INFORMATICA, 2008, Vol. 19, No. 2, 213–226

On the Use of the Formant Features in the Dynamic Time Warping Based Recognition of Isolated Words

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

Received: June 2007

Abstract. A possibility to use the formant features (FF) in the user-dependent isolated word recognition has been investigated. The word recognition was performed using a dynamic time-warping technique. Several methods of the formant feature extraction were compared and a method based on the singular prediction polynomials has been proposed for the recognition of isolated words. Recognition performance of the proposed method was compared to that of the linear prediction coding (LPC) and LPC-derived cepstral features (LPCC). In total, 111 Lithuanian words were used in the recognition experiment. The recognition performance was evaluated at various noise levels. The experiments have shown that the formant features calculated from the singular prediction polynomials are more reliable than the LPC and LPCC features at all noise levels.

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

1. Introduction

Speech recognition systems usually demonstrate a good performance under laboratory conditions, but often fail under real, noisy conditions. This is due to the features used for the speech recognition. They are not sufficiently robust in noisy environments. Even the most-popular Mel-frequency cepstral features (Davis and Mermelstein, 1980) often do not meet the performance requirements. Researchers are looking for the new kinds of features that could be more reliable for the speech recognition.

It is known (Huang *et al.*, 2001) that expert spectrogram readers are able to recognize speech by looking at a spectrogram, particularly at the formants. However, it is not clear whether the formants are very useful features for the speech recognition. Numerous attempts have been made to find low-dimensional, formant-related representations of speech signals that are suitable for the automatic speech recognition (ASR). For example, formant-like features were evaluated on the vowel classification task (De Wet *et al.*, 2004). The results showed that, for both clean and noisy data, the Mel-frequency cepstral features achieved the same or superior results as the formant features. The price for that was a much higher dimensionality of the Mel-frequency cepstral features. Similar results were obtained in (Welling and Ney, 1998). In (Holmes *et al.*, 1997) and (Garner and Holmes, 1998), the recognition results were compared between a recognizer that used the conventional cepstrum features and the one that used three formant frequencies, combined with fewer cepstrum features. These results showed that the inclusion of the formant features can increase the recognition accuracy when the total number of features is fixed. The formant features were also used for the speech recognition in fixed and cellular network applications (De Wet *et al.*, 2000). Weber *et al.* (2001) investigated the potential of formant information to enhance state-of-the-art noise-robust HMM-based ASR. The HMM-based speech recognition was used in all papers mentioned above.

A dynamic time warping (DTW) approach was used in the formant-based recognition of isolated words in Lipeika (2005). 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 experiment consisted of 12 Lithuanian words. The recognition results were encouraging; however, a more accurate method for the formant frequency estimation was desired. The following part of the paper presents attempts to get a more suitable formant estimation method for the formant feature based recognition of isolated words.

2. Formant Feature Estimation

One way to obtain estimates of the formants at a frame level is to compute the roots of the pth order LPC polynomial (Huang *et al.*, 2001). Each complex root z_i can be represented as

$$z_i = \exp(-\pi b_i + j2\pi f_i), \tag{1}$$

where f_i and b_i are the respective estimates of the formant frequency and the formant bandwidth, corresponding to the ith root. Real roots are discarded. Complex roots are sorted by increasing f, discarding negative values.

An alternative approach (Markel, 1972; McCandless, 1974) is to use the peaks of the linear prediction spectra. This method, however, has drawbacks. Frequently, two peaks merge, or spurious peaks appear. It is difficult to recognize such situations and to deal with them.

Various approaches were used to improve the formant frequency estimation. Snell and Milinazzo (1993) derived a new method for locating roots within the unit circle. Welling and Ney (1996) proposed a new method for the estimation of the formant frequencies. There the formant model is based on a digital resonator. The complete spectrum is modeled by a set of digital resonators connected in parallel. An algorithm based on dynamic programming produces both the model parameters and the segment boundaries that optimally match the spectrum. Yegnanarayana and Veldhuis (1998) investigated the formant extraction from the voiced speech consistent across successive pitch periods. Acero (1999) used hidden Markov models for the formant tracking. The Viterbi search is carried out to find the most likely path of the formants. Watanabe (2001) proposed a new method to estimate the formant frequencies. It is based on the inverse-filter control and the zero-crossing frequency distributions. In total, 32 basic inverse filters that are mutually controlled by the weighted means of the zero-crossing frequency distributions are used. Diankha and Shimamura (2002) proposed an improved method for the formant frequency estimation based on the complex autocorrelation function of the speech signal. Instead of using the incoming signal as an input for the LPC analysis, the analytic signal of the autocorrelation function of the speech signal is computed and used itself as an input for the LPC analysis.

We concentrated our attention on improving the LPC-based formant frequency estimation. We looked for a method to estimate the LPC parameter that would yield formant positions more reliably. The previously used autocorrelation method was compared to the covariance method (Markel and Gray, 1976), the Marple method (Marple, 1980) and the modified Split Levinson algorithm (Willems, 1987) named as the RF (robust formants) algorithm. These methods were compared with the Praat software (Boersma *et al.*). The autocorrelation method, the covariance method and the Marple method demonstrated similar performance. On the other hand, the RF algorithm produced more-stable estimates of the formant frequencies, and it always yielded the prescribed number of formants. Instead of directly applying the root solving procedure to a standard LPC polynomial to obtain the frequency positions of the spectral maxima, a so-called singular predictor polynomial is constructed, from which the zeroes are determined by an iterative procedure (De Wet *et al.*, 2004). All zeroes of this singular predictor polynomial lie on the unit circle, thus the number of the found maxima is guaranteed to be equal to a half of the LPC order under all circumstances. The maxima that are located in this manner are referred to as "formants" found by the RF algorithm. The roots of the singular predictor polynomial are close to the roots found in the classical root solving procedure, provided that these are close to the unit circle. This property ensures that the most important formants are properly represented.

The RF algorithm has two major advantages over the standard root solving of the LPC polynomial or the search for the maxima in the spectral envelope derived from the LPC coefficients (De Wet *et al.*, 2004). First, the algorithm guarantees to find a fixed number of complex poles corresponding to the formants for each speech frame. This prevents from the labeling errors (e.g., F3 labeled as F2), because there are no missing formants. Second, the algorithm tends to distribute the complex poles along the unit circle uniformly and the formant tracks are guaranteed to be fairly smooth and continuous, as one would expect the vocal tract resonances to be. The formant frequency estimation algorithm is proposed below.

3. Formant Frequency Estimation Algorithm

The Split Levinson algorithm (Delsarte and Genin, 1986; 1987) to estimate the LPC parameters was developed aiming to reduce the computational amount of the autocorrelation method used for the estimation of the LPC parameters. Replacing the standard

Levinson algorithm (Hayes, 1996) in the autocorrelation method by the Split Levinson algorithm, we obtain the same estimate of the LPC parameters; however, the number of multiplications and additions is reduced by approximately a factor of two.

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

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

they obey the recurrence relation (Delsarte and Genin, 1986)

$$A_{k+1}(z) = A_k(z) + \rho_{k+1} z^{-(k+1)} A_k(z^{-1}), \tag{3}$$

where $\rho_1, \rho_2, \ldots, \rho_p$ are the reflection coefficients. Assuming ρ_{k+1} to be either 1 or -1, we obtain two singular prediction polynomials from (2):

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

and

$$\begin{aligned} Q_{k+1}(z) &= A_k(z) - z^{-(k+1)} A_k(z^{-1}) \\ &= 1 + \big(a_k(1) - a_k(k)\big)z^{-1} + \big(a_k(2) - a_k(k-1)\big)z^{-2} \\ &\quad + \cdots + (a_k(k) - a_k(1))z^{-k} - z^{-(k+1)}. \end{aligned} \tag{5}$$

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

$$A_{k+1}(z) = 1/2\big[P_{k+1}(z) + Q_{k+1}(z)\big]. \tag{6}$$

Eqs. 4 and 5 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. The analysis can proceed twofold:

- Use some standard algorithm to compute the complex roots (1) of the LPC polynomial with real coefficients (Press *et al.*, 1988). The roots in the range of $0 < f_i < 1/2$, are regarded as the formant frequency estimates.
- Compute the magnitude spectrum of a singular prediction polynomial and find the spectral peaks.

If the singular prediction polynomial is

$$P_p(z) = 1 + s(1)z^{-1} + s(2)z^{-2} + \cdots + s(p)z^{-p}, \tag{7}$$

then the magnitude spectrum $S(f)$ can be calculated using the following expression

$$S(f) = \frac{1}{|1 + s_1 e^{-j2\pi f} + s_2 e^{-j4\pi f} + \cdots + s_p e^{-j2p\pi f}|}, \quad 0 < f < 1/2. \tag{8}$$

4. Experiments on Recognition of Isolated Words

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). It was developed earlier to use the LPC and LPC-derived cepstral features (LPCC). We use the system for the recognition of isolated words.

To illustrate the formant feature based dynamic time warping method, two utterances of the Lithuanian word "mokslas" (science) is presented in Fig. 1. The word was pronounced by the same speaker. For the analysis, symmetric singular prediction polynomials were calculated. The first two formants F_1 and F_2 were used for the pattern matching. The average distance between the formant frequencies of the two words was 23.6 Hz. The global path constraints, a linear time alignment path, a dynamic time warping path, and the formant trajectories are displayed in the picture. This constitutes the dynamic time-warping (DTW) picture. Below the DTW picture, local distances on the DTW path and both the reference and the test utterances are displayed.

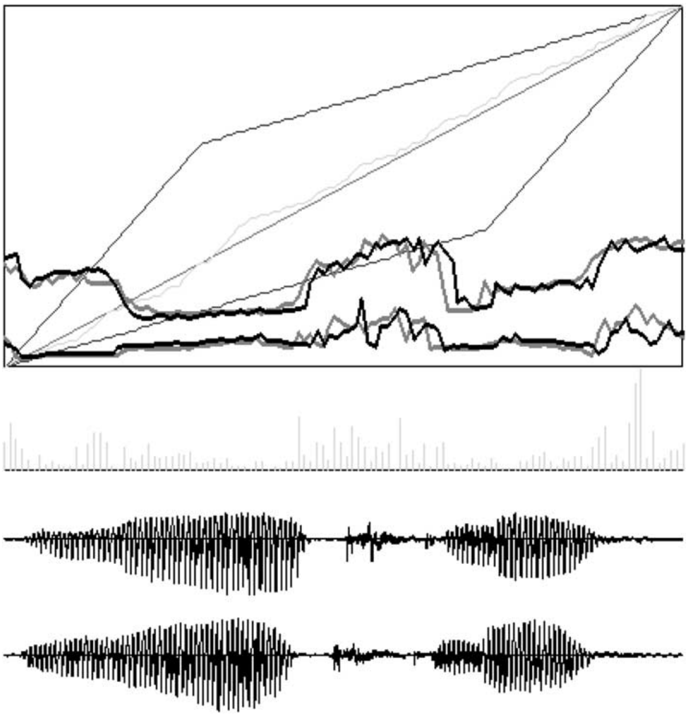

Fig. 1. Illustration of the formant feature-based dynamic time warping process. Two utterances of the Lithuanian word "mokslas" are compared. The average distance is 23.6 Hz. Black lines are the reference templates, and grey lines are the test templates.

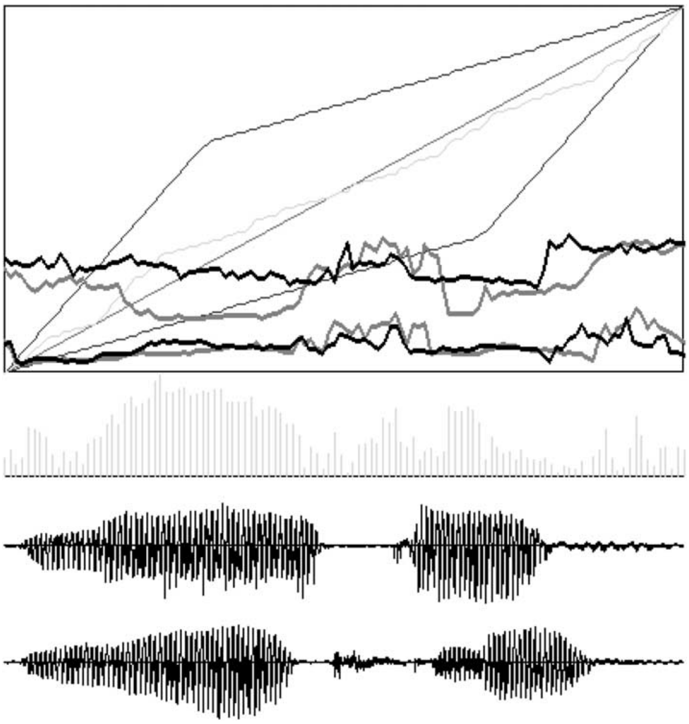

Fig. 2. Illustration of the formant feature-based dynamic time warping process. The utterances of the Lithuanian words "mokslas" and "metas" are compared. The average distance is 80.1 Hz. Black lines are the reference templates, and grey lines are the test templates.

This illustration can be compared to the one presented in Fig. 2. There a dynamic time warping of two different Lithuanian words "mokslas" and "metas" (time) is displayed. The average distance between the formants is much larger. It amounts to 80.1 Hz. An important advantage of the formant features is the simplicity to visualize and interpret the recognition results. One can see the parts of the word that are matched better or worse and investigate the influence of the recoding conditions on the recognition results.

4.1. *The Data*

To compare the formant feature (FF) based recognition to the LPC and LPCC feature based recognitions at different noise levels, several speech recognition experiments were performed. In total, 111 Lithuanian words were used for recognition. Speech recordings were taken from the Lithuanian speech corpus which was collected for small vocabulary, isolated word recognition experiments. Utterances of digits from 0 to 9, Lithuanian words "pradžia" (start), "pabaiga" (end) and 99 most-frequent Lithuanian words chosen from the frequency dictionary of the modern written Lithuanian (Grumadienė and Žilinskienė, 1997) were recorded in this corpus. Recordings were made by students under various recording conditions (mostly at home) and using the sampling rate of 11025 Hz and the resolution of 16 bits. This sampling rate and the resolution were the ones used in the dynamic time-warping based isolated word recognition system (Tamulevicius and Lipeika, 2004) which was adopted for the formant feature based isolated word recognition and in other systems (Lipeika *et al.*, 2002). Although such sampling rate introduces additional distortions in the speech spectrum, it allows to increase the recognition speed of the real-time speech recognition system and is typically used in the dynamic time warping-based isolated word recognition systems (Rabiner and Juang, 1993).

4.2. *Speech Recognition Experiments*

To test the approach, two sessions of the same male speaker were used in the experiments: one for reference templates, another for testing. 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. The analysis parameters were following: the frame length was 250 (22 msec); the frame step was 125 (11 msec); the preemphasis was 0.95. The hamming window was applied for each analysis frame. The 10th order LPC analysis was used and the 10th order symmetric singular prediction polynomials were calculated for the formant feature extraction. In total, 400 values of the magnitude spectrum of a singular prediction polynomial were calculated for each analysis frame. The LPCC analysis considered 15 cepstral coefficients. The FF-based recognition used either two formants $(F_1$ and $F_2)$ or three formants $(F_1, F_2$ and $F_3)$. The results of the experiment are summarized in Table 1.

As we see from Table 1, the performance of the formant features using three formants is the same to that of LPC and LPCC features in the conventional room environment. Therefore it is preferable to use three formants for recognition under those conditions. In a noisy environment the results are different. The use of the formant features gives better results at all noise levels.

The recognition results based on the symmetric singular prediction polynomial calculation to extract the formant features were compared to the results based on the cal-

Table 1

Recognition errors (%) of 111 words at different SNR levels. Four different features were applied: the formant frequencies (first two or three formants) calculated from the symmetric singular prediction polynomials (FFSSPP), the LPC features (LPC), and the LPC-derived cepstral features (LPCC)

Recognition conditions	Recognition error (%)			
	FFSSPP, 2 formats: F_1, F_2	FFSSPP, 3 formats: F_1, F_2, F_3	LPC features	LPCC features
(SNR = 30 dB)	2.7	0	0	0
+65 dB white noise	11.7	11.7	42.3	38.7
+60 dB white noise	21.6	29.7	63.0	62.1
+55 dB white noise	51.3	63.9	85.5	81.9

Table 2

Recognition errors (%) of 111 words at different SNR levels using the formant frequencies (first two or three formants) calculated from symmetric singular prediction polynomials (FFSSPP) and the formant features calculated from the standard LPC polynomials estimated by the autocorrelation method (FFLPC)

Recognition conditions	Recognition error (%)			
	FFSSPP, 2 formats: F_1, F_2	FFSSPP, 3 formats: F_1, F_2, F_3	FFLPC 2 formats: F_1, F_2	FFLPC 3 formats: F_1, F_2, F_3
(SNR = 30 dB)	2.7	0	18.9	17.1
+65 dB white noise	11.7	11.7	30.6	39.6
+60 dB white noise	21.6	29.7	42.3	47.7
+55 dB white noise	51.3	63.9	65.7	64.8

culation of the formant features from the standard LPC polynomials estimated by the autocorrelation method. The recognition results are summarized in Table 2.

Table 2 illustrates that in the conventional room environment (SNR = 30 dB) the recognition results are poor if the formant features, calculated from the standard LPC polynomials estimated by the autocorrelation method, are used. The difference in the recognition errors between the symmetric singular prediction polynomial and the LPC polynomial decreases as the noise level increases. However, even at the highest noise level the recognition using the singular prediction polynomial is more accurate.

The formant feature extraction based on the antisymmetric singular prediction polynomial calculation is also possible. We have compared the recognition results for the formant feature extraction based on the symmetric singular prediction polynomial calculation with that based on the antisymmetric singular prediction polynomial calculation. The results are summarized in Table 3.

Table 3

Recognition errors (%) of 111 words at different SNR levels using the formant frequencies calculated from the symmetric singular prediction polynomials (FFSSPP) and the formant frequencies calculated from the antisymmetric singular prediction polynomials (FFASPP)

Recognition conditions	Recognition error (%)			
	FFSSPP, 2 formats: F_1, F_2	FFSSPP, 3 formats: F_1, F_2, F_3	FFASPP 2 formats: F_1, F_2	FFASPP 3 formats: F_1, F_2, F_3
(SNR = 30 dB)	2.7	0	1.8	0
+65 dB white noise	11.7	11.7	25.2	19.8
+60 dB white noise	21.6	29.7	42.3	44.1
+55 dB white noise	51.3	63.9	60.3	68.4

Based on the results of the recognition experiments we conclude that the performance of both methods is similar in the conventional room environment. However, it is preferable to use the symmetric singular prediction polynomials in the noisy environment.

The formant frequency estimates based on the symmetric singular prediction polynomial calculation are different from those based on the antisymmetric singular prediction polynomial calculation. The formant frequency estimates based on the antisymmetric singular prediction polynomial calculation are always higher in frequency. This situation is illustrated in Fig. 3 for the Lithuanian word "įmonė" (company), to which the white noise of 60 dB was added. The reference template was calculated using the symmetric singular prediction polynomials, and the test template was calculated using the antisymmetric singular prediction polynomials. These estimates were calculated for the same speech wave. The average distance between the test and the reference templates is 107.26 Hz.

It is reasonable to think that the real formant frequencies are somewhere between the estimates based on the symmetric and antisymmetric polynomials. We have introduced a new estimate of the formant features as an average of the estimates based on the symmetric and antisymmetric polynomials. We have compared the recognition results for the formant feature extraction based on this new estimate. The results are summarized in Table 4.

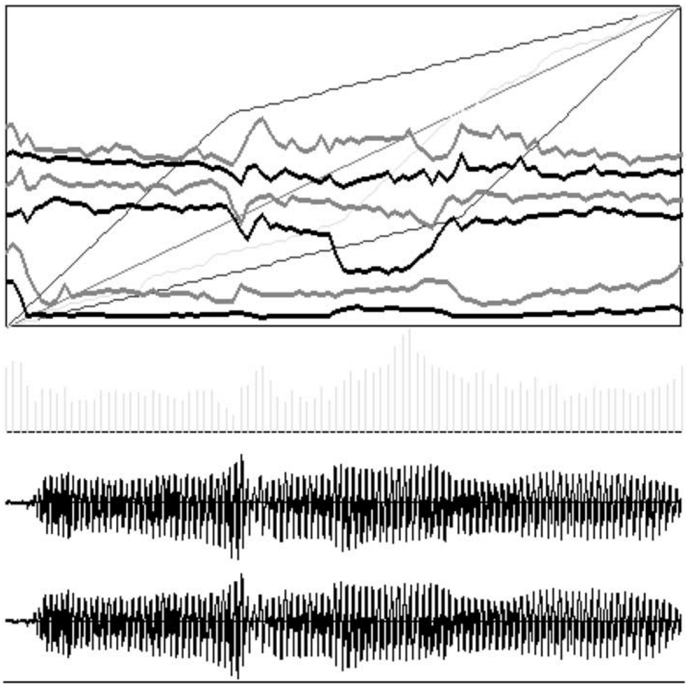

Fig. 3. Lithuanian word "įmonė", to which the white noise at 60 dB was added. The reference and the test templates were calculated using, respectively, the symmetric singular prediction polynomials (FFSSPP) and the antisymmetric singular prediction polynomials (FFASPP). Black lines are the reference templates, grey lines are the test templates.

Table 4

Recognition errors (%) of 111 words at different SNR levels using the formant frequencies calculated from the symmetric singular prediction polynomials (FFSSPP) and the formant frequencies calculated as an average of the estimates based on the symmetric and antisymmetric polynomials (FFSSPP + FFASPP)

Recognition conditions	Recognition error (%)			
	FFSSPP, 2 formats: F_1, F_2	FFSSPP, 3 formats: F_1, F_2, F_3	FFSSPP + FFASPP 2 formats: F_1, F_2	FFSSPP + FFASPP 3 formats: F_1, F_2, F_3
(SNR = 30 dB)	2.7	0	0.9	0
+65 dB white noise	11.7	11.7	13.5	14.4
+60 dB white noise	21.6	29.7	36.9	36.9
+55 dB white noise	51.3	63.9	60.3	72.9

The results of the recognition experiments show that these new estimates are comparable with the symmetric singular polynomial-based estimates at higher SNR levels and are worse than the symmetric singular polynomial based estimates at low SNR levels, especially for the two formant case. Thus one can conclude that the symmetric singular polynomial based estimates are the most suitable ones for the formant estimates in the dynamic time warping based isolated word recognition. Three formants should be used for recognition at higher SNR levels, two formants suffice at lower SNR levels.

4.3. *Recognition in Real Noisy Environments*

To test recognition performance in real noisy environments recognition experiments were performed for noisy speech recordings. The noisy recordings were selected from the same Lithuanian speech corpus of 111 Lithuanian words. Noisy speech recordings of 3 male and 3 female speakers were used in the experiments. SNR and description of noise type of the speech recordings are listed in Table 5.

In this experiment the FF-based recognition (two formants and three formants) was compared to the LPC and LPCC feature based recognition. The results of the experiment are summarized in Table 6.

Table 5

Speech recordings and corresponding SNR for male (M) and female (F) speakers

Speaker	Description of noise, SNR (dB)
M 1	Pseudo-periodic noise 9 dB
M 2	Wideband noise 23 dB
M 3	Wideband pseudo-periodic noise 12 dB
F 1	Pseudo-periodic noise 26 dB
F 2	Wideband pseudo-periodic noise 18 dB
F 3	Pseudo-periodic noise 14 dB

Table 6

Recognition errors (%) of 111 words of 3 male and 3 female speakers recorded in real noisy environments. Four different features were applied: the formant frequencies (first two or three formants) calculated from the symmetric singular prediction polynomials (FFSSPP), the LPC features (LPC), and the LPC-derived cepstral features (LPCC)

Speaker	Recognition error (%)			
	FFSSPP, 2 formants: F_1, F_2	FFSSPP, 3 formants: F_1, F_2, F_3	LPC features	LPCC features
M 1	1.8	0	0.9	0.9
M 2	2.7	2.7	2.7	3.6
M 3	9.0	7.2	5.4	7.2
F 1	0	0	0.9	0.9
F 2	4.5	2.7	2.7	0
F 3	0.9	0	0	0

From the results of this experiment we can conclude that in real noisy environments the recognition accuracy using three formants is similar to that using the LPC or LPCC features. It is preferable to use three formants for the formant feature based recognition. Also, we note that the wideband and wideband pseudo-periodic noise induces large recognition errors.

5. Conclusions and Future Work

A possibility to use the formant features in the dynamic time warping based recognition of isolated words has been investigated. Several formant feature extraction methods have been compared. A method based on the singular prediction polynomials has been proposed for the recognition of isolated words. Many speech recognition experiments have been done. In total, 111 Lithuanian words were used in the recognition experiments. The recognition performance was evaluated at various noise levels. From the experiments one can conclude:

1. The recognition performance of the formant features based on the singular symmetric prediction polynomials was compared to that of the linear prediction coding and the linear prediction coding derived cepstral features. Two and three formants were used in the formant-based speech recognition. The recognition results have shown that the performance of the formant features is similar to that of the LPC the LPCC features in the conventional room environment. It is preferable to use three formants for the recognition under those conditions. In the white noise environment, the results are different. Better results have been obtained using the formant features at all noise levels. It is preferable to use two formants for the recognition under white noise conditions.

2. The recognition results based on the symmetric singular prediction polynomial calculation to extract the formant features have been compared with that based on the calculation of the formant features from the standard LPC polynomials by the autocorrelation method. In the conventional room environment (30 dB), poor recognition results were obtained for the formant features calculated from the standard LPC polynomials estimated by means of the autocorrelation method. The difference in the recognition errors from those of the singular prediction polynomial calculation decreases as the noise level increases. However, even at the highest white noise levels the recognition using the singular prediction polynomials is more accurate.
3. The recognition results for the formant feature extraction, based on the symmetric singular prediction polynomial calculation, were compared to the results based on the antisymmetric singular prediction polynomial calculation. The recognition results have shown that the performance of both methods is similar in the conventional room environment. However, in the noisy environment, it is preferable to use the symmetric singular prediction polynomials.
4. We introduced a new estimate of the formant features as an average of the estimates based on the symmetric and antisymmetric polynomials. The recognition results have shown that these new estimates are comparable to the estimates based on the symmetric singular polynomials at higher SNR levels and are worse than the symmetric singular polynomial based estimates at low SNR levels, especially for the two formant case.
5. From the results of the experiment with speech signals recorded in real noisy environments we can conclude that the recognition accuracy using three formants is similar to that using the LPC or LPCC features. It is preferable to use three formants for the formant feature based recognition. Also, we note that the wideband and wideband pseudo-periodic noise induces large recognition errors.

From all speech-recognition experiments we can conclude that the symmetric singular polynomial based estimates are the most suitable for the formant estimates in the dynamic time warping based recognition of isolated words and three formants should be used for the recognition.

Future work should be concentrated on the robustness of the formant features to the variability of the recording conditions and the environmental noise. Also, we have some ideas how to make our system use-independent. The vocal tract normalization and the formant feature clustering techniques should help to reach this goal.

The formant features were widely used in the acoustic-phonetic speech recognition approach, however without a great success. To our mind, the dynamic time warping approach provides a new insight to the usage of the formants and other acoustic-phonetic features for the isolated word recognition.

References

Acero, A. (1999). Formant analysis and synthesis using hidden Markov models. In *Proceedings of Eurospeech*. pp. 1047–1050.

Boersma, P., and D. Weenink. Praat: doing phonetics by computer. Available at: `www.praat.org`.

Davis, S., and P. Mermelstein (1980). Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech and Signal Processing*, **ASSP-28**(4), 357–366.

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

Delsarte, P., and Y.V. Genin (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., B. Cranen, J. de Veth and L. Boves (2000). Comparing acoustic features for robust ASR in fixed and cellular network applications. In *Proceedings of ICASSP*. pp. 1415–1418.

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

Diankha, O., and T. Shimamura (2002). Noise robust formant frequency estimation based on complex autocorrelation function. In *Proceedings of IEICE International Technical Conference onCircuits/Systems, Computers and Communications*, vol. 13. pp. 1799–1802.

Garner, N., and W. Holmes (1998). On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proceedings of ICASSP*. pp. 1–4.

Grumadienė, L., and V. Žilinskienė (1997). *Frequency Dictionary of Modern Written Lithuanian*. Mokslo aidai (in Lithuanian).

Hayes, M.H. (1996). *Statistical Digital Signal Processing and Modelling*. John Wiley & Sons, INC.

Holmes, J., W. Holmes and P. Garner (1997). Using formant frequencies in speech recognition. In *Proceedings of Eurospeech*. pp. 2083–2086.

Huang, X., A. Acero and H-W. Hon (2001). *Spoken Language Processing*. Prentice Hall PTR.

Lipeika, A., J. Lipeikienė and L. Telksnys (2002). Development of isolated word speech recognition system. *Informatica*, **13**(1), 37–46.

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

Markel, J. (1972). Digital inverse filtering – a new tool for formant trajectory estimation. *IEEE Trans. on Audio and Electroacoustics*, **AU-20**(2), 129–137.

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

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

McCandless, S. (1974). An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Trans. Acoustics, Speech and Signal Processing*, **ASSP-22**(2), 135–141.

Press, W.H. *et al.* (1988). *Numerical Recipes in C*. Cambridge University Press.

Snell, R., and F. Milinazzo (1993). Formant location from LPC analysis data. *IEEE Trans. on Speech and Audio Processing*, **1**(2), 129–134.

Rabiner, L., and B.-H. Juang (1993). *Fundamentals of Speech Recognition.* Prentice Hall.

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

Watanabe, A. (2001). Formant estimation method using inverse-filter control. *IEEE Transactions on Speech and Audio Processing*, **9**(4), 317–326.

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

Welling, L., and H. Ney (1996). A model for efficient formant estimation. In *Proceedings of ICASSP*. pp. 797–800.

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

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

Yegnanarayana, B., and R. Veldhuis (1998). Extraction of vocal tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, **6**(4), 313–327.

A. Lipeika is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics, an associate professor at the Fundamental Sciences Department of Vilnius Gediminas Technical University and a professor at the Mathematics and Informatics Department of Vilnius Pedagogical University. His scientific interests include processing and recognition of random processes, detection of changes in the properties of random processes, signal processing, speech processing, speech and speaker recognition.

J. Lipeikienė is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and a professor, the head of the Informatics Department at Vilnius Pedagogical University. Her scientific interests embrace processing of random signals, including speech signals, and modern technologies in education.

Apie formantinių požymių naudojimą dinaminiu laiko skalės kraipymu grįstame atskirai sakomų žodžių atpažinime

Antanas LIPEIKA, Joana LIPEIKIENĖ

Darbe yra nagrinėjama formantinių požymių panaudojimo galimybė dinaminiu laiko skalės kraipymu grįstame atskirai sakomų žodžių atpažinime. Keletas formantinių požymių išskyrimo metodų buvo palyginta ir išsigimusiais prognozės polinomais grįstas formantinių požymių išskyrimo metodas pasiūlytas naudoti atskirai sakomų žodžių atpažinime. Pasiūlytas metodas žodžių atpažinimo tikslumo požiūriu buvo palygintas su atpažinimu naudojant tiesinės prognozės kodavimo ir iš tiesinės prognozės modelio parametrų apskaičiuotus kepstrinius požymius. Atpažinimo eksperimente buvo naudojama 111 lietuvių šnekos žodžių. Atpažinimo tikslumas buvo vertinamas esant įvairiems kalbos signalo užtriukšminimo lygiams. Eksperimentinis tyrimas parodė, kad prie visų užtriukšminimo lygių iš išsigimusių prognozės polinomų apskaičiuotais formantiniais požymiais grįstas atskirai sakomų žodžių atpažinimas yra patikimesnis už atpažinimą naudojant tiesinės prognozės kodavimo arba iš tiesinės prognozės modelio parametrų apskaičiuotus kepstrinius požymius.