14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

WAVELET METHOD OF SPEECH SEGMENTATION

Bartosz Ziółko[*], *Suresh Manandhar*[*], *Richard C. Wilson*[*] *and Mariusz Ziółko* [**]

[*]Department of Computer Science, University of York
Heslington, YO10 5DD, York, UK
[**]Department of Electronics, AGH University of Science and Technology
phone:+(44) 01904 432757, fax: +(44)1904 432767 , email:bziolko@cs.york.ac.uk
web: http://www-users.cs.york.ac.uk/bziolko/

ABSTRACT

In this paper a new method of speech segmentation is suggested. It is based on power fluctuations of the wavelet spectrum for a speech signal. In most approaches to speech recognition, the speech signals are segmented using constant-time segmentation. Constant segmentation needs to use windows to decrease the boundary distortions. A more natural approach is to segment the speech signals on the basis of time-frequency analysis. Boundaries are assigned in places where some energy of a frequency band rapidly changes. Most methods of non-constant segmentation need training for particular data or are realized as a part of modelling. In this paper we apply the discrete wavelet transform (DWT) to analyse speech signals, the resulting power spectrum and its derivatives. This information allows us to locate the boundaries of phonemes. It is the first stage of speech recognition process. Additionally we present an evaluation by comparing our method with hand segmentation. The segmentation method proves effective for finding most phoneme boundaries. Results are more useful for speech recognition than constant segmentation.

1. INTRODUCTION

As information technology has an impact on more and more aspects of our daily lives, the problem of communication between human beings and information-processing devices become increasingly important. Up to now such communication has been run almost entirely by means of keyboards and screens, but speech is by far the most widely used, natural and fast means of communication for people. Unfortunately, machine capabilities for interpreting speech is still poor in comparison to what a human can achieve.

In the vast majority of approaches to speech recognition, the speech 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 of a single part of speech.

The most often used current method is to use constant-time framing, for example into 25 ms blocks [12]. 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. Moreover, boundary effects provide additional distortion (which is typically reduced by applying the Hamming window). Obviously framing creates more boundaries than phoneme segmentation. 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.

A more satisfactory approach is an attempt to find the phoneme boundaries from the time-varying speech signal properties. A number of approaches have been previously suggested for this task [6, 11, 13] but these utilise features derived from acoustic knowledge of the phonemes. Such methods need to be optimised to particular phoneme data and cannot be performed in isolation from phoneme recognition. Neural networks [9] have also been tested, but they require time consuming training. Segmentation can be applied by the segment models (SM) instead of the hidden Markov models (HMM) [7]. This solution groups frames into sequences of frames using modelling. Such a solution means segmentation and recognition are conducted at once and there is a set of possible observation lengths. In a general SM, duration distribution gives the segment length likelihood so in fact it describes the likelihood of a particular segmentation of an utterance. SM for a given label is also characterised by a family of output densities. It gives information about observation sequences of different lengths. These features of SM solution allow to locate boundaries only on several fixed positions dependent on framing (on multiplied length of one frame).

Spectral analysis is a very efficient method for extracting information from speech signals. Discrete wavelet transform (DWT) has been successfully used in many speech processing applications [3, 4, 5, 8, 10] for the spectral analysis of signals. For the speech recognition case, it was mainly used to improve accuracy of parameterisation. Experimental results show superiority of DWT methods over more classic ones like mel-frequency cepstral coefficients (MFCC) [3, 4, 5]. The analysis of the power in different frequency subbands gives an excellent opportunity to distinguish the beginning and the end of phonemes. For many boundaries, there is no discernible drop in overall power, and at some frequencies, the power is broadly constant over the phoneme duration. However, many phonemes exhibit rapid changes in particular subbands which can determine their beginnings and endpoints. Our method differs from most of others because it analyses the signal itself in frequency domain. This means we do not use any information based on modelling or phonemes recognition. The segmentation step can be conducted independently and finished before the recognition step. Additionly it does not need any training or adaptation to the user.

The outline of this paper is as follows: in section 2, we describe the DWT decomposition, as the main tool of our method. In section 3 we present the formation of envelopes and subband filtering to get rate-of-change information. The

segmentation process as a comparison of smoothed power and its rate-of-change function is described as are some other general rules of the segmentation and the method. In section 4 we describe the implemented algorithm. Section 5 presents details of our experiments including comparison of applying different wavelets and constant segmentation (framing). The new general evaluation method for phoneme segmentation is described, and results are presented for a database of Polish words.

2. THE DISCRETE WAVELET TRANSFORM, ITS POWER AND AN ENVELOPE

The human ear uses a frequency processing in the first step of sound analysis [2]. This encourages us to use a DWT in an artificial method of speech analyzing as perceptually motivated solution.

The original signal and its wavelet spectrum are of 16 bits accuracy. The wavelet transform belongs to the group of frequency transforms. As a result, it is easy to find speech parameters which are important for the human hearing system [10]. In order to obtain DWT, the coefficients of series

$$s(t) = \sum_i c_{m+1,i}\, \phi_{m+1,i}(t) \tag{1}$$

need to be computed, where $\phi_{m+1,i}$ is the ith wavelet function at the $(m+1)$th resolution level.

The coefficients of the lower level are calculated by applying the well-known formulae [2, 8]

$$c_{m,n} = \sum_i h_{i-2n}\, c_{m+1,i} \tag{2}$$

$$d_{m,n} = \sum_i g_{i-2n}\, c_{m+1,i} \tag{3}$$

where h and g are the constant coefficients which depend on the assumed pair: scale function ϕ and wavelet ψ. The formulae (2) and (3) are used for the signal decomposition by digital filtering of wavelet coefficients. If there are given the wavelet coefficients $c_{m+1,i}$ of the $(m+1)$th resolution level, we can apply (2) and (3) to compute the coefficients of the mth resolution level. 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 coefficients of other resolution levels are calculated recursively by applying formulae (2) and (3). The multiresolution analysis leads naturally to a hierarchical and fast scheme for the computation of the wavelet coefficients for a given speech signal s. In this way the values

$$\mathrm{DWT}(s) = \{\mathbf{d}_M, \mathbf{d}_{M-1}, ..., \mathbf{d}_1, \mathbf{c}_1\} \tag{4}$$

of the DWT for $M+1$ levels are obtained. The wavelet spectra are produced by using a filter bank (cascading the filtering and downsampling operations). The wavelet transformation can be viewed as a tree. The root of the tree consists of the coefficients of wavelet series (1) of the original speech signal. The next level of the tree is the result of one step of the DWT. Subsequent levels in the tree are constructed by recursively applying the wavelet transform step to split the signal into the low (approximation) and high (detail) parts. The undertaken experiments showed that the speech signal should be decomposed into six levels, which cover the frequency band of a human voice (see Table 1). The energy of the speech signal above 5.5 kHz and below 86 Hz is very low.

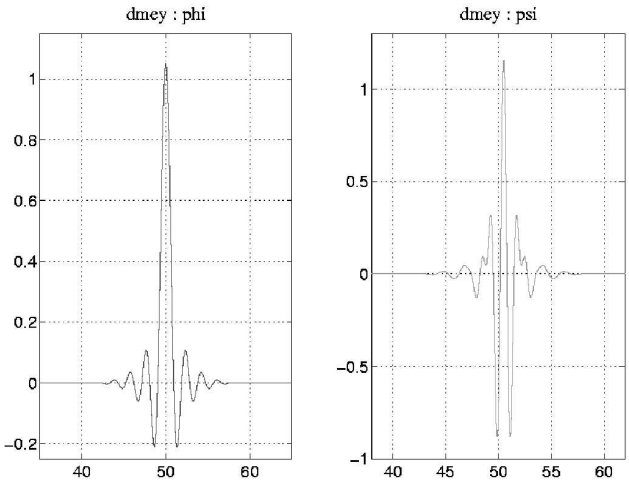

Figure 1: Discrete Meyer Wavelet

The usefulness of six wavelet functions was verified. The obtained results for different wavelets (see Table 2) shows small differences in their efficiency. It seems that discrete Meyer wavelet (Fig. 1) [1] or symlets should be chosen as a basis for the DWT because of their symmetry in time domain and compact support in the frequency domain.

3. SEGMENTATION

Clearly, we would expect the absolute value of the rate-of-change of power to be large at the beginning and at the end of phonemes. However, this does not uniquely define start and end points, for two reasons. Firstly, the power can rise over a considerable length of time at the start of a phoneme, leading to an ambiguous start time. Secondly, there may also be rapid changes in power in the middle of a segment. A better method of detecting the boundary of phonemes relies on power transitions between the DWT subbands.

A properly chosen segmentation method should increase the efficiency of speech recognition. Our approach is based on six level DWT analysis (i.e. $M = 6$) of a speech signal (Fig. 2).

The amount $2^{-M+n-1}N$ of wavelet spectrum samples in n-level (where $n = 1, \ldots, M$) depends on the length N of speech signal in time domain, assuming N is a power of 2. Table 1 presents their number at each level relative to the lowest resolution level. For each n-level decomposition the power waveform

$$p_n(i) = \sum_{j=1}^{2^{n-1}} d^2_{n,j+2^{n-1}i} \qquad \text{where } i = 0, \ldots, 2^{-M}N - 1, \tag{5}$$

is computed in a different way to obtain the equal number of power samples.

The DWT subband power shows rapid variations (see Fig.2). Despite smoothing (5) power waveforms change rapidly. The first order differences in the power are inevitably noisy, and so we calculate the envelopes p'_n for power fluctuations in each subband by choosing the highest values of p_n in a window of given size ω to obtain a power envelope (Fig.3 and Table 1). Additionally we use a smoothed differencing operator. The subband power p_n is convolved with the mask

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

DWT Level	Frequency band (Hz)	Number of samples in compare with level 1	Window size ω
6	2756–5512	32	3
5	1378–2756	16	3
4	689–1378	8	3
3	345–689	4	5
2	172–345	2	5
1	86–172	1	5

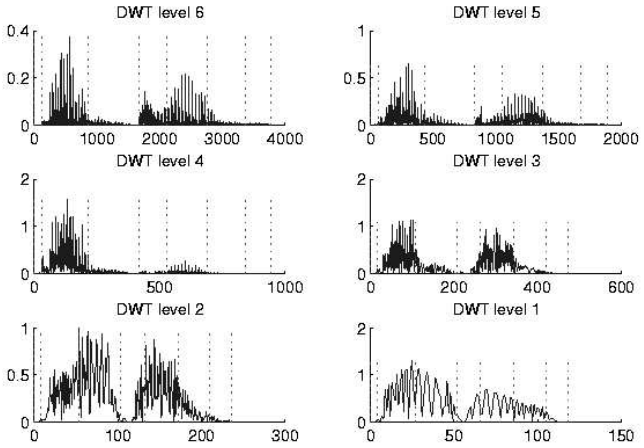

Figure 2: Power of DWT sub-bands of the name 'Andrzej' /ʌ:ndʒɛɪ/. Dotted lines are hand segmentation boundaries.

Table 2: Comparison on constant segmentation and proposed method using different wavelets

Method	av. ε_n	av. ε_p	Overall error
Const 23.2 ms	2.9018	5.6380	20.1472
Const 92.8 ms	0.0796	5.2479	5.6459
Meyer	0.1602	3.2325	4.0334
db2	0.2325	2.8531	4.0157
db6	0.1927	3.0752	4.0385
db20	0.1716	3.2724	4.1305
sym6	0.1816	3.0581	3.9660
haar	0.2663	2.8783	4.2099

$[1,2,-2,-1]$ to obtain smoothed rate-of-change information $r_n(i)$.

The start of a phoneme should be marked by an initially small but rapidly rising power level in one or more of the DWT levels. In other words, we should expect the power to be small and the derivative to be large. We can detect phoneme boundaries searching for *i*-points for which the inequality

$$p \geq |\beta|r_n(i)| - p'_n(i)| \tag{6}$$

holds for the phoneme boundaries, where constant p is a value of threshold which accounts for the time scale and sensitivity of the crossing points. Rate-of-change function r_n is multiplied by scaling factor β approximatelly equal to 1. In practice we seek indexes for which the smoothed power and rate-of-change function approach close to each other and not necessarily cross them. We found the threshold p of the distance between smoothed power and rate-of-change function as 0.02 for the best results. Another condition improving an accuracy is overrunning of a minimal threshold p_{min} of subband DWT power which was chosen experimentally as 0.003. It prevents us from analysing noise instead of the speech signal.

4. PHONEME DETECTION ALGORITHM

The presented above method without additional details would not precisely detect the phoneme boundaries for a number of reasons. Firstly, the precise positions of the boundaries may vary slightly between levels. For some phonemes, only one frequency band may show significant variations in power, for others several. In the second case, each subband analysis will detect separate boundary. They may differ slightly. Secondly, despite smoothing the derivative, near the threshold there may be a number of transitions which represent the same boundary. These problems are overcome by grouping together all transition points across all the bands, provided they are less than time α, apart where α represents the minimum length of a phoneme. We put 5 as its value in the discretised power which represents 29 ms. Neighboring values gave worse results in the evaluation test. The boundary position is the centre of these grouped transition points. Surprisingly we found pre-emphasis filtering as a step degradating quality so we did not use it in the final version of the algorithm.

The algorithm consists of following steps:

1. Normalise a speech signal by dividing by its maximum value.
2. Decompose a signal into six levels of the DWT.
3. Calculate the sum of power samples in all frequency subbands according to Table 1 to obtain (5), the power representations $p_n(i)$ of the nth subband.
4. Calculate the envelopes p'_n for power fluctuations in each subband by choosing the highest values of p_n in a window of a given size ω (Fig. 3 and Table 1).
5. Calculate the rate-of-change function $r_n(i)$ by filtering $p_n(i)$ with [1, 2, -2, -1] mask.
6. Given a threshold p of the distance between $r_n(i)$ and p'_n and a threshold p_{min} of minimal p'_n, find indexes for which $|\beta|r_n(i)| - p'_n(i)| < p$ *AND* $(|\beta|r_n(i+1)| - p'_n(i+1)| > p$ *OR* $|\beta|r_n(i-1)| - p'_n(i-1)| > p)$ *AND* $p'_n(i) > p_{min}$, where $\beta = 1$. Write such indexes in one vector (marked as asterisks in Fig. 3).
7. Find and group indexes where there is no space between neighboring ones longer than attribute α.
8. Calculate an average index value (rounded to the nearest integer) for each group found in the previous step as the

representative of a group. They are indexes of phonemes' boundaries in indexing order of DWT level 1.

5. EXPERIMENTAL RESULTS AND EVALUATION METHOD

In our implementation we assumed the sampling frequency $f_0 = 11025$ Hz. This gives sampling period $t_0 = 90.7$ μs. In order to assess the quality of our results, we have hand-segmented 50 Polish words for comparison. The hand segmentation itself is not an entirely accurate process because of human ear errors. Additionally the phonemes typically overlap each other. The reason for this is that voiced sounds are produced by modulation of the airflow from the lungs by vibration of vocal cords. This modulation react on changes in vocal cords vibrations with a delay. There may be a degree of uncertainty precisely where the phoneme starts and ends, to within a few samples.

The words are segmented not only using our automatic technique. Constant segmentation method where the speech is broken into fixed length segments was also evaluated as a baseline. The quality of segmentation may be assessed on two criteria. Firstly, the right number of segments should be found - the number of segments should correspond to the number of phonemes present in the speech. The error in the number of segments for word w is defined to be

$$\varepsilon_n(w) = \frac{|n_a - n_h|}{n_h} \tag{7}$$

where n_a and n_h are the number of segments in the automatic and hand segmentation respectively.

The second criterion is accuracy of the position of the segmentation. This is based on the closeness of the boundary to the hand-segmented boundary. Since we do not know which boundary corresponds to a particular boundary in the hand segmentation, we take the closest boundary as the correct one. The error in placement for word w is

$$\varepsilon_p(w) = \sum_j \min_i |p_j - q_i| \tag{8}$$

where p_i is the position of the i-th boundary in the automatic segmentation, and q_j is the j-th boundary position in the hand segmentation. Finally, we construct an overall error of

$$\varepsilon(w) = \frac{1}{n_w} \sum_w \alpha \varepsilon_n(w) + \varepsilon_p(w), \tag{9}$$

where n_w is the number of words in evaluation set (50 in our example) and α equals 5 which represents 29 ms. The error in the number of segments $\varepsilon_n(w)$ has a larger impact on the further recognition than the boundary shift represented by $\varepsilon_p(w)$. We decided to scale $\varepsilon_n(w)$ by α the minimum length of a phoneme because boundary displacement smaller than α is typically less degradating than missing the boundary at all. Such a criterion describes the possible inaccuracy of segmentation. It takes into account all important issues however the solution is not without flaws. It counts small differences between hand segmentation and automatic segmentation as errors while such shifts should not be necessarily considered in that way. As it was mentioned before it is difficult to show statistics of correct segmentation because we cannot compare them with the ideal ones. Hand segmentation is not perfect enough to be a fully convincing template.

Table 3: Comparison of detected phoneme boundaries with hand segmentation for word 'Andrzej' /ʌ:ndʒɛɪ/

	Segment boundaries positions									
Auto	0	6	38	45	55	63	86	97	107	118
Hand	0	4	27		52	66	86		105	118

Table 4: The effect of introducing white noise to detected phoneme boundaries (using sym6 wavelet) on error inplacement ε_p. In columns marked by $\pm$ we presented maximal value of noise.

$\pm$	av. ε_p	$\pm$	av. ε_p	$\pm$	av. ε_p
0	3.0581	2	3.3600	6	3.9704
0.5	3.0780	3	3.4931	7	4.0035
1	3.2340	4	3.7579	8	4.0326
1.5	3.2881	5	3.8002	9	4.2065

The evaluation method would be improved if several people do hand segmentation. In such case the template would be not an average value over different observers but rather a short range of possible correct answers for each boundary. The error would be counted if automatic segmentation went outside the range and its value was the distance to the closer end of the range. It would protect, at least partly, the evaluation method from human error of limited accuracy of human ear. The overall error for our data set is summarised in Table 2. The greatest error for the constant segmentation 23.2 ms and much smaller for 92.8 ms is caused by the fact that the length of phonemes is typically around 100 ms in average.

Fig. 3 shows an example of the segmentation process. The six smoothed wavelet bands are shown, along with the automatic and hand segmented boundaries. For this word, the boundary positions are shown in Table 3. The method finds nearly all the boundaries accurately, to within 2 samples, but misplaces one boundary too far to the end of one phoneme and the phoneme was split twice on separate elements.

We did an additional experiment for the further evaluation. We introduced white noise to automatically found phoneme boundaries. We increased the power of noise by adding or substracting larger random values to each boundary in successive comparisons. We found the average error in placement ε_p growing as presented in (Table 4). This shows that evaluation is degradated by introducing white noise to results.

6. CONCLUSIONS

We loose some information on the meaning of speech by constant segmentation. An effective and fast algorithm of speech segmentation allows us to introduce an opportunity of better automatic speech recognition.

The proposed method is based on DWT analysis. It is efficient because some phonemes have power variations in the narrow band only. It is much easier to detect them analysing DWT subsignals than the power of the whole signal. Envelopes of DWT subsignal power should be calculated for an easier and faster analysis of power variations. Rate-of-change information is a crucial parameter for the method. It is easy to detect most single phonemes by using our algo-

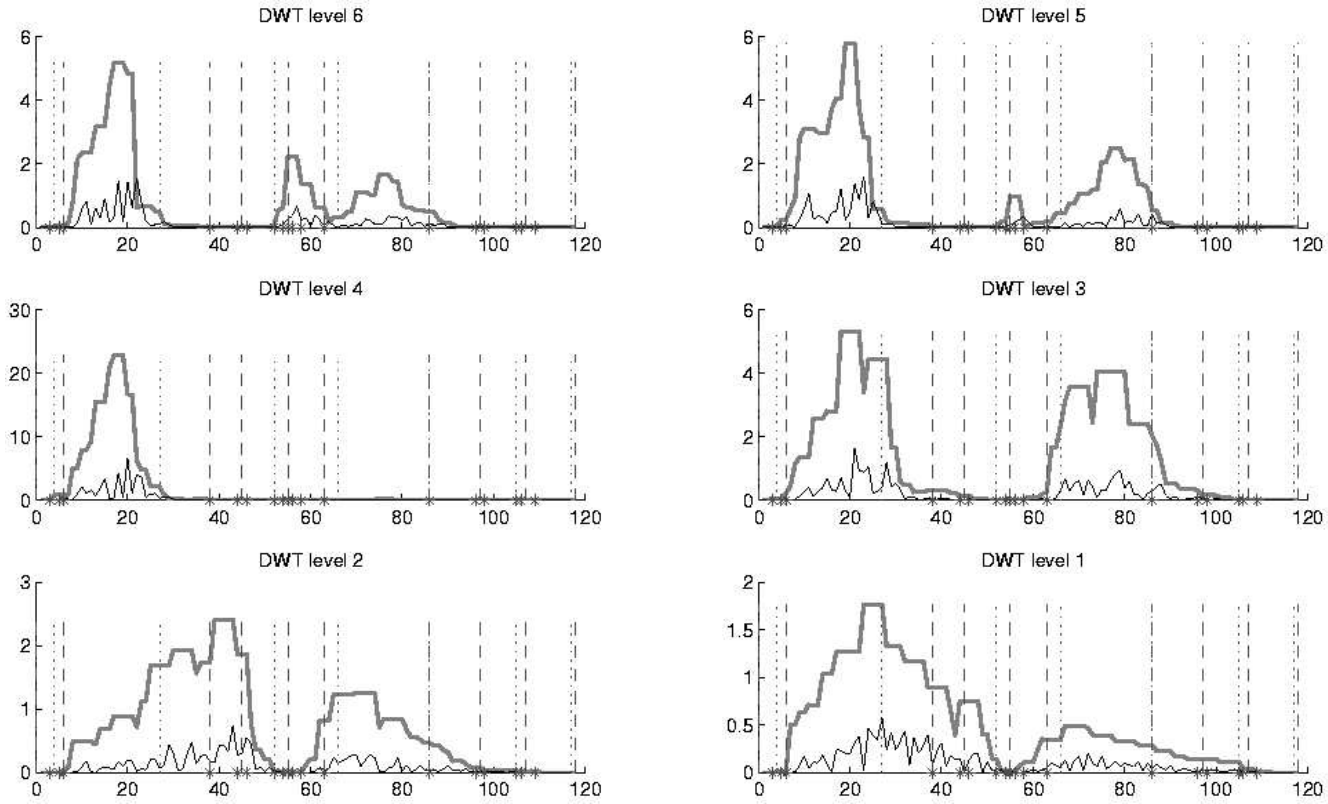

Figure 3: An example of the segmentation of a name 'Andrzej' /ʌ:ndʒɛɪ/. Dotted lines are hand segmentation boundaries; dashed lines are automatic segmentation boundaries, bold grey lines are envelopes and thin lines are smoothed rate-of-change functions. Asterisks are candidates for boundaries for which $r_n(i)$ and $p'_n(i)$ come close to each other or cross (compare with 6-th step of the algorithm).

rithm. Additionally a simple evaluation method of segmentation based on comparing with hand segmentation is presented.

REFERENCES

[1] P. Abry. *Ondelettes et turbulence*. Diderot ed., 1997.

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

[3] M. Deviren and K. Daoudi. Frequency and wavelet filtering for robust speech recognition. *Joint International Conference on Artificial Neural Networks (ICANN)/International on Neural Information Processing (ICONIP)*, Istanbul, 2002.

[4] O. Farooq and S. Datta. Wavelet based robust subband features for phoneme recognition. *IEE Proceedings: Vision, Image and Signal Processing*, 151(3):187–193, 2004.

[5] J.N. Gowdy and Z. Tufekci. Mel-scaled discrete wavelet coefficients for speech recognition. *Proc. of ICASSP*, Istanbul, 2000.

[6] D. B. Grayden and M. S. Scordilis. Phonemic segmentation of fluent speech. *Proc. of ICASSP*, pages 73–76, 1994.

[7] M. Ostendorf, V.V. Digalakis, and O.A. Kimball. From hmm's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4:360–378, 1996.

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

[9] Y. Suh and Y. Lee. Phoneme segmentation of continuous speech using multi-layer perceptron. In *ICSLP 96*, 1996.

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

[11] C. J. Weinstein, S. S. McCandless, L. F. Mondshein, and V. W. Zue. A system for acoustic-phonetic analysis of contimuous speech. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 23:54–67, 1975.

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

[13] V. W. Zue. The use of speech knowledge in automatic speech recognition. *Proc. of the IEEE*, 73:1602–1615, 1985.