectories using mixture spline models. In Proceedings Eurospeech'95, V.1, Madrid, 1995.
8. B.Lobanov, T.Levkovskaya. Continuous speech recognizer for aircraft application. In Proceedings SPECOM'97, Clui-Napoca, Romania, 1997
9. B.Lobanov et al. An intelligent answering system using speech recognition. . In Proceedings Eurospeech'97, Rhodes - Greece, 1997
10. R.A.Cole at al. New telephone speech corpora at CSLU. In Proceedings Eurospeech'95, V.1, Madrid, 1995.