Unsupervised Segmentation of Continuous Speech Using Vector Autoregressive Time-Frequency Modeling Errors

Petri Korhonen, Unto K. Laine

Laboratory of Acoustics and Audio Signal Processing
Helsinki University of Technology, Espoo, Finland
petri@acoustics.hut.fi, unto.laine@hut.fi

Abstract

A vector autoregressive (VAR) model is used in the auditory time-frequency domain to predict spectral changes. Forward and backward prediction errors increases at the phone boundaries. These error signals are then used to study and detect the boundaries of the largest changes allowing the most reliable automatic segmentation. Using a fully unsupervised method yields segments consisting of a variable number of phones. The quality of performance of this method was tested with a set of 150 Finnish sentences pronounced by one female and two male speakers. The performance for English was tested using the TIMIT core test set. The boundaries between stops and vowels, in particular, are detected with high probability and precision.

1. Introduction

Many subfields of speech technology need robust methods for automatic phonetic speech segmentation. Preferably these methods would be fully speaker and language independent. They should perform segmentation without any prior information about the speaker or the utterance in question. These methods should not apply any type of prior learning, and they should be able to process unknown utterances in a fully unsupervised manner. This paper describes a preliminary test of a novel method for automatic speech segmentation, which fulfills the hard demands mentioned to a certain degree.

Segmentation methods described in the literature can be classified into explicit and implicit methods. They also vary in terms of segmentation units (e.g. phonemes, syllables, words). In explicit methods, the underlying phoneme sequence is known prior to the segmentation. These methods are used in speech synthesis for example. Implicit methods split the utterance into smaller units without using any information about the underlying phoneme sequence. These methods are based on analyzing the acoustic properties of the signal and detecting either spectrally stable parts or rapid variations of signal. An example of a method based on locating spectrally stable parts is in [1] where the correlation between parameters computed from nearby frames has been used as a measure of stability. In [2], segment boundaries are implicitly detected comparing the means of frames around potential boundaries using "jump-function." In [3], the variations of short-term energy function is used as a measure to produce syllable-like units using minimum phase group delay functions.

In the case of continuous speech, the signal cannot be strictly divided into stable and varying parts which would correspond one-to-one with phones and segment boundaries. No phone in continuous speech produces steady spectra, but instead within a phone there are always slow spectral movements which are, to some degree, possible to predict. The method proposed in this paper does not detect these slow spectral variations, but rather is based on detecting unpredicted changes in auditory time-frequency picture of speech at phone boundaries. These unpredicted changes happen most often when moving from one phoneme class to another. Change in the speech production mechanism changes the acoustic signal in an unpredictable manner. Knowing that not all transitions produce a large or rapid spectral change, a question of this study is which kind of phone boundaries allows the most reliable and robust detection by the method.

When facing speaker-independent unlimited vocabulary (e.g. inflectional languages) continuous speech recognition, the words have to be split into smaller units such as morphemes; hence, not every phone boundary needs to be detected. Segments similar to syllables or morphemes consisting of one to many phones do apply as well as long as the total number of different segments is not too high for modeling purposes.

The novel method presented in this paper produces segments consisting of phone clusters of different lengths. The core idea is to model the spectral variation by using Vector Autoregressive model (VAR). The model performs forward and backward predictions in the auditory time-frequency domain with associated prediction errors. The segment boundary candidates are found based on these error signals.

2. VAR Model

The VAR(p) model is defined as

$$\mathbf{y}_t = \mathbf{A}(1)\mathbf{y}_{t-1} + \ldots + \mathbf{A}(p)\mathbf{y}_{t-p} + \mathbf{v} + \mathbf{u}_t \qquad (1)$$

where $\mathbf{y}_t$ is a $(K \times 1)$ vector of observations at time t, $\mathbf{A}(i)$ are fixed $(K \times K)$ coefficient matrices, $\mathbf{v}$ is a fixed $(K \times 1)$ vector of intercept terms allowing for the possibility of nonzero mean $E(\mathbf{y}_t)$ and $\mathbf{u}_t$ is a $(K \times 1)$ vector of white noise with non-singular covariance matrix $\mathbf{C}_u$. The coefficients $\mathbf{A}(1), \ldots, \mathbf{A}(p)$, and $\mathbf{C}_u$ are unknown parameters, which will be estimated from the time series data using multivariate least squares estimation, as in [4]. The VAR(1) is called stable if all eigenvalues of $\mathbf{A}(1)$ have modulus less than 1. It is worth pointing out that the process $\mathbf{y}_t$ for $t = 0, 1, 2, \ldots$ may also be defined even if the stability condition is not satisfied.

The algorithm presented here uses only first order VAR models. VAR(1) model predicts the vector at time t from a

vector at time $t-1$.

$$\hat{\mathbf{y}}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{v} \tag{2}$$

Model $\mathbf{A}$ is estimated from a sequence of vectors $\mathbf{y}$ using the least squares estimate method. The error used for estimation of the model $\mathbf{A}$ is the average one-step prediction error between subsequent vectors within the data window.

3. Algorithm

The digital speech signal $s(n)$ is converted into a sequence of short-time features $\mathbf{y}_t$ with each being a $(K \times 1)$ vector. The short-time features should be computed with relatively short intervals to obtain adequate time resolution for the purpose. This typically leads to overlap between the subsequent frames. The feature vectors $\mathbf{y}_t$ are then median filtered along the time axis in a short time window around each vector.

Let us define $\mathbf{A}_t$ as the VAR(1) model computed from the L data vectors ending at vector at time t.

$$\mathbf{A}_t = VAR_{LSE}(\mathbf{y}_{t-L+1}, \ldots, \mathbf{y}_t) \tag{3}$$

The value for L should correspond to the average length of a predictable part of a phoneme in speech. For each vector $\mathbf{y}_t$ we recursively compute M estimates, and cumulative relative prediction errors $e_{t1}, \ldots, e_{tM}$ with models $\mathbf{A}_{t-M}, \ldots, \mathbf{A}_{t-1}$.

$$e_{tj} = \sum_{i=1}^{j} \left(\frac{\sum^{K} (\mathbf{y}_{t-i} - \mathbf{A}^{i}_{t-j}\mathbf{y}_{t-j})^2}{\sum^{K} \mathbf{y}^2_{t-i}} \right) \qquad j = 1, \ldots, M \tag{4}$$

The median value of the errors represents the final error at time t

$$e_t = \mathsf{median}(e_{t1}, \ldots, e_{tM}) \tag{5}$$

The small values of the e_t are emphasized taking the logarithm of the error signal e_t

$$E_t = 10 * \log_{10}(1 + e_t) \tag{6}$$

Model $\mathbf{A}$ is used to predict the values for $\mathbf{y}$ outside the window, from which the model was estimated. Until this point the model $\mathbf{A}$ has been used to recursively produce vectors for time instances $t \ldots t + M$. This means that the model predicts the *future* values of $\mathbf{y}$. The model can be used to predict the values *before* the window as well. This can be easily done by time reversing the original sequence of vectors $\mathbf{y}$, and performing the same VAR analysis. The signals E_{t+} and E_{t-}, denoting forward and backward prediction error, respectively, are produced in the aforementioned manner. These signals are presented in Fig. 1 (b). To help with visualization, E_{t-} has been negated.

In the next step of the algorithm, the errors E_{t+} and E_{t-} are combined to a single error E_{t*} by

$$E_{t*} = E_{t+} - E_{t-} \tag{7}$$

The resultant error should have a large negative peak before, and large positive peak after, a segment boundary as shown in Fig. 1 (c). The candidates for segment boundaries are now located between these two points. In order to help detection, the signal E_{t*} is filtered with

$$h(t) = \begin{cases} \frac{t}{d} + 1 & -d < t < 0 \\ 0 & t = 0 \\ \frac{t}{d} - 1 & 0 < t < d \end{cases} \tag{8}$$

where d is set to be approximately the average width of peaks in the error signal E_{t*}. Filtering E_{t*} with $h(t)$ gives a signal with peaks at the segment boundaries (Fig. 1 (d)). In this work the selection of the peaks from E_{t*} was done hierarchically by manually setting a minimum segment length and minimum peak height.

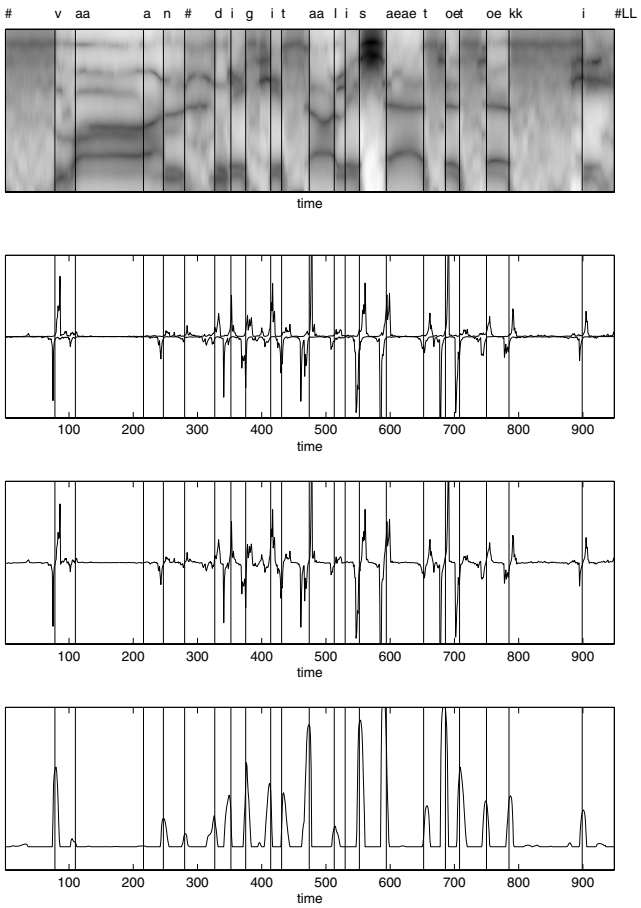

Figure 1: Algorithm step-by-step: (a) median filtered auditory spectrogram of speech, (b) prediction errors E_{t+} and $-E_{t-}$, (c) combined prediction error E_{t*}, (d) E_{t*} filtered with $h(t)$. Vertical lines correspond to manually assigned phonetic boundaries.

4. Evaluation Results

4.1. Evaluation criterion

The evaluation of performance of a segmentation algorithm is not straightforward. The method presented here detects acoustic landmarks, and it would be desirable that these time instances would correspond to phonetic landmarks. Thus phonetic transcription was used to evaluate the performance.

At the present phase of the study, the ultimate goal is not to produce a comprehensive phonetic segmentation by detecting every phone boundary. However, it may be expedient to see how far the basic method leads us. It is worth noting that differences between manual segmentation and automatic segmentation are not necessarily errors, especially if the differences occur in a systematic manner. Differences between automatic segmentation and the phonetic transcription might also arise from inconsistencies in the nature and culture of the phonetic transcription performed by humans.

For the evaluation of the proposed method the often used performance measures - precision, recall, and F-score - were computed. Precision is defined as the number of correct boundaries found by the method divided by the total number of boundaries found by the method. Recall is defined as the number of correct boundaries found by the method divided by the true number of boundaries in the data. F-score is defined as

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

4.2. Evaluation data

The performance of the proposed method was tested on both Finnish and English material. The Finnish material consisted of 150 sentences spoken by two males and one female (50 sentences each) recorded in an anechoic chamber in using 22,05 kHz sampling frequency. Median length for a phone was 105ms, and mean was 115ms. Short-time feature vectors consisted of 14th order frequency warped line spectrum frequencies (WLSF) and logarithmic energy computed every 3ms with a 25ms hamming window. Frequency warping was used to produce auditory representation, and to reduce the linear prediction model order p [6]. The short step-size 3ms was required in order to have enough data for estimation of the VAR model, and also to achieve higher time resolution for the segmentation. This also matches the time resolution of the human hearing more accurately. Features were then median filtered in a 33 ms time window, and each value was normalized between -1 and 1. The method was also tested on English material, which consisted of the core test set of the TIMIT [5] database. WLSF order for English material was set to 12, due to the different original sampling frequency. In both test sets the signal is not corrupted with background noise.

4.3. Segmentation Results

4.3.1. Overall results

Table 1 summarizes the segmentation results for the three speakers from Finnish test set. The highest precision probability was achieved with the female, whereas the highest F-score was obtained with the male 1. Recall for male 1 was better than for the other speakers. The F-score did not vary radically between the speakers. The result confirms that the method is, by and large, speaker independent. The segmentation results for English material are shown in Table 2 for three different values for maximum allowed segmentation error. Recall for the English test set is lower than that of the Finnish, due to the differences in segmentation conventions. In the Finnish test set, the bursts of stop consonants were not labeled. Precision for the English test set is slightly better than for Finnish. Results obtained with the English material suggests that the method is, by and large, language independent. Figure 2 summarizes the

Table 1: *Segmentation results for three speakers in Finnish test set. ($M = 7$, $L = 66ms$, minimum segment length $45ms$, minimum $E_{t*} = 1.0$, median filter width = $33ms$, maximum allowed deviation = $15ms$)*

	Precision	Recall	F-score
Male 1	82.5%	71.3%	76.5%
Male 2	85.3%	65.6%	74.2%
Female	85.6%	63.1%	72.7%

Table 2: *Segmentation results for TIMIT core test set. ($M = 7$, $L = 66ms$, minimum segment length $45ms$, minimum $E_{t*} = 1.0$, median filter width = $33ms$)*

Max. deviation	Precision	Recall	F-score
20 ms	91.5%	56.8%	70.0%
15 ms	87.0%	54.0%	66.6%
10 ms	69.6%	43.1%	53.2%

effect of data window length L for performance for Finnish (male 1). Three different maximum allowed deviations from manual segmentation are included. The model data window length L does not radically affect the F-score, but the precision and recall increase and decrease respectively as we increase L.

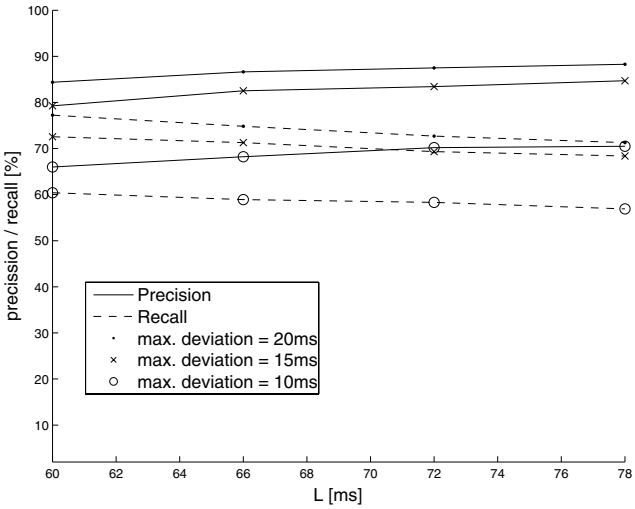

Figure 2: Effect of selection model data window length L for performance with different allowed deviations from manual segmentation.

Figure 3 shows the deviations from manually assigned segment boundaries for male speaker 1. Over 71% of the detected segment boundaries fall within 6 ms from the manually assigned boundaries, and 87.5% of the detected boundaries are within 9ms from manual segmentation, indicating that the method has a good temporal resolution.

Performance of the method was tested for different types of phoneme transitions. The set of Finnish phonemes was divided into seven subclasses based on their phonetic similarity (Table 3). Theoretically, there are 49 different kind of transitions between the classes. Six transitions were not present in

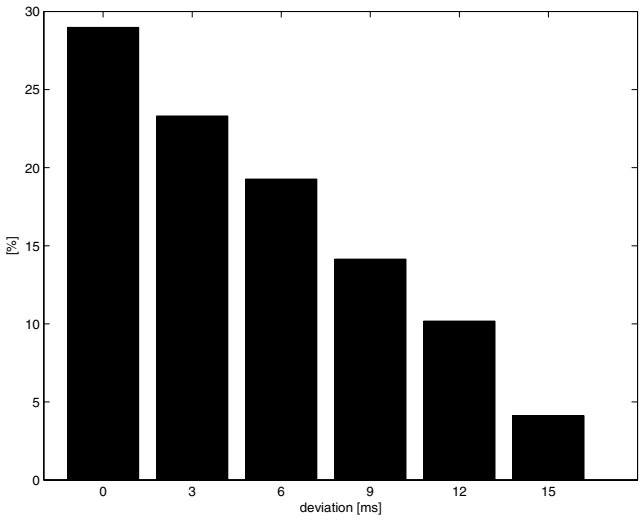

Figure 3: Deviations from manually assigned segment boundaries.($M = 7$,$L = 66$ms, minimum segment length $45ms$, minimum $E_{t*} = 1.0$, median filter width = 33ms)

the material. Three of them are not realizable at all, or there are conflicting phonological rules of Finnish (marked with $\times$)). The cases not present in the material is marked with 0. The stop-vowel and nasal-vowel are detected with over 93% of the cases. Out of all commonly occurring transitions, the vowel-vowel boundaries are the most difficult to detect. Insertions rate is highest in trills.

Table 3: Recall [%] between different phoneme classes, and number of insertions in each class. n is the total number of occurances. Male speaker 1, $M = 7$, $L = 66ms$, min. E_{t*} = 1.0, max. allowed deviation = 15ms. vow: /a/, /e/, /i/, /o/, /u/, /y/, /ae/, /oe/, stop: /b/, /d/, /g/, /k/, /p/, /t/, nas: /n/, /m/, /ng/, fri: /f/, /h/, /s/, sem: /j/, /l/, /v/, tri: /r/, sil: silence

	vow	stop	nas	fri	sem	tri	sil
vow	12.3	82.0	81.4	89.4	70.1	52.6	87.5
$n =$	114	189	113	94	87	38	40
stop	95.7	62.5	100.0	91.6	100	100	0
$n =$	235	8	5	2	6	8	1
nas	93.1	81.5	50.0	90.0	42.9	100	16.7
$n =$	72	27	2	10	7	1	12
fri	88.6	100	100	0	100	66.7	0
$n =$	105	13	3	1	8	3	1
sem	62.1	50	100	80	25.0		
$n =$	103	4	1	5	4	$\times$	0
tri	25.6	83.3	100	33.3			
$n =$	43	6	2	3	0	$\times$	0
sil	100	28.5	100	55.6	100	50	
$n =$	3	28	5	9	5	4	$\times$
Ins.	82	69	17	18	7	25	7
$n =$	675	275	131	134	117	54	104

5. Conclusion and Perspectives

A novel method to detect unpredictable auditory time-frequency changes in acoustic signals was introduced. The method is based on VAR-modeling of auditory spectrograms and thus does not apply any *a priori* knowledge of the signals chosen for segmentation. Therefore, the method is fully unsupervised and immediately applicable without any prior training. Due to its universal character, the method may be applied for segmentation of other types of audio material as well.

6. Acknowledgments

This research was funded by the Academy of Finland. The authors would like to thank Jouni Pohjalainen, M.S.(Eng), for his implementation of VAR model Matlab codes used in the study.

7. References

[1] Jan P. van Hemert, "Automatic Segmentation of Speech", IEEE Transactions on Signal Processing, Vol. 39, No. 4. April 1991.

[2] Guido Aversano, Anna Esposito, Antonietta Esposito, Maria Marinaro, "A New Text-Independent Method for Phoneme Segmentation", in Proc. the 44th IEEE Midwest Symposium on Circuits and Systems, vol. 2, pp. 516–519, 2001.

[3] V. Kamakshi Prasad, T. Nagarajan, Hema A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions", Speech Communications, 42 (2004) 429-446.

[4] H. Lütkepohl, "Introduction to Multiple Time Series Analysis", 2nd edition, Springer-Verlag, 1993.

[5] W.M. Fisher, G.R. Doddington, K.M. Goudie-Marshal, 1986. "The DARPA speech recognition research database: specification and status", In: Proc. DARPA Workshop on Speech Recognition. pp. 93-99.

[6] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency warped signal processing for audio applications", Journal of Audio Engineering Society, vol. 48, pp. 1011-1031, November 2000.