INFORMATICA, 2010, Vol. 21, No. 3, 361–374

Optimization of Formant Feature Based Speech Recognition

Antanas LIPEIKA

Institute of Mathematics and Informatics
Akademijos 4, LT-08663 Vilnius, Lithuania
e-mail: lipeika@ktl.mii.lt

Received: February 2010; accepted: June 2010

Abstract. The paper deals with the use of formant features in dynamic time warping based speech recognition. These features can be simply visualized and give a new insight into understanding the reasons of speech recognition errors. The formant feature extraction method, based on the singular prediction polynomials, has been applied in recognition of isolated words. However, the speech recognition performance depends on the order of singular prediction polynomials, whether symmetric or antisymmetric singular prediction polynomials are used for recognition and as well on the fact even or odd order of these polynomials is chosen. Also, it is important to know how informative separate formants are, how the speech recognition results depend on other parameters of the recognition system such as: analysis frame length, number of the formants used in recognition, frequency scale used for representation of formant features, and the preemphasis filter parameters. Properly choosing the processing parameters, it is possible to optimize the speech recognition performance.

The aim of our current investigation is to optimize formant feature based isolated word recognition performance by varying processing parameters of the recognition system as well as to find improvements of the recognition system which could make it more robust to white noise. The optimization experiments were carried out using speech records of 111 Lithuanian words. The speech signals were recorded in the conventional room environment (SNR = 30 dB). Then the white noise was generated at a predefined level (65 dB, 60 dB and 55 dB) and added to the test utterances. The recognition performance was evaluated at various noise levels.

The optimization experiments allowed us to improve considerably the performance of the formant feature based speech recognition system and made the system more robust to white noise.

Keywords: formant features, singular prediction polynomials, isolated word recognition, dynamic time warping, optimization.

1. Introduction

Speech recognition is widely used in various applications. Continuous speech recognition is used in dictation, information systems. Isolated word recognition is mostly applied in control by voice systems. In many applications control by voice can be accomplished using a simple speaker-dependent dynamic time warping technique. Thus, the focus of our research is reliable isolated word recognition, using the dynamic time warping approach.

The speech recognition performance highly depends on the features used for speech recognition. Traditional features, in this case, are the parameters of a linear prediction model (LPC features) and cepstral features. These features are sensitive, especially LPC, to background noise, recoding conditions, etc. However, transformations of LPC features can significantly improve recognition performance (Bastys *et al.*, 2010). We looked for an alternative feature set and chose estimates of formant frequencies as features for speech recognition. These features can be simply visualized and give us a new insight in understanding the reasons of speech recognition errors.

Several attempts were made to use formant features in HMM based speech recognition (De Wet *et al.*, 2004; Welling *et al.*, 1998; Weber *et al.*, 2001). The results have showed that inclusion of the formant features can increase the recognition accuracy.

Earlier we have investigated the use of formant features in dynamic time warping based speech recognition (Lipeika, 2005). The first three formants were used for the pattern matching. The formants were estimated from the roots of the 10th order LPC polynomials. The autocorrelation method was used to estimate the LPC parameters. The recognition results were encouraging; however, a more accurate method was desired for the formant frequency estimation.

The reliability of formant feature extraction depends on the method used for the linear prediction model parameter estimation. The autocorrelation method previously used for the linear prediction model parameter estimation was not reliable enough for formant feature extraction. Therefore we were looking for a more reliable method of the linear prediction model parameter estimation. The previously used autocorrelation method was compared with covariance (Markel and Gray, 1976), Burg (Kay and Marple, 1981), Marple (Marple, 1980) methods and the modified Split Levinson algorithm (Willems, 1987). It has been concluded that autocorrelation, covariance, Burg and Marple methods are similar from the point of view of formant feature extraction. The modified Split Levinson algorithm is a modification of the Split Levinson algorithm (Delsarte and Genin, 1986, 1987) and provided the best formant feature estimates (Lipeika, 2007).

The Split Levinson algorithm is based on the calculation of the singular prediction polynomials. If we have a set of the LPC polynomials

$$A_k(z) = 1 + a_k(1)z^{-1} + a_k(2)z^{-2} + \cdots + a_k(k)z^{-k}, \quad k = 1, \ldots, p, \tag{1}$$

the polynomials obey the recurrence relation

$$A_{k+1}(z) = A_k(z) + \rho_{k+1} z^{-(k+1)} A_k(z^{-1}), \tag{2}$$

where $\rho_1, \rho_2, \ldots, \rho_p$ are reflection coefficients. Assuming ρ_{k+1} to be either 1 or -1, from (1) we obtain two singular prediction polynomials:

$$\begin{aligned} P_{k+1}(z) &= A_k(z) + z^{-(k+1)} A_k(z^{-1}) \\ &= 1 + \big(a_k(1) + a_k(k)\big)z^{-1} + \big(a_k(2) + a_k(k-1)\big)z^{-2} + \cdots \\ &\quad + \big(a_k(k) + a_k(1)\big)z^{-k} + z^{-(k+1)} \end{aligned} \tag{3}$$

and

$$\begin{aligned}Q_{k+1}(z) &= A_k(z) - z^{-(k+1)}A_k(z^{-1})\\ &= 1 + \big(a_k(1) - a_k(k)\big)z^{-1} + \big(a_k(2) - a_k(k-1)\big)z^{-2} + \cdots\\ &\quad + \big(a_k(k) - a_k(1)\big)z^{-k} - z^{-(k+1)}.\end{aligned} \tag{4}$$

The polynomial $P_{k+1}(z)$ is symmetric, and $Q_{k+1}(z)$ is antisymmetric. Thus, we have

$$A_{k+1}(z) = 1/2\big[P_{k+1}(z) + Q_{k+1}(z)\big]. \tag{5}$$

Equations (3) and (4) mean that even without using the Split Levinson algorithm, we can calculate the pth order singular prediction polynomials $P_p(z)$ or $Q_p(z)$ from the $(p-1)$th order standard LPC polynomial, as well as the magnitude spectrum $S(f)$from these polynomials and find spectral peaks which are referred as estimates of formant frequencies. The formant feature extraction based on the calculation of the singular prediction polynomials from the standard LPC polynomials, the magnitude spectrum calculation, and the spectral peaking was implemented into a dynamic time warping based recognition system (Tamulevicius and Lipeika, 2004) and used for dynamic time warping based isolated word recognition experiments.

The speech recognition performance depends on whether symmetric or antisymmetric singular prediction polynomials are used for recognition (Lipeika and Lipeikienė, 2008) and even or odd order of these polynomials is chosen (Kabal and Ramachandran, 1986). Also, it is important to investigate how informative separate formants are and how speech recognition results depend on other parameters of the recognition system such as: analysis frame length, number the of formants used in recognition, frequency scale used for representation of formant features, and preemphasis filter parameters.

The aim of our current investigation is to optimize the formant feature based isolated word recognition performance by varying the above mentioned processing parameters of the recognition system and to find improvements of the recognition system which can make the recognition system more robust to white noise.

In the following section, the optimization experiments of formant feature based speech recognition are described. Section 3 concludes the paper.

2. Optimization Experiments of Formant Feature Based Speech Recognition

In Lipeika and Lipeikienė (2008) isolated word recognition experiments were done using the records of 111 Lithuanian words. The sessions were recorded in the conventional room environment (SNR = 30 dB). Then the white noise was generated at a predefined level (65 dB, 60 dB and 55 dB) and added to the test utterances. To make the results, obtained in (Lipeika and Lipeikienė, 2008), comparable with the current research, we used the same speech data in our experiments.

In Lipeika and Lipeikienė (2008), the recognition performance of the formant features, based on the singular symmetric prediction polynomials, was compared to that of

Table 1

Recognition errors (%) of 111 words using different single formants. The formant frequencies calculated from the 10th order symmetric singular prediction polynomials (9th order linear prediction polynomials)

Formants	SNR = 30 dB	+65 dB white noise
First formant	19.8	30.6
Second formant	1.8	28.8
Third formant	17.1	63.9
Fourth formant	17.1	87.3

the linear prediction coding, and the linear prediction coding derived cepstral features. The recognition results have shown that the performance of the formant features is similar to that of the LPC and the LPCC features in the conventional room environment. In the white noise environment, the results were different. Better results have been obtained using the formant features at all noise levels, but the number of recognition errors drastically increased when additional white noise was added to the test utterances. Our current investigation was to optimize the formant feature based isolated word recognition performance and to find improvements of the recognition system which can make the recognition system more robust to white noise.

2.1. *Recognition Performance Using Separate Formants*

Contribution of separate formants to the speech recognition performance is different. One can guess that lower formants contribute more the to speech recognition accuracy than the higher ones. Also, higher formants are more sensitive to noise. We estimated formant frequencies using the 10th order symmetric singular polynomials, and compared the recognition results, using different single formants. In this investigation, we used the test utterances recorded in the conventional room environment (SNR = 30 dB) and with additional 65 dB white noise. Here and in the following experiments the reference utterances were recorded in the conventional room environment (SNR = 30 dB). The recognition results are summarized in Table 1.

As we see from Table 1, the most informative is the second formant. The recognition performance of other formants is similar in the conventional room environment, but the third and especially the fourth formants are very sensitive to white noise. It is surprising that, using the second formant in the conventional room environment, only 1.8% recognition error was obtained.

2.2. *Recognition Performance Using Different Number of Formants*

The speech recognition performance depends on the number of formants used for recognition, order of singular prediction polynomials (even or odd) and on the fact which singular prediction polynomial is used – symmetric or antisymmetric. We investigated this dependence using the 8th, 9th and 10th order singular prediction polynomials. The

Table 2

Recognition errors (%) of 111 words using different number of formants. The formant frequencies calculated from the 8th order singular prediction polynomials (7th order linear prediction polynomials)

Polynomial, number of formants	SNR = 30 dB	+ 65 dB white noise
Symmetric, first 2 formants	0.9	16.2
Symmetric, first 3 formants	0	21.6
Antisymmetric, first 2 formants	2.7	23.4
Antisymmetric, first 3 formants	2.7	33.3

Table 3

Recognition errors (%) of 111 words using different number of formants. The formant frequencies calculated from the 9th order singular prediction polynomials (8th order linear prediction polynomials)

Polynomial, number of formants	SNR = 30 dB	+65 dB white noise
Symmetric, first 2 formants	0	13.5
Symmetric, first 3 formants	0.9	10.8
Symmetric, first 4 formants	1.8	28.8
Antisymmetric, first 2 formants	3.6	33.3
Antisymmetric, first 3 formants	1.8	32.4
Antisymmetric, first 4 formants	4.5	64.8

Table 4

Recognition errors (%) of 111 words using different number of formants. The formant frequencies calculated from the 10th order singular prediction polynomials (9th order linear prediction polynomials)

Polynomial, number of formants	SNR = 30 dB	+65 dB white noise
Symmetric, first 2 formants	2.7	11.7
Symmetric, first 3 formants	0	11.2
Symmetric, first 4 formants	0	19.8
Antisymmetric, first 2 formants	1.8	25.2
Antisymmetric, first 3 formants	0	19.8
Antisymmetric, first 4 formants	0.9	42.3

recognition results of the experiment for the 8th order are displayed in Table 2, for the 9th order – in Table 3, and for the 10th order – in Table 4.

From Table 2 one can see that for the 8th order of singular prediction polynomials (7th order linear prediction polynomials), the common best result (for SNR = 30 dB and when 65 dB white noise is added) is obtained when the symmetric singular prediction polynomials and the first 2 formants are used for recognition. For the 9th (Table 3) and 10th (Table 4) order singular prediction polynomials (8th and 9th order linear prediction polynomials) the common best result is obtained when the symmetric singular prediction polynomials and the first 3 formants are used for recognition. The common best result

Table 5

Recognition errors (%) of 111 words using different frame length. The formant frequencies calculated from the 10th order symmetric singular prediction polynomials (9th order linear prediction polynomials)

Frame length	SNR = 30 dB	+65 dB white noise	+60 dB white noise	+55 dB white noise
$N = 250$	0	11.2	29.7	63.9
$N = 400$	0	10.8	27.0	55.8
$N = 500$	0	8.1	27.9	51.3
$N = 600$	0.9	9.0	27.0	51.3

(recognition error 11.2%) is obtained when formant frequencies are calculated from 10th order symmetric singular polynomials (9th order linear prediction polynomials) and the first 3 formants are used for recognition.

When two formants are used for recognition, the common best result (recognition error 13.5% for SNR = 30 dB and when 65 dB white noise is added) is obtained when formant frequencies are calculated from 9th order symmetric singular polynomials (8th order linear prediction polynomials). Concerning model complexity, when we use two formants in recognition instead of three, we simply do not include the third formant into the feature vector. It helps to save memory consumed for reference patterns.

2.3. *Recognition Performance Using Different Analysis Frame Length*

In the dynamic time warping speech recognition system (Tamulevicius and Lipeika, 2004), using the linear prediction (LPC) or LPC derived cepstral features, traditionally the frame length of 250 data points (22 ms) was used for the analysis at 11025 Hz sampling frequency. However, when formant features were implemented in this system, we have noticed that formant tracks are not always smooth enough, especially when noise is added. We tried to increase the analysis frame length. For the 10th order symmetric singular prediction polynomials, we investigated the recognition performance at different analysis frame length. The first three formants were used for recognition. The results of the experiment are displayed in Table 5.

From Table 5 we can notice that in the formant frequency based recognition it is better to increase the frame length up to $N = 500$.

2.4. *Recognition Performance Calculating Formant Frequencies in the Mel-Scale*

We noticed that formant tracks of higher formants were more scattered than that of lower formants. So we tried to represent formant tracks in the Mel-scale which corresponds to the auditory sensation of tone height (Furui, 2001). The relationship between the frequency f in kilohertz and the Mel-scale is approximated by the equation

$$\mathrm{Mel} = 1000 \log_2(1 + f). \tag{6}$$

Table 6

Recognition errors (%) of 111 words using formant frequency representation in the Mel frequency scale. The formant frequencies calculated from the 9th and 10th order symmetric singular prediction polynomials (9th order linear prediction polynomials). Frame length $N = 250$ and $N = 500$.

Frame length N Order of the polynomial M	SNR = 30 dB	+65 dB white noise	+60 dB white noise	+55 dB white noise
$N = 250,\ M = 9$	0	7.2	24.3	56.7
$N = 250,\ M = 10$	0	6.3	23.4	47.7
$N = 250,\ M = 10$ Linear scale	0	11.2	29.7	63.9
$N = 500,\ M = 9$	0	9.9	21.6	52.2
$N = 500,\ M = 10$	0	5.4	22.5	45.0
$N = 500,\ M = 10$ Linear scale	0	8.1	27.9	51.3

We represented formant frequencies in the Mel-scale and investigated the recognition performance for the 9th and 10th order symmetric singular prediction polynomials and for different frame lengths $N = 250$ and $N = 500$. The first three formants were used for recognition. The performance was investigated at different noise levels and compared with the results obtained in the linear frequency scale. The results of experiments are displayed in Table 6.

Comparing the recognition results we can notice that with frame length $N = 250$, in all cases the recognition results for the Mel-scale representation are better. With $N = 500$, the results are better than that of the linear scale only for the 10th order singular prediction polynomials at all noise levels. For frame length $N = 500$ and the 9th order singular prediction polynomials at higher noise levels, the recognition results are worse than that of the linear frequency scale.

2.5. *Impact of Preemphasis on the Recognition Performance*

The preemphasis filter is usually used in speech recognition to flatten the speech signal spectrum and to make a speech signal less sensitive to finite precision effects later in the speech signal processing. In our speech recognition system (Tamulevicius and Lipeika, 2004), a first order finite impulse response filter with the system function

$$H(z) = 1 - p_1 z^{-1} = 1 - 0.95z^{-1} \tag{7}$$

was used for preemphasis. Amplitude response and the pole/zero plot of the filter are shown in Fig. 1.

We used the 9th order symmetric singular prediction polynomials and 3 formants for the recognition experiment. The analysis frame length was $N = 500$. These conditions were kept in all the following experiments. The recognition results are displayed in Table 7.

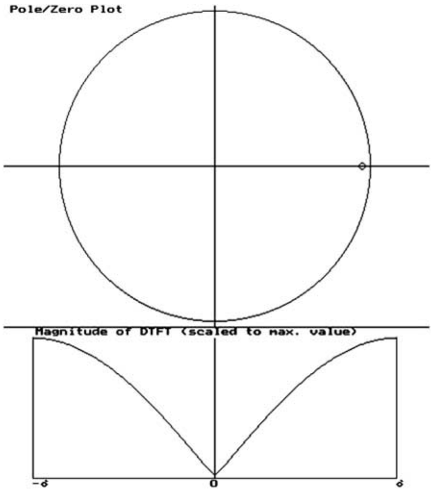

Fig. 1. Amplitude response and the pole/zero plot of the preemphasis filter (7).

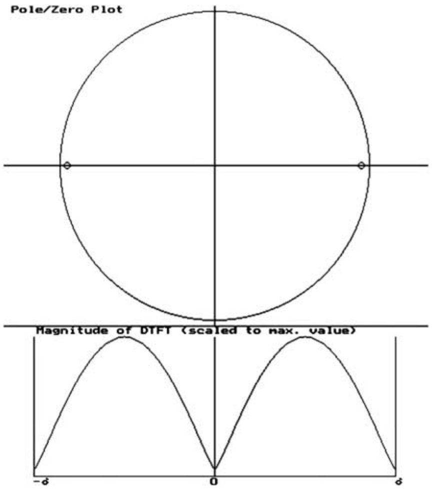

Fig. 2. Amplitude response and the pole/zero plot of the preemphasis filter (8).

Next, to reduce the impact of white noise on recognition errors, we included additional real zero $p_2 = -0.95$ in to the system function of the preemphasis filter getting the system function

$$H(z) = (1 - p_1 z^{-1})(1 - p_2 z^{-1}) = (1 - 0.95z^{-1})(1 + 0.95z^{-1}). \tag{8}$$

Amplitude response and the pole/zero plot of the filter are shown in Fig. 2.

The recognition results are displayed in Table 7. From the recognition results we can notice that the recognition performance has improved at all noise levels but slightly decreased at SNR = 30 dB.

Encouraged by the results, we have placed two additional complex conjugate zeroes near the unit circle at frequencies close to $-\pi$ getting the filter below (Fig. 3)

$$H(z) = (1 + 1.959998z^{-1} + 0.067506z^{-2} - 1.769247z^{-3} - 0.876533z^{-4}). \quad (9)$$

The recognition results are displayed in Table 7. They illustrate that the recognition performance has improved under additional white noise with SNR = 60 and 55 dB.

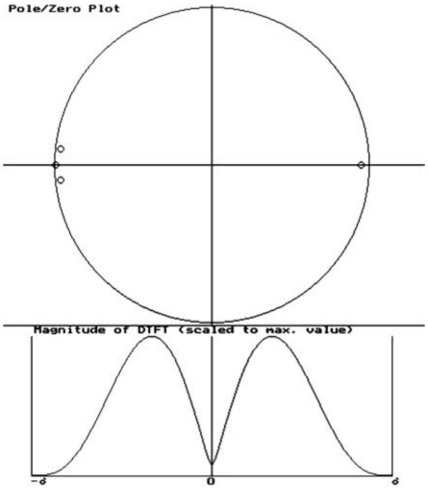

Fig. 3. Amplitude response and the pole/zero plot of the preemphasis filter (9).

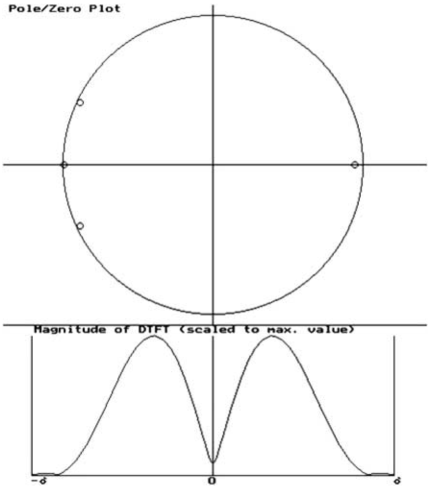

Fig. 4. Amplitude response and the pole/zero plot of the preemphasis filter (10).

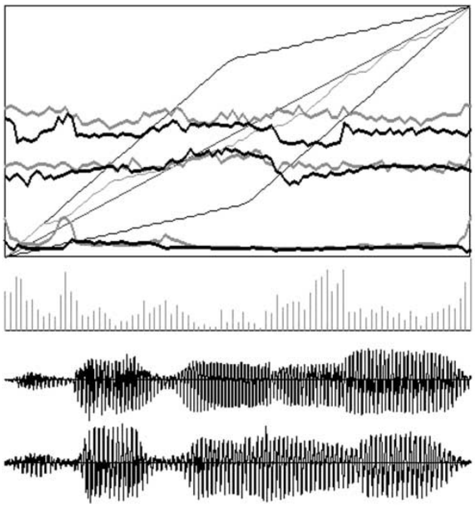

Fig. 5. Illustration of the formant feature based dynamic time warping process using the traditional preemphasis filter (7). Two utterances of the Lithuanian word "devyni" are compared. The additional 55 dB white noise was added to the test utterance. The average distance between the test and reference templates is 57.2. Black lines are formants of the reference templates and grey lines of the test templates. The word was not recognized.

Finally, we shifted complex conjugate zeroes farther from $-\pi$ getting the filter (Fig. 4)

$$H(z) = (1 + 1.799998z^{-1} + 0.072006z^{-2} - 1.618268z^{-3} - 0.886789z^{-4}) \quad (10)$$

and repeated the experiment. The recognition results (Table 7) have showed a further improvement at additional white noise with SNR $=$ 60 and 55 dB. The best result was obtained using the preemphasis filter (10).

To illustrate the difference between the traditional preemphasis filter (7) and the preemphasis filter (10), two utterances of the Lithuanian word "devyni" are compared using the traditional preemphasis filter (7) (Fig. 5) and preemphasis filter (10) (Fig. 6). The additional 55 dB white noise was added to the test utterance. For preemphasis filter (7), the average distance between the test and reference templates is 57.2. The word was not recognized. Using preemphasis filter (10) the word was recognized. The average distance is 25.5. The global path constraints, a linear time alignment path, a dynamic time warping path, and the formant trajectories (black – trajectories of the reference templates, grey – of the test templates) are displayed in the picture. This constitutes the dynamic time warping (DTW) picture. Below the DTW picture, local distances on the optimal DTW path and both the reference and the test utterances are displayed.

In Fig. 5, the estimates of formant trajectories of the test templates are shifted from true values due to the impact of additional white noise. In Fig. 6, the preemphasis filter (10) compensates the impact of white noise and formant trajectories of the test template are close to that of the reference template.

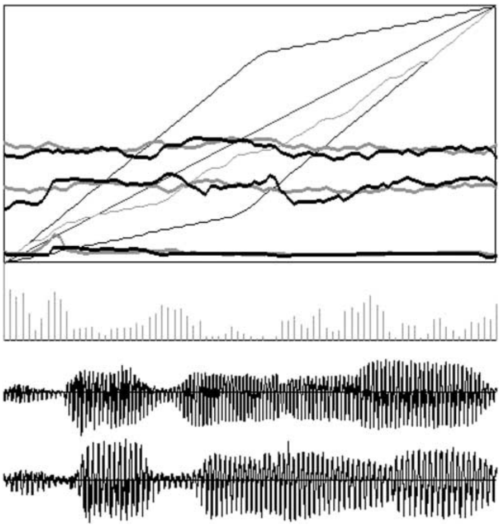

Fig. 6. Illustration of the formant feature based dynamic time warping process using preemphasis filter (10). Two utterances of the Lithuanian word "devyni" are compared. The additional 55 dB white noise was added to the test utterance. The average distance between the test and reference templates is 25.5. Black lines are formants of the reference templates and grey lines of the test templates. The word was recognized.

Table 7

Recognition errors (%) of 111 words using different preemphasis filters. Formant frequencies calculated from the 9th order symmetric singular prediction polynomials (8th order linear prediction polynomials). Frame length $N = 500$. Three formants used for recognition

Preemphasis filter	SNR $=$ 30 dB	$+65$ dB noise	$+60$ dB noise	$+55$ dB noise
(7)$H(z) = 1 - p_1 z^{-1} = 1 - 0.95z^{-1}$	0	11.7	26.1	56.7
(8)$H(z) = (1 - 0.95z^{-1})(1 + 0.95z^{-1})$	0.9	4.5	18.0	37.8
(9)$H(z) = (1 + 1.959998z^{-1} + 0.067506z^{-2} - 1.769247z^{-3} - 0.876533z^{-4})$	0	4.5	12.6	29.7
(10)$H(z) = (1 + 1.799998z^{-1} + 0.072006z^{-2} - 1.618268z^{-3} - 0.886789z^{-4})$	0	4.5	11.7	27.0
LPC features (10th order)	0	42.3	63.0	85.5
LPC derived cepstral features (15th order)	0	38.7	62.1	81.9
LPC derived cepstral features with cepstral mean subtraction (15th order)	0	7.2	13.5	35.1

We repeated the same experiments using traditional features: LPC, LPC derived cepstral features and LPC derived cepstral features with cepstral mean subtraction. The results are also displayed in Table 7.

Basing on the results of the recognition experiments (Table 7), we can conclude that, using the preemphasis filter (10), we have obtained the best result at all noise levels.

When a speech signal is corrupted with white noise, the recognition performance at all noise levels is higher using formant features and preemphasis filter (10) than the best result for the traditional feature set – LPC derived cepstral features with cepstral mean subtraction.

3. Conclusions and Future Work

A possibility to optimize the formant feature based isolated word recognition has been investigated. The singular prediction polynomials were used for formant feature extraction. However, the speech recognition performance depends on order of singular prediction polynomials, whether symmetric or antisymmetric singular prediction polynomials are used for recognition as well as on the fact even or odd order of these polynomials is chosen. Also, it is important to investigate how informative separate formants are, how the speech recognition results depend on other parameters of the recognition system, such as analysis frame length, number of the formants used in recognition, frequency scale used to represent the formant features and the preemphasis filter parameters.

The optimization experiments were carried out using speech records of 111 Lithuanian words. The speech signals were recorded in the conventional room environment (SNR = 30 dB). Then the white noise was generated at a predefined level (65 dB, 60 dB and 55 dB) and added to the test utterances. The recognition performance was evaluated at various noise levels. The optimization experiments have shown that:

1. Most informative is the second formant. The recognition performance of other formants is similar in the conventional room environment, but the third and, especially, fourth formants are very sensitive to the white noise.
2. Recognition performance was investigated for the 10th, 9th and 8th order of singular prediction polynomials. The common best result (recognition error 11.2%) is obtained when formant frequencies are calculated from 10th order symmetric singular polynomials (9th order linear prediction polynomials) and the first 3 formants are used for recognition. The similar result is obtained for the 9th order of symmetric singular prediction polynomials (8th order linear prediction polynomials).

   When two formants are used for recognition, the common best result (recognition error 13.5% for SNR = 30 dB and when 65 dB white noise is added) is obtained when formant frequencies are calculated from 9th order symmetric singular polynomials (8th order linear prediction polynomials).
3. In the formant frequency based recognition, it is better to increase the frame length up to $N = 500$.
4. Calculating the formant frequencies in the Mel frequency scale with frame length $N = 250$, in all cases (of the 10th and 9th order of the singular prediction polynomials) the recognition results are better. At frame length $N = 500$, the recognition results are better than that of the linear scale only for the 10th order singular prediction polynomials at all noise levels.
5. The recognition results highly depend on the parameters of the preemphasis filter. Four different preemphasis filters were used in the recognition experiments. The

best recognition results were obtained using the preemphasis filter (10) at all noise levels.

6. The recognition results using formant features were compared with that obtained using the traditional features: LPC, LPC derived cepstral features, and LPC derived cepstral features with cepstral mean subtraction. Using the preemphasis filter (10) for formant features, the recognition results are better even than that obtained for the LPC derived cepstral features with cepstral mean subtraction for the traditional features.

The optimization experiments enabled us to considerably improve the performance of the formant feature based speech recognition system.

Future work should be concentrated on including pitch information in the feature set, since voiced/unvoiced discrimination can also improve the recognition performance.

References

Bastys A., Kisel, A., Šalna, B. (2010). The use of group delay features of linear prediction model for speaker recognition. *Informatica*, 21(1), 1–12.

Delsarte, P., Genin, Y.V. (1986). The split Levinson algorithm. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-35(5), 645–653.

Delsarte, P., Genin, Y.V. (1987). On the splitting of classical algorithms in linear prediction theory. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-34(3), 470–478.

De Wet F., Weber, K., Boves, L., Cranen, B., Bengio, S., Bourland, H. (2004). Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America*, 116(3), 1781–1792.

Furui, S. (2001). *Digital Speech Processing, Synthesis, and Recognition*. Dekker, New York.

Kabal, P., Ramachandran, R.P. (1986). The computation of line spectral frequencies using Chebyshev polynomials. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-34(6), 1419–1426.

Kay, S.M., Marple, S.L. (1981). Spectrum analysis – a modern perspective. *Proc. IEEE*, 11, 1380–1419.

Lipeika, A. (2005). Formant features in speech recognition. *Information Sciences*, 34, Vilnius University Publishing House, 215–219 (in Lithuanian).

Lipeika, A. (2007). Formant feature extraction methods. *Information Sciences*, 42–43, Vilnius University Publishing House, 201–206 (in Lithuanian).

Lipeika, A., Lipeikienė, J. (2008). On the use of the formant features in the dynamic time warping based recognition of isolated words. *Informatica*, 19(3), 213–226.

Markel, J.D., Gray, A.H. (1976). *Linear Prediction of Speech*. Springer, Berlin.

Marple, L. (1980). A new autoregressive spectrum analysis algorithm, *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-28(4), 441–454.

Tamulevičius, G., Lipeika, A. (2004). Dynamic time warping based speech recognition system. In: *The First Baltic Conference "Human language technologies"*, Riga, pp. 156–161.

Weber, K., Bengio, S., Bourland, H. (2001). HMM2-extraction of formant structures and their use for robust ASR. In: *Proceedings of Eurospeech*, pp. 607–610.

Welling, L., Ney, H. (1998). Formant estimation for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 6(1), 36–48.

Willems, L.F. (1987). Robust formant analysis for speech synthesis applications. *Proceedings of the European Conference on Speech Technology*, 1, 250–253.

A. Lipeika is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics, a professor of the Fundamental Sciences Department at the Vilnius Gediminas Technical University and a professor of the Information Technology Department at the Vilnius Pedagogical University. His scientific interests include digital signal processing, pattern recognition, speech processing, speech and speaker recognition.

Formantiniais požymiais grįsto kalbos atpažinimo optimizavimas

Antanas LIPEIKA

Darbe yra nagrinėjamas formantinių požymių naudojimas dinaminiu laiko skalės kraipymu grįstame kalbos atpažinime. Šie požymiai yra lengvai vizualizuojami ir suteikia naujų žinių stengiantis suvokti kalbos atpažinimo klaidų priežastis. Izoliuotų žodžių atpažinimui buvo pritaikytas išsigimusiais prognozės polinomais grįstas formantinių požymių išskyrimo metodas. Tačiau, kalbos atpažinimo tikslumas priklauso nuo to, ar atpažinimui naudojami išsigimę prognozės polinomai yra simetriniai ar antisimetriniai ir ar jie yra lyginės ar nelyginės eilės. Taip pat svarbu ištirti, kiek informatyvios yra atskiros formantės, kaip kalbos atpažinimo rezultatai priklauso nuo kitų atpažinimo sistemos parametrų: analizės kadro ilgio, atpažinimui naudojamų formančių skaičiaus, formantinių požymių atvaizdavimui naudojamos dažnių skalės, pradinės filtracijos filtro parametrų. Tinkamai parinkus šiuos atpažinimo sistemos parametrus galima pagerinti sistemos atpažinimo tikslumą.

Šio tyrimo tikslas yra keičiant atpažinimo sistemos parametrus optimizuoti formantiniais požymiais grįsto izoliuotų žodžių atpažinimo tikslumą ir surasti atpažinimo sistemos patobulinimo būdus, kurie padidintų sistemos atsparumą balto triukšmo poveikiui. Optimizavimo eksperimentai buvo atliekami naudojant 111 lietuvių kalbos žodžių kalbos signalo įrašus. Kalbos signalai buvo įrašyti įprastinėje kambario aplinkoje esant 30 dB signalo/triukšmo santykiui. Tada buvo generuojamas 65 dB, 60 dB ir 55 dB baltas triukšmas ir pridedamas prie testinių ištarimų balso įrašų. Atpažinimo tikslumas buvo vertinamas esant įvairiems užtriukšminimo lygiams.

Atlikti optimizavimo eksperimentai padėjo žymiai pagerinti formantiniais požymiais grįstos kalbos atpažinimo sistemos tikslumą ir padaryti ją atsparesnę balto triukšmo poveikiui.