INFORMATICA, 2003, Vol. 14, No. 4, 487–496

# Word Endpoint Detection Using Dynamic Programming

Antanas LIPEIKA, Joana LIPEIKIENĖ
*Institute of Mathematics and Informatics*
*Akademijos 4, 2600 Vilnius, Lithuania*
*e-mail: lipeika@ktl.mii.lt, joanal@ktl.mii.lt*

Received: September 2003

**Abstract.** The paper deals with the use of dynamic programming for word endpoint detection in isolated word recognition. Endpoint detection is based on likelihood maximization. Expectation maximization approach is used to deal with the problem of unknown parameters. Speech signal and background noise energy is used as features for making decision. Performance of the proposed approach was evaluated using isolated Lithuanian words speech corpus.

**Key words:** endpoint detection, change-point detection, dynamic programming, likelihood maximization, expectation maximization.

## 1. Introduction

Speech recognition performance highly depends on endpoint detection accuracy. In speech recognition based on dynamic time warping it is assumed that word boundaries are known precisely (Rabiner *et al.*, 1993), and incorrect endpoint detection often is a reason of recognition errors. Similar situation occurs in speech recognition based on Hidden Markov Models.

Most speech recognition systems usually use frame level signal energy or energy contour as features for endpoint detection (Lamere *et al.*, 2003; Young *et al.*, 2002). Background noise energy level is evaluated at the beginning and the end of speech signal and energy thresholds are applied to detect speech beginning and end points. Simple heuristic logic is used to make a decision. Up to three thresholds are used to make a decision (Young *et al.*, 2002). Such approach works well enough when background noise is stationary. When background noise is nonstationary due to nonspeech events introduced by the speaker, recording environment, transmission system, a number of endpoint detection errors occurs. Several heuristic improved techniques for the detection of the endpoints were studied (Lamel *et al.*, 1981; Wilpon *et al.*, 1984). These techniques were broadly classified as either explicit, implicit, or hybrid in concept. The explicit techniques locate endpoints prior to and independent of the recognition and decision stages of the speech recognition system. For the implicit methods, the endpoints are determined solely by the recognition and decision stages of the system, i.e., there is no separate stage for endpoint detection. The hybrid techniques incorporate aspects from both the explicit and implicit methods. Investigations showed that the hybrid techniques consistently provided the best estimates for both of the word endpoints and, correspondingly, the highest recognition accuracy of the three classes studied.

Other kinds of features also can be used for endpoint detection. Zhu *et al.* (1999) reported about the use of autocorrelation function, Navarro-Mesa *et al.* (1999) proposed to use third-order spectra for endpoint detection. In both cases threshold is used to make a decision.

Junqua *et al.* (1994) proposed the time-frequency (TF) parameter. They used the frequency energy in the fixed frequency band $250 \sim 3500$ Hz to enhance the time energy information. Based on the TF parameter, an algorithm was proposed to get more precise word boundary in noisy environment. This algorithm includes noise classification, a refinement procedure, and some preset thresholds. Although this algorithm outperforms several commonly used algorithms for word boundary detection in the presence of noise, it needs to determine thresholds empirically and ambiguous rules which are not easily determined by human (Wu *et al.*, 2000). Qi *et al.* (1993) used multilayer neural networks to classify the speech signal into voiced, unvoiced and silence segments.

Wu *et al.* (2000) modified TF parameter approach by proposing adaptive time-frequency (ATF) parameter for extracting both time and frequency features of noisy speech signals. The ATF parameter extends the TF parameter from single band to multiband spectrum analysis, where the frequency bands help to make the distinction of speech and noise signals clear. The ATF parameter can extract useful frequency information by adaptively choosing proper bands of the mel-scale frequency bank. New word boundary detection algorithm by using a neural fuzzy network (SONFIN) for identifying islands of word signals in noisy environment was proposed (Wu *et al.*, 2000). Due to the self-learning ability of SONFIN, the proposed algorithm avoids the need of empirically determined thresholds. This approach reduced the recognition error rate due to endpoint detection to about 10% as compared with TF-based algorithm.

According to Li *et al.* (2002), change-point detection methods also can be applied to word endpoint detection. Our endpoint detection investigation is also based on change-point detection (Lipeika, 2000). It is assumed that there are two change-points in the signal, the beginning and the end of the spoken word. Change-point detection is based on likelihood maximization. The likelihood maximization is based on dynamic programming. Expectation maximization approach is used to deal with the problem of unknown parameters. In formal solution of the problem it is assumed that speech signal and background noise are described by linear prediction (LP) model. Since model parameters are unknown, we used fixed length segments at the beginning and the end of the signal to estimate initial background noise parameters and all remaining signal to estimate initial spoken word parameters. Further expectation maximization procedure is applied. We estimate endpoints using initial estimates of spoken word and background noise parameters, reestimate spoken word and background noise parameters according to obtained endpoint estimates etc. Calculations continues until endpoint estimates does not change.

Our experiments showed that LP model does not suit well for endpoint detection. We discovered that more preferably is to apply our approach to signal energy. In further investigations we adopted our approach to frame level signal energy to match frame by frame speech preprocessing.

Main advantage of our approach is that it avoids the need of thresholds and heuristic decision rules. Performance is higher as compared with threshold based method (Lipeika *et al.*, 2002).

## 2. Statement of the Problem for Known Model Parameters

Let us consider random sequence $x = \{x(1), x(2), ..., x(N)\}$, which is an output of linear discrete time system with time-varying parameters. $v(n)$ is input of the system. The system structure satisfies linear prediction (LP) equation of the form

$$x(n) = -a_1(n)x(n-1) - a_2(n)x(n-2) - \cdots - a_p(n)x(n-p) + b(n)v(n), \quad (1)$$

where $p$ is an order of autoregressive system; $A'(n) = [a_1(n), a_2(n), \dots, a_p(n), b(n)]$ is a vector of time-varying parameters of the system, at every time instant satisfying system stability conditions. At this point parameters $A(n)$ of the system are known a priori and are changing according to the rule

$$A(n) = \begin{cases} A_1, & n = \dots, 1, 2, \dots, u_1, \\ A_2, & n = u_1 + 1, \dots, u_2, \\ A_3, & n = u_2 + 1, \dots, N, \end{cases} \quad (2)$$

where $u = [u_1, u_2]$ are unknown change points, satisfying the condition $p < u_1 < u_2 < N$ and our problem is to find their maximum likelihood estimates $\hat{u} = [\hat{u}_1, \hat{u}_2]$.

## 3. Maximization of the Likelihood Function

Estimates of change-points can be found by maximizing the logarithm of the likelihood function

$$\hat{u} = \arg\max_u \log p(x|u), \quad (3)$$

where

$$\begin{aligned} \log p(x|u) &= \log p(x(1), x(2), \dots, x(p)) - (N-p)/2 \log(2\pi) \\ &\quad -(u_1 - p)\log b(1) - (u_2 - u_1)\log b(2) - (N - u_2)\log b(3) \\ &\quad -\frac{1}{2b^2(1)} \sum_{n=p+1}^{u_1} \left[\sum_{j=0}^{p} a_j(1)x(n-j)\right]^2 - \frac{1}{2b^2(2)} \sum_{n=u_1+1}^{u_2} \left[\sum_{j=0}^{p} a_j(2)x(n-j)\right]^2 \\ &\quad -\frac{1}{2b^2(3)} \sum_{n=u_2+1}^{N} \left[\sum_{j=0}^{p} a_j(3)x(n-j)\right]^2. \end{aligned} \quad (4)$$

Straightforward maximization of (4) is problematic due to a large computation score. Instead of maximizing (4) we can maximize objective function $\Theta(u|x)$, which differs from (4) by an additive constant not depending on $u$, i.e.,

$$\Theta(u|x) = L_1(u_1|x) + L_2(u_2|x), \tag{5}$$

where

$$\begin{aligned} L_i(k|x) \;=\; & -(k-p)\log b(i) - (N-k)\log b(i+1) \\ & -\frac{1}{2b^2(i)} \sum_{n=p+1}^{k} \left[\sum_{j=0}^{p} a_j(i)x(n-j)\right]^2 \\ & -\frac{1}{2b^2(i+1)} \sum_{n=k+1}^{N} \left[\sum_{j=0}^{p} a_j(i+1)x(n-j)\right]^2, \\ & i = 1,2; \quad k = p+1, 2, \ldots, N. \end{aligned} \tag{6}$$

Functions $L_i(k|x),\ i = 1,2$ are partial likelihood functions and each of the functions depends only on one change point.

To reduce computation amount expression (6) can be calculated recursively

$$\begin{aligned} & L_i(k|x) = L_i(k-1|x) - \log b(i) + \log b(i+1) \\ & -\frac{1}{2b^2(i)} \left[\sum_{j=0}^{p} a_j(i)x(n-j)\right]^2 + \frac{1}{2b^2(i+1)} \left[\sum_{j=0}^{p} a_j(i+1)x(n-j)\right]^2, \\ & i = 1,2; \quad k = 2, \ldots, N, \end{aligned} \tag{7}$$

Since initial conditions does not depend on change points, we can let initial conditions to be equal to zero.

Since the function $\Theta(u|x)$ consists of the sum of partial functions $L_i(u_i|x), i = 1,2$, and each of these partial functions depends only on one variable, we can use the dynamic programming method to determine place of the global maximum (Cooper *et al.*, 1981) of this function. Let us define the Bellman functions

$$g_1(u_2|x) = \max_{\substack{u_1 \\ p<u_1<u_2}} L_1(u_1|x), \quad u_2 = p+2, \ldots, N, \tag{8}$$

$$g_2(u_3|x) = \max_{\substack{u_2 \\ p+1<u_2<u_3}} \left[L_2(u_2|x) + g_1(u_2|x)\right], \quad u_3 = p+3, \ldots, N. \tag{9}$$

Estimates of the change points can be found from the Bellman functions

$$\hat{u}_k = \min\left[\arg \max_{\substack{n \\ p+k \leqslant n \leqslant \hat{u}_{k+1}}} g_k(n|x)\right], \quad k = 2,1, \tag{10}$$

where, for convenience, we made a notation $\hat{u}_3 = N$.

Up to this point we assumed that LP parameters were known. We need to adopt this approach to unknown parameters.

## 4. Maximization of the Likelihood Function when the Model Parameters are Unknown

For the maximization of the likelihood function when parameters are unknown we used generalized expectation maximization (GEM) approach. We used fixed length segments at the beginning and the end of the signal to estimate initial background noise parameters and all remaining signal to estimate initial spoken word parameters. Then using equations (7)–(10) and initial values of unknown parameters we estimated change-points (endpoints) and applied them to get improved estimates of the parameters. Calculations continued iteratively until change-point estimates does not change.

For tentative evaluation of our approach we performed the following experiment. 50 different Lithuanian words were pronounced four times. Isolated word recognition based on dynamic time warping (Lipeika *et al.* 2002) was used for evaluation. The first pronouncing was used to create reference templates. Recognition error rate was used for endpoint detection performance evaluation.

Using energy threshold based approach for endpoint detection 12% of recognition error were obtained. Using our approach for 0th order LP model (only energy features) error rate reduced to 4%. When using higher order LP model error rate increased dramatically, e.g., for 4th order LP model error rate increased up to 26%. It means that LP model does not provide relevant information for endpoint detection, and for further development we will use only energy features.

## 5. Endpoint Detection for Frame Level Speech Processing

Up to this point we used sample by sample based speech processing and endpoint detection was applied to this kind of calculations. This approach was used only for development of the endpoint detection algorithm. In speech recognition systems usually frame based calculations are used. In this section we apply our approach for frame based speech processing using energy features.

Let $x = \{x(1), x(2), \ldots, x(N)\}$ is frame based, independent and normally distributed signal energy samples. So (2) expression can be written as follows:

$$A(n) = \begin{cases} A_1 = N(\mu_1, \sigma_1^2), & n = \ldots, 1, 2, \ldots, u_1, \\ A_2 = N(\mu_2, \sigma_2^2), & n = u_1 + 1, \ldots, u_2, \\ A_3 = N(\mu_3, \sigma_3^2), & n = u_2 + 1, \ldots, N, \end{cases} \tag{11}$$

where $\mu_i, \sigma_i^2, i = 1, 2, 3$ is mean and variance of the signal energy in frames. $A_1$ and $A_3$ are background noise parameters, $A_2$ is spoken word parameter vector.

Using this notations the likelihood function can be written as

$$\begin{aligned}\log p(x|u) &= \log \prod_{n=1}^{N} p(x(n)) = -N/2\log(2\pi) - u_1 \log\sigma_1 \\ &\quad -(u_2-u_1)\log\sigma_2 - (N-u_2)\log\sigma_3 \\ &\quad -\frac{1}{2\sigma_1^2}\sum_{n=1}^{u_1}[x(n)-\mu_1]^2 - \frac{1}{2\sigma_2^2}\sum_{n=u_1+1}^{u_2}[x(n)-\mu_2]^2 \\ &\quad -\frac{1}{2\sigma_3^2}\sum_{n=u_2+1}^{N}[x(n)-\mu_3]^2. \end{aligned} \tag{12}$$

Partial likelihood functions are

$$\begin{aligned} L_i(k|x) &= -k\log\sigma_i - (N-k)\log\sigma_{i+1} \\ &\quad -\frac{1}{2\sigma_i^2}\sum_{n=1}^{k}[x(n)-\mu_i]^2 - \frac{1}{2\sigma_{i+1}^2}\sum_{n=k+1}^{N}[x(n)-\mu_{i+1}]^2, \\ &\quad i=1,2;\quad k=1,2,\ldots,N, \end{aligned} \tag{13}$$

which can be calculated recursively

$$\begin{aligned} L_i(k|x) &= L_i(k-1|x) - \log\sigma_i + \log\sigma_{i+1} \\ &\quad -\frac{1}{2\sigma_i^2}[x(n)-\mu_i]^2 + \frac{1}{2\sigma_{i+1}^2}[x(n)-\mu_{i+1}]^2, \\ &\quad i=1,2;\quad k=2,\ldots,N, \end{aligned} \tag{14}$$

using initial conditions as in (7).

Other calculations are the same as in sample by sample based endpoint detection. We simply use (8), (9) for calculation of the Bellman functions and (10) for the change-point estimation. Also, expectation maximization procedure is used to solve the problem of unknown parameters, as described in the previous section.

## 6. Experiments and Results

To compare our proposed dynamic programming based endpoint detection approach with threshold based approach we performed the following experiment. 50 different Lithuanian words were pronounced by one female and two male speakers, each word by each speaker was pronounced eleven times. Isolated word speech recognition system (Tamulevičius *et al.*, 2003) based on dynamic time warping, in which dynamic programming and threshold based methods were implemented, was used for experiments.

The first session of every speaker was used for training, other ten sessions (500 words) – for recognition. Results of the experiment are summarized in Table 1.

Table 1

Results of endpoint detection experiments. DP-dynamic programming, TH-threshold

| | Recognition error % | | | | | |
|---|---|---|---|---|---|---|
| Speaker | F1 (female) | | M1 (male) | | M2 (male) | |
| Method | DP | TH | DP | TH | DP | TH |
| I session | 10 | 18 | 2 | 6 | 26 | 56 |
| II session | 10 | 12 | 2 | 8 | 24 | 56 |
| III session | 6 | 4 | 4 | 6 | 18 | 52 |
| IV session | 8 | 24 | 2 | 4 | 34 | 56 |
| V session | 8 | 16 | 10 | 6 | 18 | 56 |
| VI session | 6 | 12 | 16 | 14 | 28 | 50 |
| VII session | 6 | 14 | 14 | 16 | 22 | 60 |
| VIII session | 10 | 8 | 12 | 10 | 22 | 60 |
| IX session | 14 | 14 | 12 | 12 | 26 | 60 |
| X session | 12 | 4 | 12 | 8 | 28 | 58 |
| Average error | 9 | 12.6 | 8.6 | 9 | 24.6 | 56.4 |

Average recognition error rate for different endpoint detection methods is presented in Fig. 1.

F1 and M1 are high quality and low noise level speech records. M2 has high noise level and a lot of transient impulses. This is the main reason of high recognition error rate for speaker M2. We can resume that dynamic programming algorithm results in lower recognition error rate and outperforms threshold-based approach. In addition the difference in error rate is higher for noisy speech records.

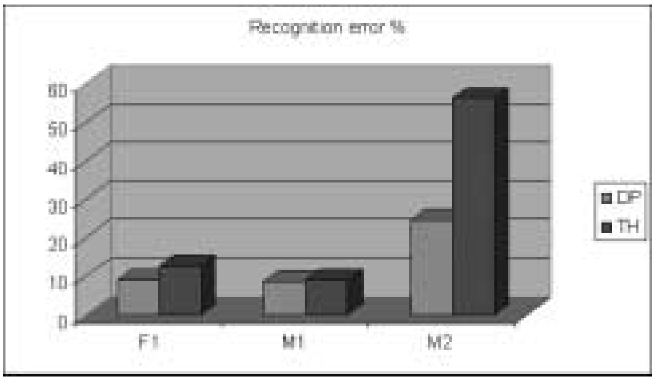

Fig. 1. Average recognition error rate for different endpoint detection methods. Speakers F1, M1 and M2. DP-dynamic programming method, TH-threshold – based method.

## 7. Conclusions

The use of dynamic programming for endpoint detection in isolated word recognition was investigated. In the first investigation stage linear prediction was used for spectral modeling of speech and background noise but our endpoint detection experiments showed that the linear prediction model does not suit well for endpoint detection.

Further frame level endpoint detection algorithm based on dynamic programming for likelihood maximization which used speech signal and background noise frame energy features was developed. Performance evaluation based on recognition error rate showed that:

- for low noise level endpoint detection based on dynamic programming slightly outperforms threshold based endpoint detection
- for high background noise level preference of dynamic programming based algorithm is more noticeable.

The main advantage of dynamic programming based approach is that this method does not need any threshold. Use of the thresholds always raises many problems (Junqua *et al.*, 1994) and reliable threshold determination for endpoint detection under noisy conditions remains an unsolved problem.

## References

Cooper, L., M. Cooper (1981). *Introduction to Dynamic Programming*. Pergamon Press.

Duda, R.O., P.E. Hart, D.G. Stork (2001). *Pattern Classification*. John Wiley & Sons.

Junqua, J.C., B. Mak, B. Reaves (1994). A robust algorithm for word boundary detection in the presence of noise. *IEEE Trans. on Speech and Audio Processing*, **2**, 406–412.

Lamel, L., L. Rabiner, A. Rosenberg, J. Wilpon (1981). An improved endpoint detector for isolated word recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, **ASSP-29**(4), 777–785.

Lamere, P., P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, P. Wolf (2003). The CMU SPHINX-4 speech recognition system. In *Proc. of the ICASSP'2003*.

Li, Q., J. Zheng, A. Tsai, Q. Zhou (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Trans. on Speech and Audio Processing*, **10**(3), 146–156.

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

Lipeika, A. (2000). Optimal segmentation of random sequences. *Informatica*, **11**(3), 243–256.

Navarro-Mesa, J., A. Moreno-Bilbao, E. Lleida-Solano (1999). An improved speech endpoint detection system in noisy environments by means of third-order spectra. *IEEE Signal Processing Letters*, **6**(9), 224–226.

Qi, Y., B.R. Hunt (1993). Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier. *IEEE Trans. on Speech and Audio Processing*, **1**, 250–255.

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

Tamulevičius, G., A. Lipeika (2003). Isolated word recognition system development. *Lithuanian Mathematical Journal*, **43** (submitted).

Wilpon, J.G., L.R. Rabiner, T. Martin (1984). An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints. *AT&T Bell Laboratories Technical Journal*, **63**(3), 479–498.

Wu, G.-D, Ch.-T. Lin (2000). Word boundary detection with male-scale frequency bank in noisy environment. *IEEE Trans. on Speech and Audio Processing*, **8**(5), 541–554.

Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland (2002). *The HTK Book*. Cambridge University Engineering Department.

Zhu, J., F.-L. Chen (1999). The analysis and application of a new endpoint detection method based on distance of autocorrelated similarity. In *Proc. of the EUROSPEECH'99*. pp. 105–108.

**A. Lipeika** is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and a professor at the Radioelectronics and Fundamental Sciences Departments of Vilnius Technical University and at the Mathematics and Informatics Department of Vilnius Pedagogical University. 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 at the Mathematics and Informatics Department of Vilnius Pedagogical University. Scientific interests embraces processing of random signals, including speech signals, methods for determination of change-points in the properties of random processes, modern technologies in education.

# Žodžio galų nustatymas naudojant dinaminį programavimą

Antanas LIPEIKA, Joana LIPEIKIENĖ

Straipsnyje nagrinėjamas dinaminio programavimo naudojimas žodžio galų nustatymui atskirai sakomų žodžių atpažinime. Žodžio galų nustatymas remiasi tikėtinumo funkcijos maksimizavimu. Nežinomų parametrų problemos sprendimui yra taikomas matematinės vilties maksimizavimo principas. Sprendimų priėmimas remiasi kalbos signalo ir foninio triukšmo energija. Pasiūlyto metodo darbingumo tyrimas yra atliekamas naudojant atskirai sakomų lietuvių kalbos žodžių duomenų bazę.