LOGITBOOST WEKA CLASSIFIER SPEECH SEGMENTATION

Bartosz Ziółko, Suresh Manandhar and Richard C. Wilson

Department of Computer Science
University of York
Heslington, YO10 5DD, York, UK
{bziolko,suresh,wilson}@cs.york.ac.uk

Mariusz Ziółko

Department of Electronics
AGH University of Science and Technology
al.Mickiewicza 30, 30-059 Kraków, Poland
ziolko@agh.edu.pl

ABSTRACT

Segmenting the speech signals on the basis of time-frequency analysis is the most natural approach. Boundaries are located in places where energy of some frequency subband rapidly changes. Speech segmentation method which bases on discrete wavelet transform, the resulting power spectrum and its derivatives is presented. This information allows to locate the boundaries of phonemes. A statistical classification method was used to check which features are useful. The efficiency of segmentation was verified on a male speaker taken from a corpus of Polish language.

Index Terms— speech segmentation, WEKA, machine learning, classifier, LogitBoost

1. INTRODUCTION

In the vast majority of approaches to speech recognition, signals need to be divided into segments before recognition can take place. The properties of the signal contained in each segment are then assumed to be constant, or in other words to be characteristic for a single part of speech. The speech segmentation is often conducted as a part of further analysis in speech recognition systems.

Segmentation is a task of splitting a sequence into meaningful units. For a speech input, these units can be phonemes, words or sentences etc. The segmentation problem can be viewed as an unlabelled splitting problem where the input sequence needs to be split into a sub-sequences.

In many applications, like speech recognition, the most frequently used segmentation is the constant-time framing, for example into 23.2 ms blocks [1]. This method benefits from simplicity of implementation and the ease of comparing blocks of the same length. However, the different length of phonemes is a natural phenomenon which cannot be ignored. Constant segmentation therefore risks losing information about the phonemes due to merging different sounds into single blocks, losing phoneme length information and losing complexity of individual phonemes. Moreover, boundary effects provide additional distortion. A more satisfactory approach is an attempt to find the natural phoneme boundaries. A number of approaches have been previously suggested for this task [2, 3, 4].

We used WEKA LogitBoost classifier [5] for speech segmentation. Experiments with many different sets of features were conducted. All of them are based on discrete wavelet transform (DWT) but several combinations of smoothing procedures, derivatives, context and normalisation techniques were checked.

2. DISCRETE WAVELET TRANSFORM

The human hearing system is equipped with frequency processing system in the first step of sound analysis. While the details are still not fully understood, it is clear that a frequency based analysis of speech reveals important information. The wavelet transform belongs to the group of frequency transforms. This encourages us to use a DWT as a method of speech analysis, since the DWT has some features similar to the principles of the operation of human hearing system [6]. The wavelet transform provides a time-frequency spectrum. The original speech signal $s(n)$ defined for the discrete tine n and its wavelet spectrum are of 16 bits accuracy. In order to obtain DWT, the coefficients of series

$$s(n) = \sum_i c_{M,i}\, \phi_{M,i}(n) \tag{1}$$

are computed, where

$$\phi_{M,i}(n) = 2^{M/2}\phi(2^M n\Delta t - i) \tag{2}$$

is the value of ith wavelet function at the Mth resolution level under the time discretisation Δt. Due to the orthogonality of wavelet functions (2) we obtain

$$c_{M,i} = \sum_{n\epsilon D_i} s(n)\phi_{M,i}(n), \tag{3}$$

where D_i are supports of $\phi_{M,i}$. The coefficients of the lower level are calculated by applying [7, 8] formulae

$$c_{M-1,n} = \sum_i h_{i-2n}\, c_{M,i} \tag{4}$$

$$d_{M-1,n} = \sum_i g_{i-2n}\, c_{M,i} \tag{5}$$

where h and g are the constant coefficients which depend on the wavelet ϕ (i.a. the base function in (2)).

There is a wide variety of possible basis functions from which a DWT can be derived. Meyer wavelets (see Fig. 1) gave the best results in our previous experiments on speech segmentation [2] so we used them also to verify the methods presented in this paper.

The speech spectrum is decomposed by digital filtering and downsampling procedures defined by (4) and (5). It means that given the wavelet coefficients $c_{m,i}$ of the mth resolution level, (4) and (5) are applied to compute the coefficients of the $(m-1)$th resolution level, where $M \leq m \leq 2$. The elements of the DWT for a particular level may be collected into a vector, for example $\mathbf{d}_m = (d_{m,1}, d_{m,2}, ...)^T$. The multiresolution analysis gives a hierarchical and fast scheme for the computation of the wavelet coefficients for a given speech signal s. The coefficients of all resolution levels are calculated recursively by applying formulae (4) and (5). In this way the values

$$\mathrm{DWT}(s) = \{\mathbf{d}_M, \mathbf{d}_{M-1}, ..., \mathbf{d}_1, \mathbf{c}_1\} \tag{6}$$

of the DWT for $M+1$ levels are obtained.

The wavelet spectra are produced by cascading filtering and downsampling operations in a tree-structure. The root of the tree consists of the coefficients (3) of wavelet series (1) for the original speech signal. The first level of the tree is the result of procedure (5). Subsequent levels in the tree are constructed by recursively applying (4) and (5) to split the spectrum into the low (approximation $c_{m,n}$) and high (detail $d_{m,n}$) parts, where $M-1 \leq m \leq 1$. The higher level component have two times wider frequency bands when compare with the subsequent lower frequency components.

Experiments that we have undertaken show that the speech signal decomposition into six levels is sufficient (see Fig. 2) to cover the frequency band of a human voice (see Table 1). The energy of the speech signal above 8 kHz and below 125 Hz is very low and can be neglected. The analysis of the power in different frequency subbands gives an excellent opportunity to distinguish the beginning and the end of phonemes.

3. EXPERIMENTAL SETUP

The six DWT subbands (see Tab. 1) were used as a basis to create frequency features. We combined DWT values in high frequency subbands to have units of the same lengths as for the lowest subband. They are 4 ms long, this is a distance between two samples in DWT spectrum for the lowest frequency band. It is very reasonable to include first and/or second derivatives to improve results in any speech analysis task [1]. We used one or two derivatives and we included also left and right context, namely features of signal before and after an analysed part. Features describing a whole power of signal where also used. We smoothed analysed functions and some of derivatives. We also tested how normalisation can improve the number of correctly classified boundaries.

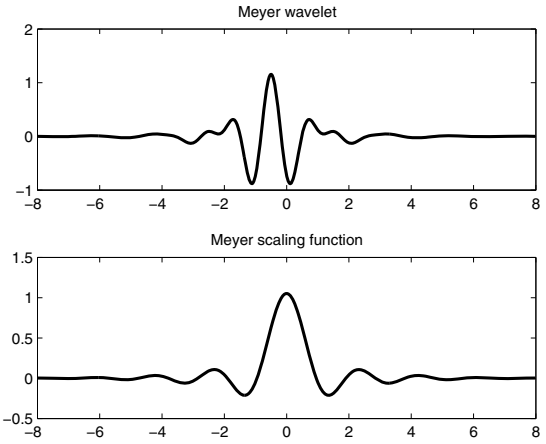

Fig. 1. The Meyer wavelet of order 4 was used in our experiments. Our previous experiments proved it is the best choice in speech analysis

Procedure called "boosting" is the important classification methodology. The WEKA LogitBoost classifier is based on well known AdaBoost procedure [9]. The AdaBoost procedure trains the classifiers on weighted versions of the training samples. It gives higher weights for those which are misclassified. That part of procedure is conducted for a sequence of weighted samples. Afterwards the final classifier is defined to be a linear combination of the classifiers from each stage. Logistic Boost [9] uses an adaptative Newton algorithm to fit an additive multiple logistic regression model.

We trained and tested our classification model on a male speaker of a corpus of Polish, called CORPORA, created under supervision of Stefan Grocholewski in Institute of Computer Science, Poznań University of Technology in 1997 [10]. Our experiments were conducted on speech files with the sampling frequency $f_0 = 16$ kHz. This gives sampling period $t_0 = 62.5\ \mu$s. Speech was recorded in an office with a working computer in the background. The database contains 365 utterances (single letters, digits, names and simple sentences).

The hand segmentation itself is not an entirely accurate process because of uncertainties in human perception of the phoneme boundaries. Additionally, overlapping phonemes or partially merged phonemes are a natural phenomena. There is

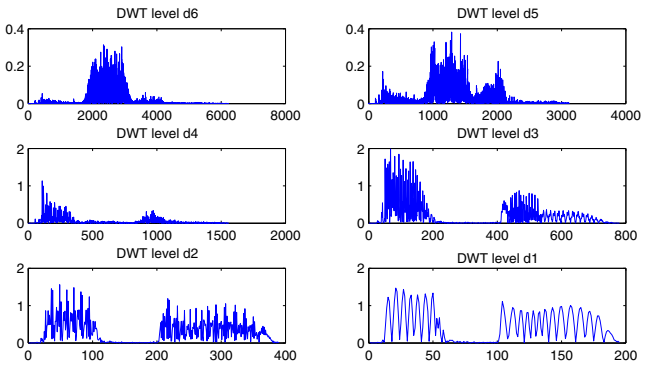

Fig. 2. Subband amplitude of DWT spectra for the Polish word 'osiem' (eng. eight). The number of samples depends on a resolution level

Table 1. Characteristics of the discrete wavelet transform levels and their envelopes

Level	Subband (kHz)	No. of samples	Window
d_6	$8-4$	32	5
d_5	$4-2$	16	5
d_4	$2-1$	8	5
d_3	$1-0.5$	4	3
d_2	$0.5-0.25$	2	3
d_1	$0.25-0.125$	1	3

therefore a degree of uncertainty where a phoneme precisely starts and ends.

There were many more non-boundary points in feature space than those which really represent boundaries. This is why we cloned all sets of features representing phoneme boundaries for 30 times to keep a similar ratio of boundaries and non-boundaries. We used 70 % of all feature points as training data and 30 % for a test in every experiment.

4. EXPERIMENTAL RESULTS

We tested 7 different sets of features for the same classifier and same test data to check which features are useful. The differences between following sets are described. The classification was evaluated using popular precision and recall measure [11] which is presented in tables and by percentage of properly classified instances which are given in text for all cases. Two evaluations are provided for every set of features to help in grading the method because we did not manage to find any other similar system to use to present as a baseline.

We started with one left and one right context subset of features to describe the surrounding part of signal. We included first and second derivatives and both of them were smoothed. Different subbands were smoothed using different windows (see Tab. 1). We found that this method is the most efficient in our previous experiments [2]. That gives 54 features in total. 64 % of test instances were correctly classified. The more exact results using recall and precision evaluation are presented in Tab. 2. The final measure is f-score presented separately for sets of features describing frames with boundaries and without. The second group is named in tables as phonemes. From practical point of view we are interested in detecting boundaries so the evaluation of classification of these frames is crucial. So for the first set of features the most important grade is f-score 0.45 (Tab. 2).

Table 2. Experimental results for the basic set of features

label	precision	recall	f-score
boundary	0.583	0.366	0.45
phoneme	0.659	0.824	0.732

We managed to slightly improve results by leaving the second derivative unsmoothed. There were no other changes in the set of features and 64 % of test instances were correctly classified like for the previous set of feature but the more exact evaluation presented in Tab. 3 indicates some improvement through higher f-score, namely 0.466.

Table 3. Results without smoothing the second derivative

label	precision	recall	f-score
boundary	0.588	0.386	0.466
phoneme	0.665	0.818	0.733

In the next approach we kept the same number and type of features but subband features were normalised by dividing by the energy. In that way 60 % of test instances were correctly classified with f-score only 0.135 (Tab. 4).

Table 4. Results with features normalised by the whole energy

label	precision	recall	f-score
boundary	0.551	0.077	0.135
phoneme	0.607	0.958	0.743

Table 5. Results with features normalised by maximum in a given subband for a given utterance

label	precision	recall	f-score
boundary	0.59	0.317	0.413
phoneme	0.649	0.851	0.737

We tried also another normalising approach, by dividing all features by a maximum in a given subband for an analysed utterence. Around 64 % of test instances were correctly classified but f-score is also quite low, namely 0.413 (Tab. 5).

Surprisingly, none of normalisation methods improved results. Finally, we experimented with wider left and right context, namely we added more subsets of features for signal around the analysed one. We have got 66 % of test instances correctly classified by including two contexts to the left and two to the right. In that case we had a set of 90 features with a relatively high f-score 0.519 (Tab. 6).

Table 6. Results for a set of features with wider context

label	precision	recall	f-score
boundary	0.618	0.447	0.519
phoneme	0.682	0.811	0.741

To use wider context, namely three to the left and three to the right, we had to skip second derivative as the number of features was too large to be operated by WEKA. In that way we had a set of 84 features. 70 % of test instances were correctly classified, but recall for boundary frames was very low, just 0.162 which caused f-score only 0.263 (Tab. 7). It means that generally this set of features is not effective.

Table 7. Results for a set of features without the second derivative but with extra 6 subsets of context features

label	precision	recall	f-score
boundary	0.699	0.162	0.263
phoneme	0.703	0.966	0.814

The three to left and one to right context was also checked. In that experiment we used the second derivatives, so we had 90 features. We received correctness of 70 % but f-score for boundaries was again quite low, only 0.302 (Tab. 8).

Table 8. Results for features with asymmetric context

label	precision	recall	f-score
boundary	0.609	0.2	0.302
phoneme	0.712	0.939	0.81

5. CONCLUSIONS

The identification of the starting and ending boundaries of voice segments in continuous speech is an important problem in different areas of speech processing such as segment-based speech recognition or automatic transcription systems. Spectral analysis is a very efficient method for extracting information from speech signals. Due to described features, the time-frequency speech segmentation and its evaluation is particularly useful in speech recognition. DWT based features can be used to detect phoneme boundaries using machine learning classification methods. The best results can be achieved by including features describing context of 8 ms to the left and 8 ms to the right.

6. REFERENCES

[1] S. Young, "Large vocabulary continuous speech recognition: a review," *IEEE Signal Processing Magazine*, vol. 13(5), pp. 45–57, 1996.

[2] B. Ziółko, S. Manandhar, and R. C. Wilson, "Phoneme segmentation of speech," *Proceedings of 18th International Conference on Pattern Recognition*, 2006.

[3] D. B. Grayden and M. S. Scordilis, "Phonemic segmentation of fluent speech," *Proceedings of ICASSP*, pp. 73–76, 1994.

[4] D.T. Toledano, L.A.H. Gómez, and L.V. Grande, "Automatic phonetic segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, 2003.

[5] I.H. Witten and E. Frank, *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*, Academic Press, 2000.

[6] D. Wang and S. Narayanan, "Piecewise linear stylization of pitch via wavelet analysis," *Proceedings of Interspeech*, 2005.

[7] I. Daubechies, *Ten lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.

[8] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, vol. 8, pp. 11–38, 1991.

[9] J.H. Friedman, T. Hastie and R. Tibshirani", *Additive logistic regression: A statistical view of boosting*, Technical report, Department of Statistics, Stanford University", 1999.

[10] S. Grocholewski, "Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom (eng. Assumptions of acoustic database for Polish language)," *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*, 1995.

[11] C. J. van Rijsbergen, *Information Retrieval*, London: Butterworths, 1979.