Speaker Identification Using Kalman Cepstral Coefficients*

Zdeněk Švenda and Vlasta Radová

University of West Bohemia, Department of Cybernetics,
Univerzitní 22, 306 14 Plzeň, Czech Republic
{svendaz, radova}@kky.zcu.cz

Abstract. In this paper an approach to speaker identification based on an estimation of parameters of a linear speech-production model is presented. The estimation is based on the discrete Kalman estimator. It is generally supposed that the vocal tract can be modeled by a system with constant parameters over short intervals. Taking this assumption into account we can derive a special form of the discrete Kalman estimator for the model of speech production. The parameters of the vocal tract model obtained by the above mentioned Kalman estimation we used then to compute a new type of cepstral coefficients which we call Kalman cepstral coefficients (KCCs). These coefficients were used in text-independent speaker identification experiments based on discrete vector quantization. Achieved results were then compared with results obtained by the using LPC-derived cepstral coefficients (LPCCs). The experiments were performed in a closed group of 591 speakers (312 male, 279 female).

1 Introduction

There are many methods of signal processing. A set of methods assumes that a combination of glottal cords, vocal tract and radiation can be modeled by a simple system with a known single all-pole transfer function which has unknown parameters. The parameters of such model have to be then estimated. For this purpose either a very popular method of linear predictive coding [2], [3] or methods based on a Kalman estimator [1] can be used. Another set of signal processing methods analyze signal in frequency domain, for example using a filterbank. In this way energy in particular frequency band can be obtained. Mel-frequency cepstral coefficients can be derived from this energy [4].

2 Model of Speech Production

The speech waveform is an acoustic pressure wave that originates from voluntary physiological movements of anatomical structures such as the vocal cords, vocal

* The work was supported by the Ministry of Education of the Czech Republic, project no. MSM235200004, and by the Grant Agency of the Czech Republic, project no. 102/96/K087.

tract, nasal cavity, tongue and lips. Human speech production can be modeled by a linear filter where the glottal cords, vocal tract, and radiation are individually modeled as linear filters [2], [3]. The input of the total filter is either a quasi-periodic impulse sequence for voiced sounds or a random noise sequence for unvoiced sounds, with a gain factor G set to control the intensity of the excitation. The glottal cords can be modeled by second-order low-pass filter $G(z)$. The vocal tract model $V(z)$ can be described as all-pole model, where each pole of this model corresponds to a formant or resonance frequency of the sound. The radiation model $R(z)$ describes the air pressure at the lips and can be reasonably approximated by a first-order backward difference. Combining the glottal pulse, vocal tract, and radiation yields a single all-pole transfer function given by

$$H(z) = G(z)\,V(z)\,R(z) = \frac{G}{1 + \sum\limits_{i=1}^{Q} a_i z^{-i}} \ , \tag{1}$$

where a_i are unknown parameters of model. With this transfer function, we get a difference equation for synthesising the speech samples $y(k)$ as

$$y(k) = -\sum_{i=1}^{Q} a_i y(k-i) + Gu(k) \ , \tag{2}$$

where $u(k)$ is the input of the filter.

It can be noted that $y(k)$ is predicted as a linear combination of the previous Q samples. Therefore, the speech production model is often called the *linear prediction* (LP) model, or the *autoregressive model.*

3 Feature Extraction

The feature extraction can be divided into two steps. First a set of predictors coefficients has to be obtained, and then this set has to be transformed into feature vector.

In practice, the predictor coefficients $\{a_i\}$ describing the autoregressive model must be computed from the speech signal. Since speech is a time-varying signal and the vocal-tract configuration changes over time, an accurate set of predictor coefficients is adaptively determined over short intervals (10-30 ms) called frames, during that time-invariance is assumed. The gain G is usually ignored to allow the parameterizations to be independent of the signal intensity [2], [3].

3.1 Linear Predictive Coding

One of the standard methods of predictor coefficients calculation is the *autocorrelation* method. The predictor coefficients can be obtained by solving the equation

$$\boldsymbol{R}\boldsymbol{a} = \boldsymbol{r} \ , \tag{3}$$

where $\boldsymbol{R}$ is Toeplitz autocorrelation matrix, $\boldsymbol{a}$ is vector of predictor coefficients, and $\boldsymbol{r}$ is a vector of autocorrelation coefficients. It is assumed that the speech samples are identically zero outside the frame of interest. A computationally efficient algorithm known as the Levinson-Durbin recursion can be used to solve this equation [3].

3.2 The Discrete Kalman Estimator

In the state space a system can be generally described by equation system

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k y_k + \Gamma_k \xi_{k+1} \ , \\ y_{k+1} &= C_k x_k + D_k y_k + \Delta_k \xi_{k+1} \ , \end{aligned} \tag{4}$$

where x_k is an internal immeasurable part of state, y_k is an external measurable part of state, A_k, B_k, C_k and D_k are matrices describing relations among vectors x_k, y_k, x_{k+1} and y_{k+1}, the term $\Gamma_k \xi_{k+1}$ represents system errors and the term $\Delta_k \xi_{k+1}$ represents an additive noise in the observation process (ξ_{k+1} is the white noise).

In arbitrary step k the state of the system is not known exactly. The estimate of actual state is given by normal distributions

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} \sim N \left\{ \begin{bmatrix} \widehat{x}_{k+1} \\ \widehat{y}_{k+1} \end{bmatrix}, \begin{bmatrix} A_k P_k A_k^T + \Gamma_k \Gamma_k^T & A_k P_k C_k^T + \Gamma_k \Delta_k^T \\ C_k P_k A_k^T + \Delta_k \Gamma_k^T & C_k P_k C_k^T + \Delta_k \Delta_k^T \end{bmatrix} \right\} \ , \tag{5}$$

where $\widehat{x}_k$ and $\widehat{y}_k$ are the estimates of x_k and y_k in step k. On the assumption that the vector of unknown parameters x is random with known mean μ_0 and covariance P_0 and independent on the random distribution of initial state s_0, the optimal estimate of the vector of parameters μ_k can be obtained by recursive algorithm

$$\begin{aligned} K_{k+1} &= \left(A_k P_k C_k^T + \Gamma_k \Delta_k^T\right) \left(C_k P_k C_k^T + \Delta_k \Delta_k^T\right)^{-1} \ , \\ \mu_{k+1} &= A_k \mu_k + B_k y_k + K_{k+1} \left(y_{k+1} - C_k \mu_k - D_k y_k\right) \ , \\ P_{k+1} &= \left(A_k P_k A_k^T + \Gamma_k \Gamma_k^T\right) - K_{k+1} \left(C_k P_k A_k^T + \Delta_k \Gamma_k^T\right) \ , \end{aligned} \tag{6}$$

with initial conditions μ_0 and P_0. K_k represents so called Kalman gain, μ_k the best a posteriori estimate and P_k the covariance of estimate dependent on observation.

3.3 A Special Form of the Discrete Kalman Estimator for the Model of Speech Production

The linear model of speech production (2) is a system with time-invariant parameters over short interval (see above). It means the transition matrix A_k is the identity over this interval and the matrices B_k and Γ_k are zero. It can be seen from (2), that matrix D_k is zero as well. Taking this assumptions into account the linear model of speech production in state representation can be described by the equation system

$$x_{k+1} = x_k = x \ , \qquad y(k) = C(y_k)\, x + \Delta_k \xi_{k+1} \ , \tag{7}$$

where $x = [a_1, a_2, \ldots, a_Q]^T$, $y_k = [y(k-1), y(k-2), \ldots, y(k-Q)]^T$ and $C(y_k) = [-y_k]^T$. If we take into account forms of the matrices A_k, B_k, C_k, D_k and Γ_k, we can derive a special form of Kalman estimator (6) for the linear model of speech production

$$\begin{aligned} K_{k+1} &= -P_k y_k \left(y_k^T P_k y_k + \Delta_k^2\right)^{-1} \ , \\ \mu_{k+1} &= \left(I + K_{k+1} y_k^T\right) \mu_k + K_{k+1} y(k+1) \ , \\ P_{k+1} &= \left(I + K_{k+1} y_k^T\right) P_k \ , \end{aligned} \tag{8}$$

where I represents the identical matrix.

3.4 Cepstrum

In many speaker identification applications the predictor coefficients are transformed into feature vectors consisting of so called *LPC-derived cepstral coefficients* (LPCCs). A recursive relation between the LPC-derived cepstral coefficients and the predictor coefficients is given as [2], [3]

$$c(k) = -a_k - \sum_{i=1}^{k-1} \left(\frac{i}{k}\right) c(i)\, a_{k-i} \ . \tag{9}$$

We used this formula for computation of so called *Kalman cepstral coefficients* (KCCs) from the predictor coefficients produced by the Kalman estimator (described in Sect. 3.2).

4 Experiments

All experiments described in this section were performed following this initial condition: $P_0 = 5\,000\,000\, I$ (I is the identical matrix) and $\mu_0 = [0, \ldots, 0]^T$.

4.1 Reconstructed Spectrum

The predictor coefficients $\{a_i\}$ allows to compute the signal spectrum. Using the substitution $z = e^{j\omega}$ in transfer function (1) we obtain $|H(j\omega)|$ that represents the spectral envelope of the speech. In the Fig. 1 the reconstructed spectra (dashed line), the predictor coefficients $\{a_i\}$ were estimated by algorithm (8), and spectra evaluated by Fourier transformation (solid line) are compared.

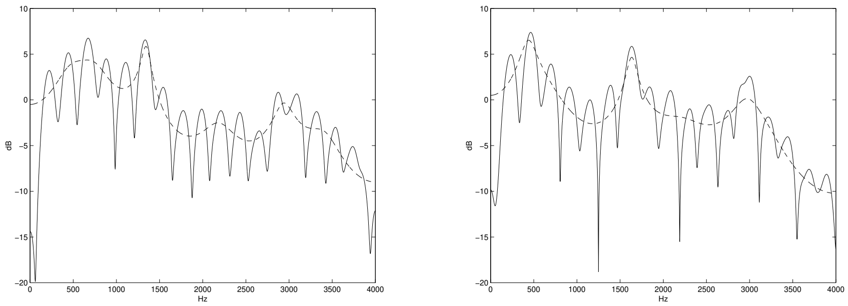

Fig. 1. Spectrum of the vowel a (left) and e (right) spoken by a female speaker

4.2 Speaker Identification

Database. The database consisted of speech signal obtained from 591 speakers (312 male, 279 female). Every speaker spoke a different set of short Czech sentences and isolated words during one session. The speech signal was transferred through a telephone channel, sampled at the 8kHz sampling rate and stored in a μ-law 8bit format. Before further processing, the 8bit μ-law digitalized samples were finally converted to linear 16bit PCM samples. About 40s of obtained speech were regarded as training data for each speaker and were used to form a reference model. The remaining speech data (on average 30s per spreaker) were used for identification tests.

Feature Extraction. The speech signal was windowed by a 16ms Hamming window (128 samples) and not pre-emphasised. For each 16ms segment a feature vector of 12th order of either LPCCs or KCCs was formed. The feature vectors were then mean normalised even thought it is not necessary because the training and testing condition were the same. Segments of silence were removed from the speech data before the feature extraction using an adaptive energy threshold speech/silence detector.

Vector Quantization. The vector quantization is an approach used for reduction of an extensive data set. The data set $X = \{x_i | i = 1, \ldots, N\}$, where N is the number of vectors in the set X, is mapped onto a finite set of M ($M \ll N$) codebook vectors $W = \{w_i | i = 1, \ldots, M\}$. Each vector is assigned to the nearest codebook vector. In the result every codebook is composed of centroid vectors (means) representing nonoverlapping regions in the feature space. Every codebook obtained in this way is included in the reference database. During an identification phase each sequence of input vectors of an unknown speaker is quantized using K codebooks corresponding to K different speakers. The unknown speaker is then identified as the reference speaker with the minimum average distortion. This method of speaker identification is described in more detail in [4].

Experimental Results. In our experiments we formed codebooks of either 80 or 320 vectors. The experimental results are summarised in Table 1.

Table 1. Identification results

Codebook size 80			Codebook size 320		
Coefficients	# correct	Correct [%]	Coefficients	# correct	Correct [%]
LPCCs	578	97.80	LPCCs	582	98.48
KCCs	576	97.46	KCCs	582	98.48

As the table shows, in the case with the codebook of 80 vectors, better results were obtained when the LPCCs were used, but the difference is almost insignificant. In the second case, using the codebook of 320 vectors, the results are absolutely identical.

5 Conclusion

Results of our experiments show that the predictor coefficients calculation based on discrete Kalman estimator can be usable in speech processing. It can be usable for the reconstruction of signal spectra (see Sect. 4.1) as well as for speaker identification application (see Sect. 4.2). The handicap of this algorithm is its very high time consuming.

References

1. Mack G. A., Jain V. K.: A Compensated-Kalman Speech Parameter Estimator. IEEE Signal Processing Magazine (1985)
2. Mammone R. J., Zhang X., Ramachandran R. P.: Robust Speaker Recognition. IEEE Signal Processing Magazine (1996)
3. Psutka J.: Communication with Computer by Speech. Academia, Prague (1995) (in Czech)
4. Radová V., Švenda Z.: An Approach to Speaker Recognition Based on Vector Quantization. First International Conference on Advanced Engineering Design, Prague (1999)