ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 *2006. Nr. 1(65)*
ELEKTRONIKA IR ELEKTROTECHNIKA

SIGNAL TECHNOLOGY
T 121
SIGNALŲ TECHNOLOGIJA

Segmentation of Words into Phones

A. Lipeika, G. Tamulevičius
Institute of Mathematics and Informatics, A. Goštauto str. 12, LT-01108 Vilnius, Lithuania; e-mail: lipeika@ktl.mii.lt

1. Introduction

Over the past years several speech segmentation methods have been proposed [1]. There are a few approaches to solving this problem. One of them utilizes explicit information, known in advance and a speech signal is then segmented using reference templates corresponding to the phonetic events. Based on this approach in [2] a procedure for automatic alignment of phonetic transcriptions with continuous speech is described. The procedure uses a standard pattern classification algorithm, a dynamic programming algorithm and the constraints imposed by acoustic-phonetic knowledge. A similar procedure is used in [3], where the dynamic time warping method, developed for isolated word recognition, is extended to the problem of time aligning sentence length utterances. In [4] waveform segmentation and labeling into broad phonetic classes prior to time alignment is used.

Another approach does not require any explicit information, but utilizes only the acoustical information that is contained within the speech signal to be segmented. This information can be e.g., amount of spectral change from one speech frame to the next. In [5] a method for decomposing the speech signals into acoustic units serially is described. In [6] the frames belonging to a segment are represented by an average spectrum. This method minimizes the spectral deviations within a segment from the corresponding average spectrum but does not require that the segments be the same as a phonetician might recognize. In [1] also is assumed that the phonetic transcription of the utterance is known to the segmentation algorithm. The speech signal information known to the segmentation algorithm is a sequence of short-time spectral information derived using a linear prediction coding (LPC) front-end analysis.

Statistical approaches for the segmentation of the speech signals are also discussed in the papers. In [7] the segmentation is performed on a sample-by-sample basis, rather than on a block-by-block one. This provides a more accurate location of the boundaries of the acoustic segments. Statistically based methods are used to detect non-stationarities in the speech signal and the nature of the segmented units is not defined in advance. Three methods: the generalized likelihood ratio test [8], the divergence test [9] and the original pulse method (a modified divergence test) are investigated in this paper. The speech signal is assumed to be described by a string of homogeneous units, each of which is characterized by the autoregressive statistical model.

Speech signal can be regarded as pseudo-stationary signal in which pseudo-stationary parts correspond to phones. Consequently, optimal segmentation methods [10] can be applied to speech signal segmentation. Segmentation of a non-stationary process consists in assuming piecewise stationarity and in detecting the instants of change. Usually is considered that all the data are available in the same time and a global segmentation is performed instead of a sequential procedure. We use a similar approach [11] in our investigations. Maximum likelihood (ML) and least mean squared error (LMSE) methods are used for determination change points of piecewise pseudo-stationary speech signal. It is assumed that change points correspond to the boundaries of the phones contained in speech signal and number of phones is known in advance. Maximization of the likelihood function and minimization of prediction error in least mean squared error approach is based on dynamic programming [12]. Expectation maximization approach is used to deal with the problem of unknown phone parameters. It is assumed that speech signal and background noise are described by linear prediction coding (LPC) model.

2. Statement of the problem for known LPC model parameters

We formulate the problem similar to one in [11]. Let us consider random sequence $x = \{ x(1), x(2), \ldots, x(N) \}$, which is an output of linear discrete time system with time-varying parameters. $v(n)$ is input of the system. The system structure satisfies LPC equation of the form

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

where p is an order of autoregressive system; $A'(n) = [a_1(n), a_2(n), \ldots, a_p(n), b(n)]$ 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

$$A(n)=\begin{cases} A_1, & n=\ldots,1,2,\ldots,u_1,\\ A_2, & n=u_1+1,\ldots,u_2\ ,\\ \ldots & \\ A_i, & n=u_{i-1}+1,\ldots,u_i\ ,\\ \ldots & \\ A_M, & n=u_{M-1}+1,\ldots,u_M,\\ A_{M+1}, & n=u_M+1,\ldots,N,\ldots\ , \end{cases} \quad (2)$$

where $\mathbf{u}=[u_1,u_2,\ldots,u_M]$ are unknown change points, satisfying the condition $p<u_1<u_2\ldots<u_M<N$ and our problem is to find their maximum likelihood estimates $\hat{\mathbf{u}}=[\hat{u}_1,\hat{u}_2,\ldots,\hat{u}_M]$, when the realization $x=\{x(1),x(2),\ldots,x(N)\}$ of the random sequence is available.

3. Maximization of the likelihood function

The maximum likelihood estimator $\hat{\mathbf{u}}$ of change points $\mathbf{u}$ is

$$\hat{\mathbf{u}}=\arg\max_{\mathbf{u}} p(\mathbf{u}\,|\,x)\,. \quad (3)$$

The likelihood function can be expressed as (for details see [11])

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

where we assume $a_0(i)=1,\quad i=1,2,\ldots,M+1$.

Instead of using likelihood function in implementation it is more convenient to use logarithmic likelihood function $\log p(x\,|\,\mathbf{u})$. It helps us to avoid exceeding dynamic range of the computer and reduces computation score. Taking $\log p(x\,|\,\mathbf{u})$ we can write (3) and (4) as

$$\hat{\mathbf{u}}=\arg\max_{\mathbf{u}} p(x\,|\,\mathbf{u})=\arg\max_{\mathbf{u}}\log p(x\,|\,\mathbf{u}) \quad (5)$$

And

$$\begin{aligned} \log p(x\,|\,\mathbf{u}) &= \log p(x(1),x(2),\ldots,x(p))-\\ &-(N-p)/2\log(2\pi)-(u_1-p)\log b(1)-\\ &-(u_2-u_1)\log b(2)-\ldots-(N-u_M)\log b(M+1)-\\ &-\frac{1}{2b^2(1)}\sum_{n=p+1}^{u_1}\left[\sum_{j=0}^{p}a_j(1)x(n-j)\right]^2-\\ &-\frac{1}{2b^2(2)}\sum_{n=u_1+1}^{u_2}\left[\sum_{j=0}^{p}a_j(2)x(n-j)\right]^2-\\ &-\ldots-\frac{1}{2b^2(M+1)}\sum_{n=u_M+1}^{N}\left[\sum_{j=0}^{p}a_j(M+1)x(n-j)\right]^2. \quad (6) \end{aligned}$$

Maximization of (6) is impossible due to very large amount of computations. For the given realization of the random sequence instead of maximizing (6) we can maximize objective function $\Theta(\mathbf{u}\,|\,x)$, which differs from (6) by an additive constant not depending on $\mathbf{u}$, i.e.,

$$\hat{\mathbf{u}}=\arg\max_{\mathbf{u}}\log p(x\,|\,\mathbf{u})=\arg\max_{\mathbf{u}}\Theta(\mathbf{u}\,|\,x)\,, \quad (7)$$

where

$$\Theta(\mathbf{u}\,|\,x)=L_1(u_1\,|\,x)+L_2(u_2\,|\,x)+\ldots+L_M(u_M\,|\,x)\,. \quad (8)$$

Each of functions $L_i(u_i\,|\,x),\quad i=1,2,\ldots,M$ depends only on one unknown change point u_i and can be expressed as

$$\begin{aligned} L_i(k\,|\,x) &= -(k-p)\log b(i)-(N-k)\log b(i+1)-\\ &-\frac{1}{2b^2(i)}\sum_{n=p+1}^{k}\left[\sum_{j=0}^{p}a_j(i)x(n-j)\right]^2-\\ &-\frac{1}{2b^2(i+1)}\sum_{n=k+1}^{N}\left[\sum_{j=0}^{p}a_j(i+1)x(n-j)\right]^2,\\ &i=1,2,\ldots,M;\quad k=p+1,2,\ldots,N \end{aligned} \quad (9)$$

or can be calculated recursively

$$\begin{aligned} L_i(k\,|\,x) &= L_i(k-1\,|\,x)-\log b(i)+\log b(i+1)-\\ &-\frac{1}{2b^2(i)}\left[\sum_{j=0}^{p}a_j(i)x(k-j)\right]^2+\\ &+\frac{1}{2b^2(i+1)}\left[\sum_{j=0}^{p}a_j(i+1)x(k-j)\right]^2,\\ &i=1,2,\ldots,M;\quad k=p+1,2,\ldots,N, \end{aligned} \quad (10)$$

with the initial conditions $L_i(p\,|\,x)=0,\quad i=1,2,\ldots,M$.

Since the function $\Theta(u\,|\,x)$ consists of the sum of partial functions $L_i(u_i\,|\,x),\quad i=1,2,\ldots,M$ 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 of this function.

According to the dynamic programming method [12] let us define the Bellman functions

$$g_1(u_2 \mid x) = \max_{\substack{u_1 \\ p<u_1<u_2}} L_1(u_2 \mid x), u_2 = p+2,...,N \quad (11)$$

$$g_2(u_3 \mid x) = \max_{\substack{u_2 \\ p+1<u_2<u_3}} [L_2(u_2 \mid x) + g_1(u_2 \mid x)], \quad (12)$$

$$u_3 = p+3, \ldots, N.$$

and for $i = 3, \ldots, M$ we have

$$g_i(u_{i+1} \mid x) = \max_{\substack{u_i \\ p+i-1<u_i<u_{i+1}}} [L_i(u_i \mid x) + g_{i-1}(u_i \mid x)], \quad (13)$$

$$u_{i+1} = p+i+1, \ldots, N.$$

For further reduction of computation amount, we can compute the functions $g_i(u_{i+1} \mid x), \quad i = 1, \ldots, M$ recursively

$$g_1(u_2 \mid x) = \max [g_1(u_2 - 1 \mid x), L_1(u_2 - 1 \mid x)], \quad (14)$$

$$u_2 = p+3, \ldots, N$$

with the initial condition $g_1(p+2 \mid x) = L_1(p+1 \mid x)$.
And for $i = 2, \ldots, M$

$$g_i(u_{i+1} \mid x) = \max \{g_i(u_{i+1} - 1 \mid x), [g_{i-1}(u_{i+1} - 1 \mid x) + L_i(u_{i+1} - 1 \mid x)]\}, \quad (15)$$

$$u_{i+1} = p+i+2, \ldots, N$$

with the initial conditions

$$g_i(p+i+1 \mid x) = L_i(p+i \mid x) + g_{i-1}(p+i \mid x), \quad (16)$$

$$i = 2, \ldots, M.$$

Now the maximum likelihood estimator $\hat{\mathbf{u}} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_M]$ of the change points $\mathbf{u}$ is obtained in the following way

$$\hat{u}_k = \min [\arg \max_{\substack{n \\ p+k \le n \le \hat{u}_{k+1}}} g_k(n \mid x)], \quad (17)$$

$$k = M, M-1, \ldots, 2, 1,$$

where, for convenience, we made a notation $\hat{u}_{M+1} = N$.

Finally we obtained maximum likelihood estimator of change points $\hat{\mathbf{u}} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_M]$.

4. Maximization of the likelihood function when the LPC model parameters are unknown

For the maximization of the likelihood function when parameters are unknown we used generalized expectation maximization (GEM) approach [13]. It is assumed that number of change points M (phone boundaries) is known in advance. Also is assumed that we have some initial information about unknown LPC parameters $A_1, ..., A_{M+1}$. Then using equations (10), (14) – (17) and initial values of unknown parameters we estimated change-points (endpoints) and applied them to get improved estimates of the parameters. Calculations continue iteratively until change-point estimates do not change and these final estimates are regarded as phone boundaries in speech signal.

5. Minimization of prediction error in least mean squared error approach

Instead of squared prediction error minimization we will maximize its negative value. Under assumption (2) negative least mean squared error can be defined as

$$E(x \mid \mathbf{u}) = -\sum_{n=p+1}^{u_1} \left[\sum_{j=0}^{p} a_j(1)x(n-j)\right]^2 - \\ -\sum_{n=u_1+1}^{u_2} \left[\sum_{j=0}^{p} a_j(2)x(n-j)\right]^2 - \ldots - \\ -\sum_{n=u_M+1}^{N} \left[\sum_{j=0}^{p} a_j(M+1)x(n-j)\right]^2. \quad (18)$$

We can rewrite (18) in the form

$$E(x \mid \mathbf{u}) = S_1(u_1 \mid x) + S_2(u_2 \mid x) + \ldots + S_M(u_M \mid x) + D, \quad (19)$$

where D is a constant, not depending on $\mathbf{u}$ and the partial functions $S_i(k \mid x), i = 1,2,...,M$ can be calculated recursively

$$S_i(k \mid x) = S_i(k-1 \mid x) - \left[\sum_{j=0}^{p} a_j(i)x(k-j)\right]^2 + \\ + \left[\sum_{j=0}^{p} a_j(i+1)x(k-j)\right]^2, \quad (20)$$

$$i = 1,2, \ldots, M; \quad k = p+1,2, \ldots, N$$

with the initial conditions $S_i(p \mid x) = 0, \quad i = 1,2, \ldots, M$.

The Bellman functions (14), (15) in this case are

$$g_1(u_2 \mid x) = \max [g_1(u_2 - 1 \mid x), S_1(u_2 - 1 \mid x)], \quad (21)$$

$$u_2 = p+3, \ldots, N$$

with the initial condition $g_1(p+2 \mid x) = L_1(p+1 \mid x)$. And for $i = 2, \ldots, M$

$$g_i(u_{i+1} \mid x) = \max \{g_i(u_{i+1} - 1 \mid x), [g_{i-1}(u_{i+1} - 1 \mid x) + S_i(u_{i+1} - 1 \mid x)]\}, \quad (22)$$

$$u_{i+1} = p+i+2, \ldots, N$$

with the initial conditions

$$g_i(p+i+1 \mid x) = L_i(p+i \mid x) + g_{i-1}(p+i \mid x),$$

$$i = 2, \ldots, M.$$

Now the least mean square estimator $\hat{\mathbf{u}} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_M]$ of the change points $\mathbf{u}$ is obtained using (17).

6. Illustrative examples

The word segmentation software was developed. Speech patterns are segmented using above formulated maximum likelihood and least mean squared error approaches. The first step after speech input is selection of frames for initial estimation of parameters values. These frames have been selected by user and marked (Fig. 1).

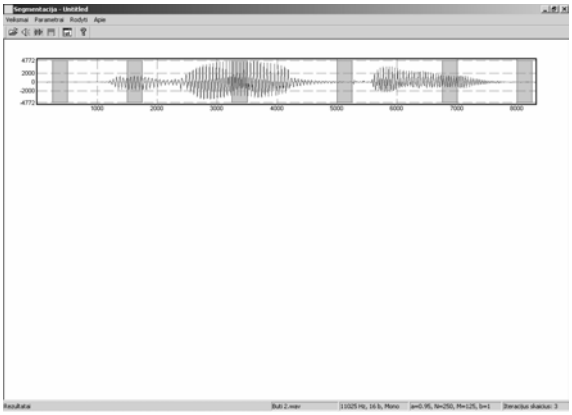

Fig. 1. Diagram of word "būti" with marked selected frames.

After selection of the frames calculations are started. Iterative calculations are repeated until change points stop changing. Segmentation results are given in Fig. 2. Here we can see detected phones boundaries (vertical lines), partial likelihood (gray curves under utterance diagram) and Bellman functions (rising black curves). Control of segmentation parameters is implemented. User can switch between maximum likelihood and least mean squared error segmentation criterion, change preemphasis ratio, analysis frame length and shift values, maximum allowed number of iterations. Besides there is possibility to control very short segment length (segment can be appended to the next or left with fixed boundaries).

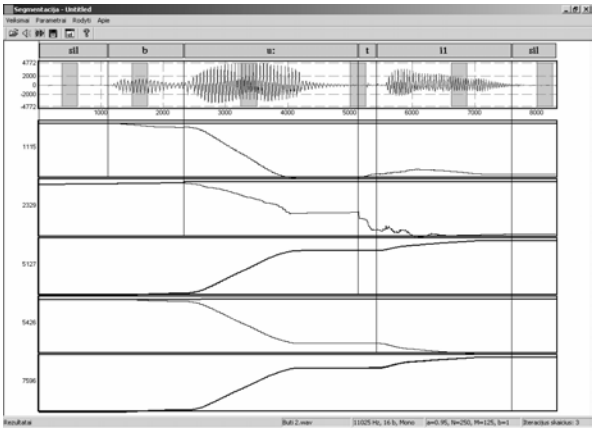

Fig. 2. Word segmentation results. Vertical lines indicate phones boundaries. Partial likelihood and Bellman functions corresponding to each detected phone are shown in lower part of the diagram. Above the diagram user defined labels are shown.

After segmentation each phone can be labeled manually. Text labels are shown above speech utterance diagram. Detected boundaries and labels can be saved into file, which can be used for word segmentation into phones preview.

7. Conclusions

Statistical speech segmentation into phones method is proposed.

1. Maximization of the objective function problem is solved using dynamic programming method.
2. Problem of unknown LPC model parameters is solved using generalized expectation maximization approach.
3. Segmentation software is developed and illustrative segmentation examples are given.

Future work should be concentrated on performance evaluation and optimization of analysis parameters.

References

1. **Svendsen T., F.K. Soong.** On the automatic segmentation of speech signals // Proc. ICASSP. – 1987. – P. 141 – 145.
2. **Leung H.C., V.W. Zue.** A procedure for automatic alignment of phonetic transcriptions with continuous speech // Proc. ICASSP. – 1984. – Vol. c-1. – P. 2.7.1 – 2.7.4.
3. **Hohne H.D., C. Coker, S.E. Levinson, L.R. Rabiner.** On temporal alignment of sentences of natural and synthetic speech // //IEEE Trans. on Acoustics, Speech and Signal Processing. – 1983. – Vol. c-31(4). – P. 807 – 813.
4. **Wagner M.** Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms // Proc. ICASSP. – 1981. – P. 1156 – 1159.
5. **Atal B.S.** Efficient coding of LPC-Parameters by temporal decomposition // Proc. ICASSP. 1983. – Vol. c-1. – P. 81 – 84.
6. **Bridle J.S., N.C. Sedgwick.** A method for segmentic acoustic patterns with application to automatic speech recognition // Proc. ICASSP. – 1977. – P. 656 – 659.
7. **Andre-Obrecht R.** A new statistical approach for the automatic segmentation of continuous speech signals //IEEE Trans. on Acoustics, Speech and Signal Processing. 1988. – Vol. c-36(1). – P. 29 – 40.
8. **Brandt A.** Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test // Proc. ICASSP. – 1983. – P. 1017 – 1020.
9. **Basseville M., A. Benveniste.** Sequential detection of abrupt changes in spectral characteristics of digital signals // IEEE Trans. on Inform Theory. – 1983. – Vol. c-29(5). – P. 709 – 723.
10. **Lavielle M.** Optimal segmentation of random processes signals // IEEE Trans. on Signal Processing. – 1998. – Vol. c-46(5). – P. 1365 – 1373.
11. **Lipeika A.** Optimal Segmentation of Random Sequences // INFORMATICA. – 2000. – Vol. c-11(3). – P. 243 – 256.
12. **Cooper L., M. Cooper.** Introduction to Dynamic Programming. – Pergamon Press, 1981.
13. **Duda R.O., P.E. Hart, D.G. Stork**. Pattern Classification. – John Wiley & Sons, 2001.

Pateikta spaudai 2005 07 04

A. Lipeika, G. Tamulevičius. Žodžių segmentavimas į garsus // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2006. – Nr. 1 (65). – P. 11 – 15.

Darbe nagrinėjama žodžių segmentavimo į garsus galimybė. Ištarimų garsų riboms nustatyti taikomas atsitiktinių sekų pasikeitimo momentų aptikimo metodas. Daroma prielaida, kad garsai yra stacionarios signalo atkarpos, o signalo parametrų pasikeitimo momentai atitinka garsų ribas. Parametrų pasikeitimo momentai aptinkami maksimizuojant jų tikėtinumo funkciją arba minimizuojant vidutinę kvadratinę prognozės klaidą (pasirinktinai). Remiantis sudaryta metodika, sukurta ištarimų segmentavimo programinė įranga. Tikslo funkcijoms optimizuoti panaudotas dinaminio programavimo metodas. Nežinomiems garsų parametrams įvertinti pritaikytas apibendrintasis matematinės vilties maksimizavimo metodas. Programinėje įrangoje numatyta galimybė keisti įvairius signalo apdorojimo parametrus, pasirinkti vieną iš dviejų segmentavimo optimalumo kriterijų. Segmentams gali būti suteikiamos teksto žymės, o jų ribos ir žymės išsaugojamos failuose, kurie vėliau gali būti panaudoti segmentavimo rezultatų peržiūrai. Il. 2, bibl. 13 (anglų kalba; santraukos lietuvių, anglų ir rusų k.).

A. Lipeika, G. Tamulevicius. Segmentation of Words Into Phones // Electronics and Electrical Engineering. – Kaunas: Technologija, 2006. – No. 1 (65). – P. 11 – 15.

Word segmentation into phones is studied in this paper. The method of change point detection in random sequences is used for phone boundaries detection in a word. It is assumed that phones are stationary signal segments and changes of signal parameters are the boundaries of these phones. Change moments are detected maximizing change points likelihood function or minimizing prediction least mean squared error. On the ground of formulated methodology word segmentation software was developed. Dynamic programming was used for object function optimization and for unknown parameters estimation generalized expectation maximization algorithm was used. Developed software allows to control speech signal processing parameters, to choose segmentation optimality criterion. Detected segments can be labeled, and these segments labels and boundaries can be saved into files for later segmentation results preview. Ill. 2, bibl. 13 (in English; summaries in Lithuanian, English and Russian).

А. Липейка, Г. Тамулявичюс. Сегментация слов на звуки // Электроника и электротехника. – Каунас: Технология, 2006. - №. 1 (65). – С. 11 – 15.

В статье рассматривается возможность сегментирования слов на звуки. Для определения границ звуков в слове применяется метод определения моментов изменения случайных последовательностей. Предполагается, что звуки слова – это стационарные сегменты сигнала, а моменты изменения параметров сигнала соответствуют границам звуков. Обнаружение моментов изменения основана на максимизации функции правдоподобия моментов изменения или минимизации ошибки прогноза. По сформулированной методике создано программное оборудование для сегментирования слов. Для оптимизации целевой функции использован метод динамического программирования, неизвестные параметры оцениваются методом максимизации математического ожидания. В созданой программе сегментации предусмотрена возможность управлять параметрами обработки речевого сигнала и выбор между двумя критериями оптимальности сегментации. Выделеным звукам слова могут быть присвоены текстовые метки. Границы и метки звуков могут быть сохранены в виде файлов для пересмотра результатов сегментации. Ил. 2, библ. 13 (на английском языке; рефераты на литовском, английском и русском яз.).

DOI: 10.5755/j02.eie.10542