5

Blind Segmentation of Speech Using Non-linear Filtering Methods

Okko Räsänen, Unto K. Laine and Toomas Altosaar
Aalto University School of Science and Technology
Finland

1. Introduction

Automated segmentation of speech into phone-sized units has been a subject of study for over 30 years, as it plays a central role in many speech processing and ASR applications. While segmentation by hand is relatively precise, it is also extremely laborious and tedious. This is one reason why automated methods are widely utilized. For example, phonetic analysis of speech (Mermelstein, 1975), audio content classification (Zhang & Kuo, 1999), and word recognition (Antal, 2004) utilize segmentation for dividing continuous audio signals into discrete, non-overlapping units in order to provide structural descriptions for the different parts of a processed signal.

In the field of automatic segmentation of speech, the best results have so far been achieved with semi-automatic HMMs that require prior training (see, e.g., Makhoul & Schwartz, 1994). Algorithms using additional linguistic information like phonetic annotation during the segmentation process are often also effective (e.g., Hemert, 1991). The use of these types of algorithms is well justified for several different purposes, but extensive training may not always be possible, nor may adequately rich descriptions of speech material be available, for instance, in real-time applications. Training of the algorithms also imposes limitations to the material that can be segmented effectively, with the results being highly dependent on, e.g., the language and vocabulary of the training and target material. Therefore, several researchers have concurrently worked on blind speech segmentation methods that do not require any external or prior knowledge regarding the speech to be segmented (Almpanidis & Kotropoulos, 2008; Aversano et al., 2001; Cherniz et al., 2007; Esposito & Aversano, 2005; Estevan et al., 2007; Sharma & Mammone, 1996). These so called blind segmentation algorithms have many potential applications in the field of speech processing that are complementary to supervised segmentation, since they do not need to be trained extensively on carefully prepared speech material. As an important property, blind algorithms do not necessarily make assumptions about underlying signal conditions whereas in trained algorithms possible mismatches between training data and processed input cause problems and errors in segmentation, e.g., due to changes in background noise conditions or microphone properties. Blind methods also provide a valuable tool for investigating speech from a basic level such as phonetic research, they are language independent, and they can be used as a processing step in self-learning agents attempting to make sense of sensory input where externally supplied linguistic knowledge cannot be used (e.g., Räsänen & Driesen, 2009; Räsänen et al., 2008).

This paper introduces a novel method for blind phonetic segmentation of speech that utilizes novel non-linear filtering methods and a short-term FFT representation of signal spectra. The method is compared to existing methods reported in literature and is shown to achieve a very similar level of performance despite the large methodological differences. A careful analysis of errors occurring in the segmentation is performed, shedding light to the question why all blind algorithms fall short of ideal segmentation performance in a similar manner.

2. A novel methodological approach to segmentation

The algorithm is based on the assumption that phonetically meaningful units are manifested as spectrally coherent, relatively steady stretches of a speech signal. To divide a speech signal into non-overlapping units, a segmentation algorithm needs to utilize parameters with specific distance metrics to estimate the similarity or changes in the signal's spectral content. The algorithm introduced here utilizes temporally integrated cross-correlation distances of feature vectors. In the basic version of the algorithm, features are produced by the Fourier transform from speech segments provided by short-term windowing. The straightforward use of FFT coefficients instead of many other possible parametric choices (e.g., MFCC or PLP) was motivated by preliminary findings made during in-house vowel-classification experiments under extremely noisy conditions. The computational simplicity of the FFT was also an influencing factor. In order to compare the effects of auditory modeling to a pure FFT representation, the use of MFCCs was tested and is reported in section 3.5.

In contrast to many prevailing approaches (e.g., Almpanidis & Kotropoulos, 2008; Aversano et al., 2001; Estevan et al., 2007), the FFT analysis is performed in a short (6 ms Hamming) window with a small window shift (2 ms) in order to detect the location of the main vocal tract excitation (after the glottal closure) for voiced sounds. These window locations provide high energy with sharp formants (good spectral contrast), which further improves the detection of formant movements at the segment boundaries as well as the noise robustness of the process. A short window also reduces the smoothing effect of formant frequency modulation during pitch periods and removes the unwanted influence of the fundamental frequency from the features.

The incoming speech signal is first pre-emphasized with a 2nd order FIR filter:

$$y[n] = b_0 x[n] + b_1 x[n-1] + b_2 x[n-2] \tag{1}$$

where values b_0 = 0.3426, b_1 = 0.4945 and b_2 = -0.64 are used according to (Nossair et al., 1995) in order to set the formants to an approximately equal amplitude level. The signal is then windowed with a 6 ms Hamming window and shifted by 2 ms steps. The linear-scale absolute value FFT is then calculated from these 96 samples in the window to create a spectral representation at each frame location, yielding a total of 48 coefficients for 16 kHz signals. The short-term energy (STE) of each 2 ms frame is also stored for further use. The FFT coefficients in each frame are then divided by the mean of their values within the frame and all coefficients are compressed using a hyperbolic tangent mapping in order to simulate the non-linear sensitivity of human hearing:

$$f'[m] = \tanh(\alpha \cdot f[m]) \tag{2}$$

where $\alpha = 0.45$ and $f[m,c]$ is the c'th coefficient at time m.
Once the entire signal has been transformed, a cross-correlation matrix **C** is calculated from the frames, i.e., each element $\mathbf{C}(m_1,m_2)$ indicates the cross-correlation of feature vectors at time m_1 and m_2:

$$C(m_1,m_2)=\frac{f'(m_1)\cdot f'(m_2)}{\|f'(m_1)\|\|f'(m_2)\|} \tag{3}$$

Now the diagonal of the correlation matrix can be considered as the linear time axis that runs through the signal, i.e., from the top-left towards the bottom-right.

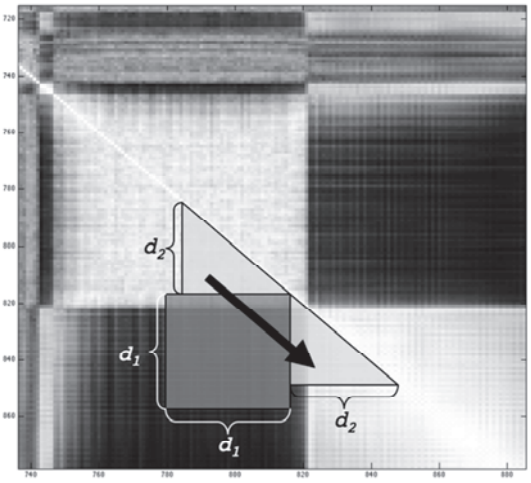

Fig. 1. Part of the correlation-matrix with a superimposed 2D-filter moving along the diagonal. The area under the square at time m corresponds to a[m] and the area under the triangles corresponds to b[m]. Signal frame indices are marked on both axes.

A special 2D-filter is applied to the correlation matrix that is composed of one square region a[m] of size d_1 x d_1 with its top-right corner placed against the diagonal, as well as two identical triangles b[m] with side lengths of d_2 where each hypotenuse is placed next to the diagonal (refer to fig. 1). As the filter moves downwards along the diagonal, the sum of the cross-correlation matrix elements under the triangles b[m] is subtracted from the sum of the elements under the square a[m] at each time step.

$$s[m]=a[m]-b[m] \tag{4}$$

This produces a representation s[m] of the speech signal where large negative peaks reflect significant spectral changes and thus indicate potential segment boundary locations, refer to fig. 2. The resolving capability of s[m] can be adjusted by varying the parameters d_1 and d_2, which is, in the end, basically a trade-off between the temporal accuracy and boundary detection reliability.
Signal s[m] can be noisy especially when using small values of d_1 and d_2 and often results in an overly detailed analysis. The application of a so-called *minmax*-filter is therefore warranted to refine the representation (the minmax-filter is a conceptual modification of the

well known maxmin-filter). As the filter passes through the signal, at each point it takes n_{mm} subsequent samples from s[m] and determines the maximum v_{max} and minimum v_{min} values of this sliding window subvector. The difference of this method compared to common maxmin-filtering is that the filter produces the difference d_{max}=v_{max}-v_{min} as an output at the point where the minimum value was located instead of the center of the time window (note that deep valleys in s[m] indicate the location of segment boundary candidates). The filtering removes small fluctuations and retains only the largest (local) changes in the signal s[m] at the points of local minima. The following pseudo-code describes the functionality of the filter:

$$\begin{aligned} d_{\max} &= \max(s[m:m+n_{mm}]) - \min(s[m:m+n_{mm}]) \\ I &= find\_index(\min(s[m:m+n_{mm}])) \\ s'[m+I] &= d_{\max} \end{aligned} \tag{5}$$

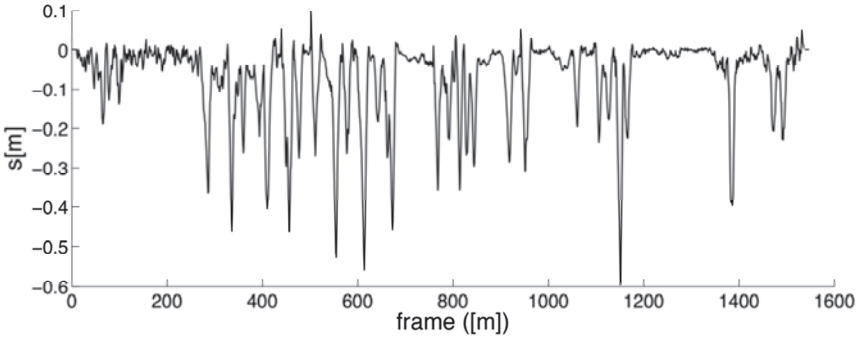

Fig. 2. Signal s[m] produced by the sliding 2D-filter of figure 1. Valleys indicate potential segment boundary locations.

As a result of filtering, signal s'[m] is obtained, refer to fig. 3, in which the estimated segment boundary locations are now represented as easily identifiable positive peaks. Peak heights are normalized to a scalar value ranging from 0 to 1 to provide a probability classification for each boundary: the higher the peak, the larger the local change in the spectral properties, and the more probable it is that a phone transition has occurred.

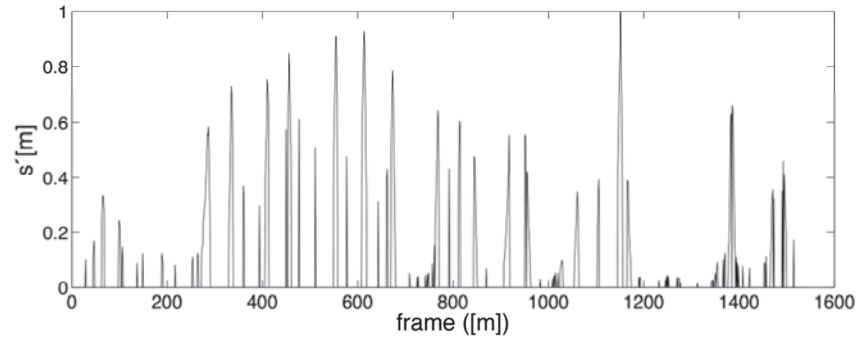

Fig. 3. s'[m] generated by minmax-filtering of s[m].

Another special operation that mimics a form of temporal masking is applied to the representation s'[m] to ensure that only the most prominent points of change are reported. For example, in the case of long spectral transitions between two adjacent phones, or due to non-correlating noise, several peaks may appear very close to one other. The inclusion of multiple points of change from several nearby frames is prevented by the following procedure: the distance between each peak in s'[m] that crosses a manually chosen threshold level p_{min} is calculated. If two or more peaks are closer than t_d to each other, the probability ratings of the peaks are compared. Only the most probable (highest) peak is retained, while its location is slightly adjusted towards the removed peak(s). The new location is situated between the old peaks and directly proportional to the ratio of probability ratings of the peaks in the region. As a result, a further refined s_r[m] is obtained.
In theory, a list of detected segments can now be created by choosing all the peaks that exceed the minimum peak probability threshold p_{min}. In practice however, this leads to splitting of the silent or quiet sections of the signal into several small segments. This can be avoided by comparing the energy of the original signal at each peak location to a minimum energy threshold e_{min} before a final decision is made. In terms of different energy thresholding mechanisms that were studied, the optimal results were obtained by using the mean energy value from –8 ms to +30 ms around the estimated boundary location for comparison to a fixed threshold, which was set to +6 dB from the minimum signal level. This asymmetry resembles the temporal masking effect present in hearing, in which effective backward masking is limited to approximately –10 ms whereas forward masking extends to a much longer time period (see page 78 in Zwicker & Fastl, 1999). All peaks exceeding the silence threshold are used as segmentation output. Figure 4 shows a schematic overview of the algorithm.

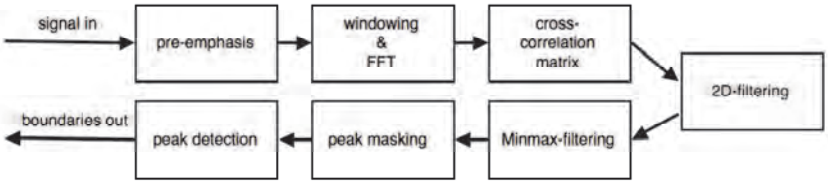

Fig. 4. Block diagram of the segmentation algorithm showing subsequent processing steps.

3. Experiments

The aim of the experiments was to obtain a good understanding of the overall performance of the algorithm so that it could be compared to earlier results found in other publications related to blind segmentation. Furthermore, determining the general effects of different parameters on segmentation results was desired. The results are presented for both genders separately in order to analyze whether gender specific differences exist, and a comparison of the obtained results to those found in existing literature is made. Additionally, noise robustness is evaluated. These results, with a brief analysis of the underlying statistics, will be covered in this section.

3.1 Evaluation measures

In order to evaluate segmentation quality, it is necessary to have a reference to which the output of the algorithm is compared. Since many well-known speech corpora are provided with a manual annotation, including TIMIT and our in-house Finnish speech corpus, a comparison to annotated segment boundaries was chosen as the primary evaluation metric. While manual segmentation is prone to the variability present in individual judgments, it is often considered as a reliable baseline for quality if it is carefully produced (Wesenick & Kipp, 1996). In addition, manual inspection of the segmentation output was performed in several phases of development and testing, yielding a more detailed insight into the phonetic details of the underlying signal in relation to the behavior of the algorithm.

A standard way to measure hits and misses in the literature is to detect whether the segmentation algorithm produces a segment boundary within a ±20 ms window (*search region*) centered around each reference boundary (Almpanidis & Kotropoulos, 2008; Aversano et al., 2001; Estevan et al., 2007; Kim & Conkie, 2002; Sarkar & Sreenivas, 2005; Scharenborg et al., 2007; Sjölander, 2003). If overlapping search regions exist, that is, adjacent regions with their reference boundaries are closer than 40 ms to each other, then the regions are asymmetrically shrunk to divide the space between two reference boundaries into two equal-width halves (see Räsänen et al., 2009). This will prevent ambiguous situations associated with overlapping search regions. Now each region can be searched for algorithmically generated boundaries: a boundary within a search region is considered as a *hit* and all additional boundaries within the same search region are counted as *insertions*. Empty regions are the source of *deletions* (or *misses*). Using this approach, the total number of hits N_{hit}, detected boundaries N_f, and reference boundaries N_{ref} are computed over the entire test material in order to derive the measures defined in table 1.

Overall segmentation accuracy is defined in terms of hit rate (HR). For some finite section of speech let N_{hit} be the number of boundaries correctly detected and N_{ref} be the total number of boundaries in the reference. HR can then be calculated using equation 6 in table 1 (Aversano et al., 2001). HR is inversely proportional to the miss (or error) rate, which is also sometimes used to indicate segmentation accuracy. Another central measure, especially in the case of blind methods, is the over-segmentation (OS) rate (7), which can be obtained if the total number of algorithmically produced boundaries N_f is included in the analysis (Petek et al., 1996). Different authors have used varying symbols for the above measures, originating from, e.g., signal detection theory. However, they have been found non-descriptive and are therefore replaced in this work by the new symbols *HR* and *OS*.

$HR = \frac{N_{hit}}{N_{ref}} * 100$ (6)	$OS = (\frac{N_f}{N_{ref}} - 1) * 100$ (7)
$PRC = \frac{N_{hit}}{N_f}$ (8)	$RCL = \frac{N_{hit}}{N_{ref}}$ (9)
$F = \frac{2.0 * PRC * RCL}{PRC + RCL}$ (10)	

Table 1. Standard quality measures used to evaluate segmentation

Precision (8) describes the likelihood of how often the algorithm identifies a correct boundary whenever a boundary is detected. Recall (9) is the same as HR (6) but without scaling to a percentage. In order to describe the overall quality of the segmentation with a single scalar between 0 and 1, the F-value can be computed from precision and recall (Ajmera et al., 2004). However, it has been shown that the F-value is not sensitive to so-called stochastic over-segmentation, where the hit rate of the algorithm can be increased by allowing higher levels of over-segmentation while the algorithm is actually producing new boundaries at random locations without any true reference to the underlying signal (Räsänen et al., 2009). A quality measure called R-value has been proposed to overcome this problem (Räsänen et al., 2009), and was therefore utilized in the evaluation process as a main criterion of quality, although the other quality measures are also reported for comparison. The R-value measures the distance between the current point of operation and the ideal performance (100% HR, 0% OS) in the HR/OS-plane (12), and the distance between the current point of operation and the case where the number of insertions is zero (12). These distances r_1 and r_2 are combined into a single scalar value between 0 and 1 according to (13), with unity indicating ideal performance.

$$r_1 = \sqrt{(100 - HR)^2 + (OS)^2} \tag{11}$$

$$r_2 = \frac{-OS + HR - 100}{\sqrt{2}} \tag{12}$$

$$R = 1 - \frac{abs(r_1) + abs(r_2)}{200} \tag{13}$$

Some authors also compute insertion rates (Cherniz et al., 2007) or ROC curves based on the ratio of insertions and total number of frames in the system (Esposito & Aversano, 2005). However, we find this type of methodology problematic since the number of frames is directly affected by the window step size, whereas the number of insertions and hits are not greatly affected since the temporal parameters (e.g., masking distance) are defined in temporal units (seconds) instead of number of frames. For example, changing the step size from 2 ms to 1 ms would basically halve the number of insertions per frame, providing very little information about the performance of the algorithm itself.

3.2 Material

The segmentation algorithm was tested on clean speech using the TIMIT speech corpus covering several American-English dialects. Additionally, a set of experiments was conducted using Finnish speech from a smaller and speaker-limited in-house corpus to detect possible language dependencies. The Finnish speech consisted of two male speakers each uttering 81 sentences of read speech, each sentence containing 28 phones on average. The sentences had been phonetically designed so that all of the naturally occurring diphones in Finnish were covered. A single phonetician then carefully segmented and labeled this material manually to produce about 4500 phones in total as well as 1680 segments (e.g., closures and releases indicated separately).

3.3 Results

Table 2 contains the evaluation results for the TIMIT test set using settings that provide optimal performance in terms of R-value (see section 3.4 for parameter dependencies). The

full test set (560 female and 1120 male sentences) was used, containing utterances from a total of 168 different speakers. A hit rate of 71.9% with -6.9% over-segmentation was obtained as a mean for both genders. The results also show that the results from both genders are nearly similar, the performance on female data being slightly higher (table 2).

gender	HR (%)	OS (%)	F-value	R-value
female	72.84	-7.9	0.78	0.79
male	71.37	-6.4	0.76	0.77
male+female	71.9	-6.9	0.76	0.78

Table 2. Segmentation results for the TIMIT test set.

The reader should note that by accepting higher values of over-segmentation (something that is not always desirable), higher hit rates are possible. The most straightforward manner to increase the over-segmentation level of the described algorithm is to adjust the length of the minmax-filter and the probability threshold p_{min} of the peak detector. Table 3 shows the results for the entire test set of TIMIT at an over-segmentation level of 54.3%. Although the overall HR has now increased notably, a large degradation of the R-value (and a relatively smaller degradation of the F-value) reflects the fact that this is simply due to an extremely high number of produced segment boundaries that start to hit search regions by chance.

gender	HR (%)	OS (%)	F-value	R-value
male+female	85.5	54.3	0.69	0.48

Table 3. Segmentation results for the TIMIT test set at a higher level of over-segmentation (male and female combined).

In general, the obtained results are well in line with the other results reported in literature regarding blind segmentation algorithms (table 4). More importantly, it seems that different blind algorithms achieve very similar levels of accuracy in terms of F- and R-values despite their methodological differences. The algorithm by Estevan et al. (2007) seems to obtain the highest R-values, but since we did not implement all of the algorithms shown in the table, it is impossible to conclude anything due to the fact that the differences in accuracy are of the same scale as the possible deviations in quality measures caused by ambiguities in evaluation methods (see Räsänen et al., 2009). The similarity of results is a topic that shall be returned to in the discussion section.

For the Finnish in-house corpus, the speech of two male speakers was automatically segmented independently to gain insight to both a) single speaker dependency, and b) the difference between rather swiftly spoken English material compared to very carefully articulated Finnish speech. The algorithm achieved 73.1% and 74.0% hit rates with over-segmentation values of 1.4% and -1.4% (F = 0.73, R = 0.77, and F = 0.75, R = 0.78, respectively) for the two Finnish speakers using the same parameters as in the TIMIT tests. These findings support the language and gender independency supposition of the algorithm and verify that excessive parameter tweaking is not necessary between languages.

Algorithm	HR (%)	OS (%)	F-value	R-value
Räsänen et al. (2009, this paper)	71.9	-6.90	0.76	0.78
Almpanidis and Kotropoulos (2008)	80.72	11.31	0.76	0.78
Aversano et al. (2001)	73.58	0.00	0.74	0.77
Esposito and Aversano (2005)	79.30	9.00	0.76	0.78
Estevan et al. (2007)	76.00	0.00	0.76	0.80

Table 4. Blind segmentation results on TIMIT from different authors.

3.4 Parameter dependency

In order to determine the impact of each parameter on overall performance in the described algorithm, parameters were adjusted and tested independently. Data used in the experiments were a randomly chosen subset of the TIMIT test set (N = 200 utterances), a set size considered sufficiently large to describe the behavior of the quality measures as a function of the parameter values. The most important parameters controlling the algorithm's behavior were the length n_{mm} of the minmax-filter, the peak masking distance t_d, and the boundary probability threshold p_{min}.

First it was verified that the FFT window length of 96 samples leads to the best performance (this corresponds to 6 ms at a 16 kHz sampling rate). As the purpose was to perform an FFT-analysis in which the window location regularly matches the location of the maximum energy of pitch periods (see section 2), this 6 ms window approximately satisfies the condition for both male and female speakers. Since the performance degraded for smaller and larger window sizes, the window length was fixed to 6 ms for the remaining parameter experiments.

During the development of the algorithm it was observed that the length n_{mm} of the minmax-filter, the threshold p_{min}, and the masking distance t_{d} of the final peak selector, were the most dominating parameters in the performance of the algorithm. As for n_{mm}, the value is mainly a tradeoff between over-segmentation and hit-rate, where approximately $n_{\mathrm{mm}} = 34$ frames (68 ms) was used in most of the tests to produce approximately OS = 0% for the entire TIMIT test material (note that the parameter experiments were performed with a subset of the test section and led to slightly different results due to a reduced set size). In the experiments it was observed that while the length n_{mm} controls the tradeoff between OS and HR, the F- and R-values are not greatly affected by these changes when OS levels are low. On the contrary, the peak selection threshold value p_{min} has a more dramatic effect on the F value. This is an expected result since it resembles the probability threshold for boundary detection: as more probable peaks are chosen, the obtained precision improves. However, when using higher values of p_{min} the algorithm starts to miss less probable boundaries (in terms of the algorithm), decreasing the recall.

For masking distance t_{d}, an optimal point can be found in the proximity of $t_{\mathrm{d}} = 25$ ms. This is a reasonable result since the rate of articulation in normal speech rarely exceeds four phones per 100 ms. There are still, e.g., some very short plosives that may exhibit bursts shorter than 20 ms, resulting in a decreased HR with longer masking distances than burst durations. On the other hand, by using values of tens of milliseconds, segmenting longer bursts into several small segments is avoided since the cross-correlation of the spectral coefficients may vary considerably within such variable transitions.

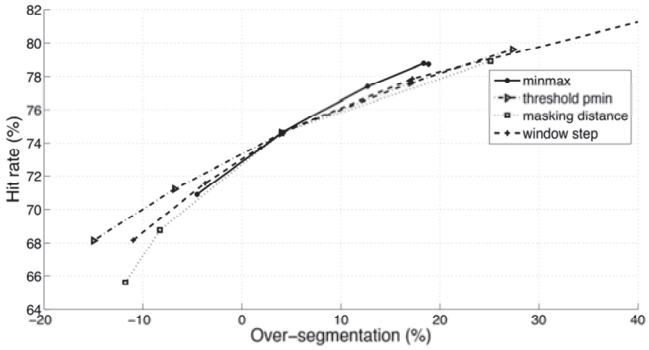

Fig. 5. Effects of different parameter values on segmentation results tested independently of each other. Parameter ranges are n_{mm} = 20-100 ms, t_d = 5-45 ms, ws = 1-3 ms and p_{min} = 0.02-0.1. Adjustment of the values changes the trade-off balance between hit-rate and over-segmentation, but the slope decreases as the value of over-segmentation increases.

To summarize, it was noted that most parameters control the tradeoff between over-segmentation and hit rate in a parallel fashion, while no parameter alone has a clear impact on improving the results (see fig. 5). Also, since many of the parameters are complementary, there are many possible combinations that achieve very similar results. Each value of choice for a parameter limits the maximum hit-rate by some amount in order to keep the over-segmentation at a reasonable level. It is possible to achieve much higher hit-rates by allowing over-segmentation to grow to very high values (see table 2). However, a large number of insertions is not usually desirable if the goal is to perform phonetic segmentation. It should be noted that once the parameters were set, the algorithm performed equally well for both genders and also for English and Finnish speech without any need for language specific optimization.

3.5 FFT versus MFCC in noise

While the FFT spectrum is a straightforward choice for use in algorithms for segment boundary detection, more popular alternative methods to describe spectral information also exist. One well-established choice in the field of speech processing is to use a parametric representation called Mel-frequency cepstral coefficients (MFCC) to obtain a simple auditory representation of the spectrum. To determine whether MFCCs enhance the performance of the segmentation algorithm when compared to the FFT, comparison tests were carried out. The first 20 static cepstral coefficients (ignoring the zeroth one) were chosen to represent the speech signal, since a further increase in their number did not yield any improvements.

Tests showed that the application of MFCCs to a 10 ms Hamming window with 2 ms steps led to optimal results in terms of windowing properties. Further increases in window size led to blurred temporal accuracy and therefore missed boundaries. Very similar results, as compared to the FFT, were obtained with noise-free signals, and led to values of HR = 74.7%, OS = 1.1% (F = 0.74, R = 0.78).

White noise and babble noise robustness of these two representations were tested with a subset of the TIMIT corpus by introducing additive white noise and babble noise to the original signals. The babble noise was generated from TIMIT data by summing together

speech signals from five different speakers speaking different utterances. Figure 6 displays the behavior of the R-value as a function of SNR. A decrease in SNR in the white noise condition leads to a small increase in the hit-rate with the FFT, but since this also starts to increase the over-segmentation level, the overall R-value drops dramatically. The hit-rate increase is explained as an increase in unintentional hits to the search regions due to increased OS (see Räsänen et al., 2009). MFCC segmentation preserves a much more conservative OS-rate at reasonable white noise levels when compared to the FFT.

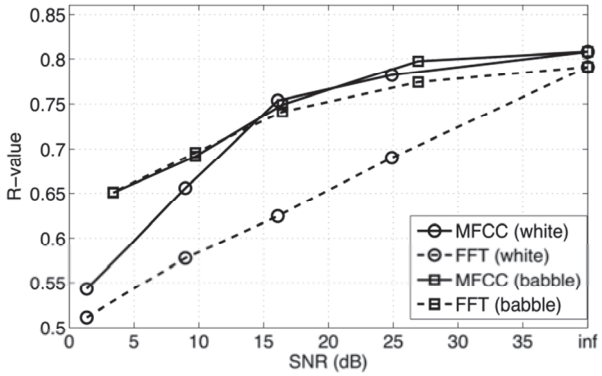

Fig. 6. The effects of white and babble noise on FFT and MFCC representations. FFT is shown with dashed lines and MFCC with solid lines. Circles denote white noise and squares babble noise.

In the case of babble noise, the difference between MFCC and FFT representations is very small. Over-segmentation at a near zero SNR level is more than 10% lower with babble noise when compared to the white noise situation, yielding much higher R-values. This is slightly surprising, since babble noise has its energy and spectral transients concentrated at the same frequency bands as the test signals.

The overall conclusion from comparing FFT and MFCC representations is that the difference is small, but MFCC seems to behave in a more stable manner especially when there is noise at the higher frequencies (e.g., white noise). This is due to the reduced spectral resolution of the MFCC's at the higher frequencies. With more natural babble noise, this difference is diminished.

4. Segmentation error analysis

4.1 Phone class-specific accuracies

Boundaries that automatic segmentation fails to detect are highly dependent on the underlying phonetic content. Some phone transitions are easy to detect due to sudden changes in the spectrum, whereas, e.g., glides and liquids may be more difficult to separate from their neighboring phones. In order to understand why and how the algorithm differs from manually produced references in the evaluated material, segmentation accuracy was estimated separately for each possible type of diphone transition defined in the reference annotation. Evaluation was performed using the FFT signal representation and TIMIT test set, yielding overall performances as reported in table 4. In order to capture an overview of

the performance and to reduce sparseness of diphone data in TIMIT, the 62 ARPABET phone classes used in TIMIT annotation were grouped into 7 larger phone classes according to Hasegawa-Johnson (2009).

		TO							
		Tense vowels	Lax vowels	Glides and liquids	Nasals	Fricatives	Stops and affricates	Closures	**Mean**
FROM	Tense vowels	48.6	25.4	44.5	85.2	94.8	N/A	65.5	60.7
	Lax vowels	80.0	17.1	37.0	82.4	89.7	N/A	76.3	63.8
	Glides and liquids	52.7	45.4	56.8	79.8	91.3	N/A	63.5	64.9
	Nasals	91.0	82.8	69.3	51.9	86.6	89.7	56.5	75.4
	Fricatives	87.8	82.1	88.4	90.5	68.1	N/A	83.7	83.4
	Stops and affricates	58.1	64.5	70.8	87.1	44.6	N/A	72.6	66.3
	Closures	45.1	34.7	58.2	73.8	77.3	80.3	55.6	60.7
	Mean	66.2	50.3	60.7	78.7	78.9	85.0	67.7	**68.8**

Table 5. Segmentation accuracy (%) for diphone transitions. Rows indicate the preceding phone while columns indicate the posterior phone of each pair. Pairs with less than 5 occurrences are excluded from the statistics.

As can be seen from table 5, there are extensive differences in accuracy between different diphone transitions. Especially problematic are across-class transitions between closures and vowels, vowels and glides, and stops and fricatives. This is understandable due to the spectral similarities of the phones in these pairs. Many sound classes also have very different segmentation accuracies depending on their relative position in the diphone. This is partly due to the fact that language specific structures impose constraints regarding which phones can precede or follow the current one. This yields different pre- and post-phone distributions for each single phone class, which is not seen in the table since it contains averaged results over entire phone groups. Another affecting factor is coarticulation that causes the segments to lose some of their spectral contrast.

Figure 7 shows histograms of segment output deviations from reference boundaries. This type of presentation reveals that transitions between spectrally contrasting segments lead to sharp distributions around, or near to, zero deviations, whereas similar speech sounds (e.g., transitions inside a phone group, the diagonals in figures and tables) have very broad distributions and low accuracies. Distributions of the majority of well-detected transitions are unimodal and fit well inside the ±20 ms time window used as an evaluation criterion.

The overall distribution of all correctly detected segment boundaries relative to the reference fits well with a normal distribution with a mean of zero and variance of approximately $\sigma_n^2 = 0.1^2$. This shows that approximately 35% of the boundaries would be located outside the search region if the deviation threshold was changed from 20 ms to 10 ms. This provides support for the convention of the ±20 ms deviation allowance that is typically found in literature (Almpanidis & Kotropoulos, 2008; Aversano et al., 2001; Estevan et al., 2007; Kim & Conkie, 2002; Sarkar & Sreenivas, 2005; Scharenborg et al., 2007; Sjölander, 2003), since the

algorithm reacts very systematically to changes in the signal in a time window of this size but rarely at larger distances.

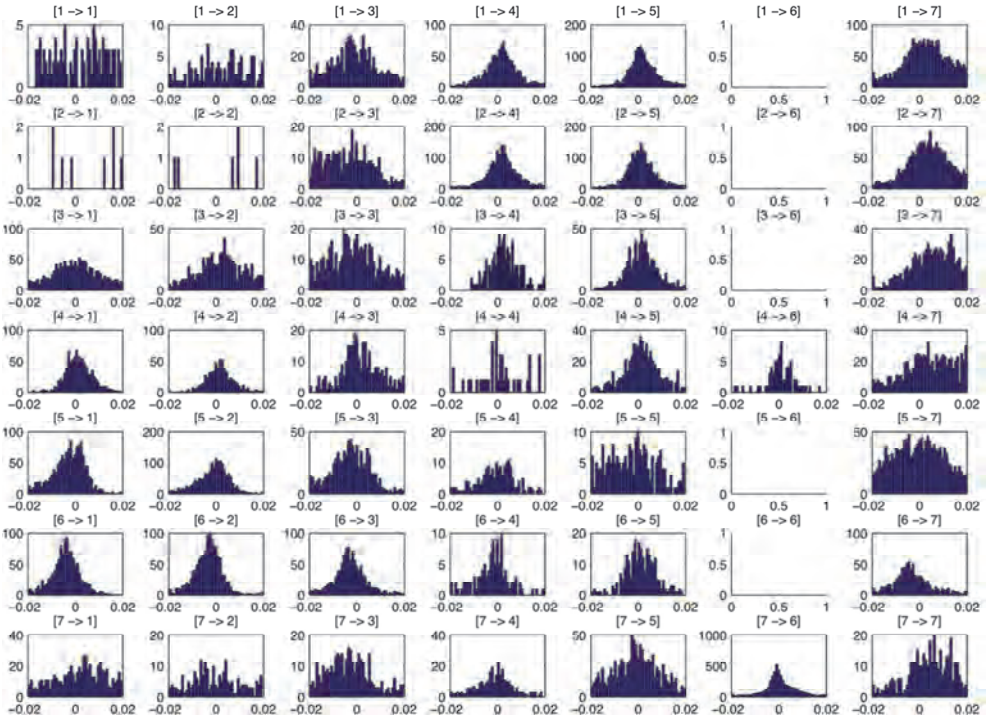

Fig. 7. Segmentation accuracy for phone classes found in Table 5 shown as temporal error distributions (seconds). Error is defined as the distance (in seconds) between produced segment boundaries and reference annotation (male + female speakers). 1: Tense vowels, 2: lax vowels, 3: glides and liquids, 4: nasals, 5: fricatives, 6: stops and affricates, 7: closures.

		To						
		Tense vowels	Lax vowels	Glides and liquids	Nasals	Fricatives	Stops and affricates	Closures
From	Tense vowels	4.2	2.1	-0.5	-3.6	0.2	N/A	-4.9
	Lax vowels	-25.0	-14.3	9.7	-1.8	-1.1	N/A	-2.4
	Glides and liquids	-0.5	0.4	-2.2	-10.7	-3.1	N/A	-7.4
	Nasals	-5.4	-7.5	4.6	11.3	1.4	-4.9	1.4
	Fricatives	-4.3	-1.1	-6.5	0.9	1.1	N/A	-3.4
	Stops and affricates	-9.5	-4.4	-1.4	3.1	-1.4	N/A	-1.8
	Closures	2.9	6.1	-1.5	-4.8	-5.6	2.3	-7.8

Table 6. Segmentation accuracy difference (%) between male and female speakers (positive value = male performance better, negative value = female performance better).

Accuracy differences for phone transitions between male and female speakers were also estimated using the FFT representation (table 6). The differences in accuracy show that some transitions (e.g., from lax vowels to tense vowels and between lax vowels) are significantly more accurately detected in female speech, whereas some others (e.g., nasal-to-nasal and lax vowel-to-glide) transitions are more readily detected in male speech. The reason for such differences is not clear, but they may arise from cross-gender differences in the anatomy of the vocal apparatus. The role of very short-term windowing in FFT may also have an impact, since the ratio of window length and one pitch period is different for the two genders.

Phone specific performance was also studied between the FFT and MFCC. It was determined that these two representations produce different results for some phone categories. The FFT segmentation performs especially well on fricatives, stops and affricates, whereas MFCC is more sensitive to vowels, glides and liquids. The FFT based segmentation seems to be much more accurate for the beginnings of stops and affricates (+14% compared to MFCC; e.g., [bcl]-[b]) whereas MFCC exhibits slightly more accuracy with post-phone transitions of the same phone classes (e.g., from [b] to [a]). These differences are somewhat expected, as the FFT has a high resolution also at the higher frequencies (fricatives and quick transitions, e.g., bursts) whereas Mel-filtering weights the low frequency range more. Despite the differences noted for different speech sound categories, both spectral representations end up exhibiting very similar results for overall segmentation accuracy (see section 3.3).

4.2 Inspection of problematic segments

As the detection of some vowel transitions is problematic for the algorithm, further studies were made to gain a deeper insight into these cases. Figure 8 illustrates an example of why it may not be possible to achieve extremely high accuracies with bottom-up approaches in general. In this example the word "*water*" is spoken by a female speaker: the time waveform is shown in the top pane while the linear-frequency spectrogram is shown below. The manually determined boundaries for phone [ao]'s transitions are indicated by dashed lines.

The segmentation algorithm is able to detect the [ao]-[dx] transition while the [w]-[ao] transition remains undetected, causing a deletion to be registered. There is no noticeable change in the spectrum, waveform, pitch, or even in signal energy, so the only possible way to place a boundary at such a location would be based on perceptual judgment. An automatic algorithm using such features, and working in a bottom-up manner, probably cannot detect such types of changes in speech.

There are also onsets of phones that do not contain sudden spectral changes but their waveform shape changes radically when compared to that of their neighbors. One such phone that is especially difficult for the present algorithm to detect is the pharyngeal fricative [q], which often contains a similar formant structure to the preceding vowel but where pitch and signal energy suddenly drop causing a perceptually creaky voice. These changes can be seen in the waveform as areas of significantly decreased amplitude and shifted phase. One example of this situation can be seen in figure 9 where a transition is occurring at the end of the word "*misquote*" and leading into "*was*". These types of deletions could be avoided by including a supplementary module with the algorithm that could track, e.g., changes in the waveform shape, pitch, or phase of the speech signal.

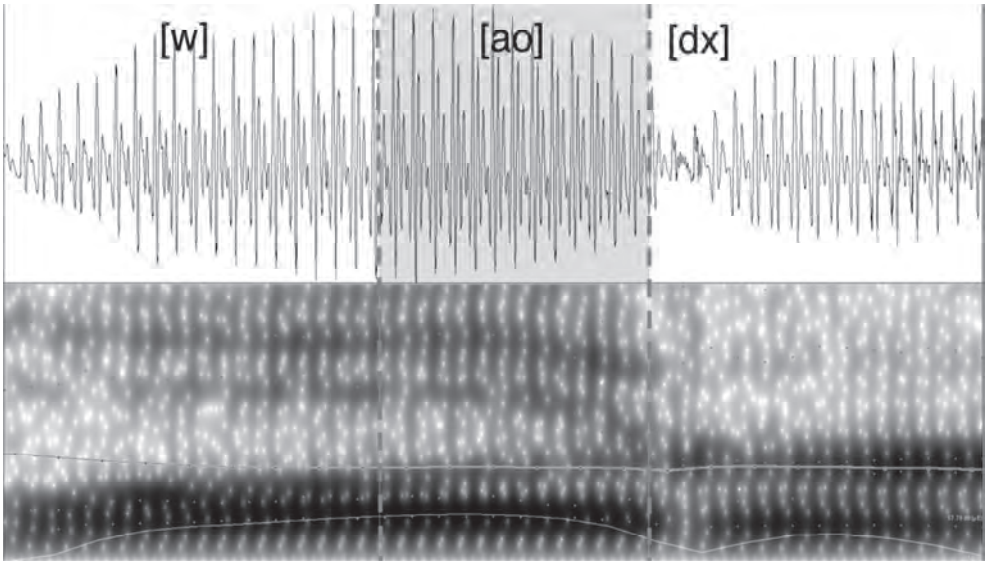

Fig. 8. A partial waveform for the word *"water"* spoken by a female speaker as well as a related spectral representation that includes F0 (upper line) and energy contours (lower line). Dashed lines indicate reference phone boundaries. The [w]-[ao] transition boundary is practically impossible to detect with the bottom-up segmentation algorithm described in this paper due to lack of changes in the feature space. Images were created using Praat software.

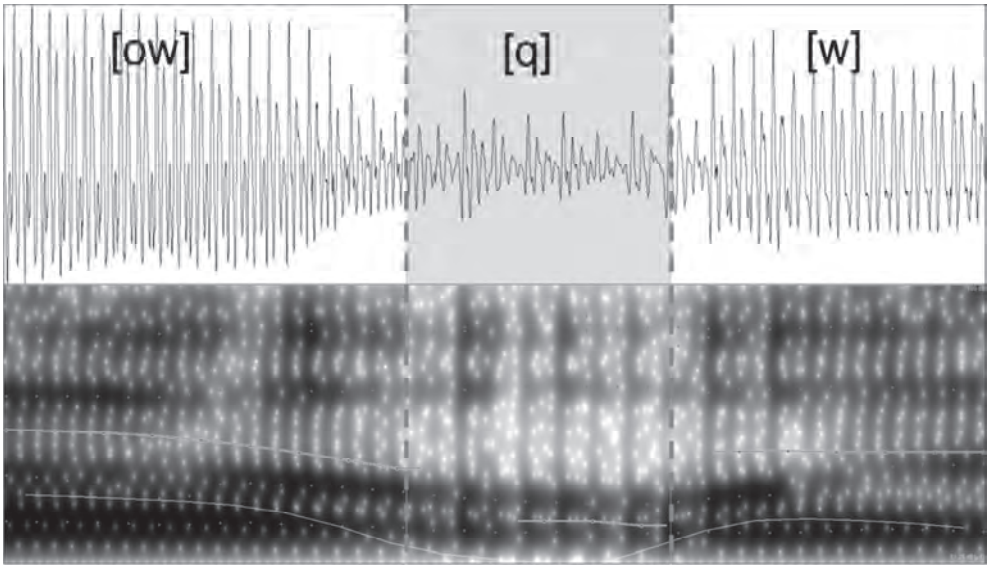

Fig. 9. The transition from [ow] to pharyngeal fricative [q] at the end of the word *"misquote"* and also from [q] to [w] in the beginning of word *"water"* are difficult to detect using spectral analysis, while changes in waveform shape are easily perceived visually.

Another general characteristic difference between the algorithm's output and the reference annotations can be found at the endings of speech signals: it is often difficult to determine where the final phone ends, and very often the perceptual ending (and annotated boundary) takes place earlier, while the spectrum of the breathy ending keeps fading away for a moment longer. As the algorithm reacts most prominently to the point where there is a structural discontinuity point in the spectrum (i.e., the signal changes from a correlating formant structure to a silence), it places a boundary where the spectrum of the exhalation finally fades to a non-existent level. This effect was observed with both English and Finnish data.

The implicit assumption underlying this work is that "optimal" automatic segmentation of continuous speech should lead to results where preferably only one phone occupies one segment. However, there seems to be a large number of cases where effective segmentation of continuous speech to phonetic units is difficult using blind bottom-up approaches. For some transitions, the changes in the features representing the signal may be gradual (e.g., in diphthongs) or almost non-existent (fig. 8), although a human listener still perceives a change from one articulatory position to another due to learned distinctions. In some other cases, like at the endings of the signals, the points of change simply cannot be unanimously defined. Real speech also contains situations where phones are spectrally split into two or more "subphones". This occurs, e.g., when an oral vowel is nasalized or a nasalized vowel is "oralized" causing rapid spectral changes to occur at first formant as well as nasal formant locations. Another example of this type of splitting is a liquid or a fricative situated between front and back vowels or some other changing phonetic context. This type of phenomena may cause the first part of such a segment to differ considerably from its remainder.

Thus, the implicit assumption behind the chosen segmentation methodology and the preferred goal is partially conflicting with the natural operation of articulatory mechanisms. Spectral change alone is not a sufficient cue for phone segment boundaries since some intra-segmental changes can be larger than some transitions from one phone class to another. This leads to an inevitable tradeoff between segmentation accuracy and over-segmentation. If more comprehensive blind phone-segmentation is required then problematic cases should be studied in more detail in order to handle them in a correct and language-universal manner. This question is left as a topic for further studies.

5. Conclusions

This paper introduced a novel blind speech segmentation algorithm that utilizes the cross-correlations of adjacent spectral representations of the signal. Local changes in the spectrum are detected using a two-dimensional filter on the cross-correlation matrix. Output from the filter is then reduced using a non-linear minmax-filtering technique, and finally a temporal masking operation is applied to the detected signal changes. The results obtained by this algorithm are comparable to those found in literature (Almpanidis & Kotropoulos, 2008; Aversano et al., 2001; Esposito & Aversano, 2005; Estevan et al., 2007; Scharenborg et al., 2007). The performed experiments also give support for the language and gender independency of the algorithm, although further evaluation on several other languages would be required to confirm this.

Experiments from several authors seem to indicate that a maximum level of segmentation accuracy with a purely bottom-up approach is already being achieved and falls below available HMM-solutions in terms of reference evaluation. The results reported by Almpanidis and Kotropoulos (2008), Aversano et al. (2001), Esposito and Aversano (2005), and Estevan et al. (2007) all produce very similar results for the TIMIT corpus material while using totally different approaches for phone segmentation - a striking discovery already noted briefly by Estevan et al. (2007). Interestingly enough, the algorithm introduced in this paper also achieves a very similar level of accuracy with yet another methodological approach. The observed asymptotic behavior from these five different methods may indicate that further improvements may not be possible without introducing linguistic or contextual knowledge, even when working in noise-free conditions. Analyzing the instantaneous properties of speech signals systematically falls short of ideal performance.

More evidence for the suggested accuracy *'limit'* existing in the bottom-up approaches can be found by analyzing the results of Cherniz et al. (2007), who attempted to improve the algorithm presented by Esposito & Aversano (2005) by replacing the original Melbank signal representation with continuous multiresolution entropy (CME) and continuous multiresolution divergence (CMD). Although the use of CMD had a statistically significant effect by lowering the number of insertions (from OS = 16.61% to OS = 13.87%), the number of detected boundaries did not change significantly ($Pr(\varepsilon < \varepsilon_{ref}) > 80.57\%$) despite employing totally different parametric representations. Similarly, here we have studied the use of FFT and MFCC in the blind segmentation task and showed that already the simple short-time FFT leads to comparable segmentation accuracy with the MFCCs (R = 0.78). One may ask whether part of the observed inaccuracies would result from the variability of the underlying reference annotation. However, the role of manual biases in overall performance should be small if ±20 ms search regions are used for evaluation (see Wesenick & Kipp, 1996, for reliability of manual

transcriptions). The boundary deviation distributions obtained in this study also support the suitability of the standard ±20 ms search regions used in evaluation.
Based on the given evidence and work already performed in the field of blind segmentation, we hypothesize that it is extremely difficult to construct a blind algorithm that analyzes the local properties of speech with universal decision parameters that could achieve notably higher segmentation accuracies than those already developed and reported in the cited literature and in this paper. In practice this would mean that grossly 70-80% of phone boundaries can be automatically and reliably detected and pinpointed in time by tracking changes in spectrotemporal features extracted from speech. The remaining 20-30% seem to be defined by changes that are too small to be detected unless the system really knows what type of signal changes it should look for in a given context. This may be the price that has to be paid with algorithms that do not learn from data or utilize expert knowledge from proficient language users.
Finally, it should also be kept in mind that perfectly matching reference boundaries is not (always) the ultimate goal of speech segmentation. In the end, the purpose of the segmentation algorithm depends on the entire speech processing system in which it is implemented, and the most important evaluation method would be then to observe and measure the functionality of the system in its entirety.

6. Acknowledgements

This research was conducted as part of the work in the Acquisition of Communication and Recognition Skills (ACORNS) project, funded by the Future and Emerging Technologies, in the Information Society Technologies thematic priority in the 6th Framework Programme of the European Union. The authors wish to thank Prof. Lou Boves from the Language and Speech Unit of Radboud University and Prof. Paavo Alku from the Dept. of Signal Processing and Acoustics of Helsinki University of Technology for giving valuable comments on this work.

7. References

Ajmera, J., McCowan, I., & Bourlard, H. (2004). Robust Speaker Change Detection. *IEEE Signal Processing Letters*, Vol. 11, No. 8, pp. 649-651

Almpanidis, G., & Kotropoulos, C. (2008). Phonemic segmentation using the generalized Gamma distribution and small sample Bayesian information criterion. *Speech Communication*, Vol. 50, pp. 38-55

Antal, M. (2004). Speaker Independent Phoneme Classification in Continuous Speech. *Studia Univ. Babes-Bolyal, Informatica*, Vol. 49, No. 2, pp. 55-64

Aversano, G., Esposito, A., Esposito, A., & Marinaro, M. (2001). A New Text-Independent Method for Phoneme Segmentation, *Proceedings of the IEEE international Workshop on Circuits and Systems*, Dayton, Ohio, USA, August, 2001

Cherniz, A.S., Torres, M.E., Rufiner, H.L., & Esposito A. (2007). Multiresolution Analysis Applied to Text-Independent Phone Segmentation. *Journal of Physics: Conference Series*, Vol. 90, pp. 1-7

Esposito, A., & Aversano, G. (2005). Text Independent Methods for Speech Segmentation, In: *Lecture Notes in Computer Science: Nonlinear Speech Modeling,* Chollet G. et al. (Eds.), pp. 261-290, Springer Verlag, Berlin Heidelberg

Estevan, Y.P., Wan, V., & Scharenborg, O. (2007). Finding Maximum Margin Segments in Speech, *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, Honolulu, Hawaii, USA, April, 2007

Hasegawa-Johnson, M. (2005). Phonetic and Prosodic Transcription Codes, In: *Lecture Notes in Speech Recognition Tools*, Page accessed in Dec. 15th, 2010, Available from: http://www.isle.illinois.edu/sst/courses/minicourse/2005/transcriptions.pdf

Hemert, J.P. (1991). Automatic Segmentation of Speech. *IEEE Trans. Signal Processing*, Vol. 39, No. 4, pp. 1008-1012

Kim, Y.-J., & Conkie, A. (2002). Automatic segmentation combining an HMM-based approach and spectral boundary correction, *Proceedings of International Conference on Spoken Language Processing (ICSLP'02)*, Denver, Colorado, September, 2002

Makhoul, J., & Schwartz, R. (1994). State of the Art in Continuous Speech Recognition, In: *Voice Communication Between Humans and Machines*, D.B. Roe & J.G. Wilpon (Eds.), pp. 165-198, National Academy Press, Washington D.C.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am.*, Vol. 58, No. 4, pp. 880-883

Nossair, Z.B., Silsbee, P.L., & Zahorian, S.A. (1995). Signal Modeling Enhancements for Automatic Speech Recognition, *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, Detroit, Michigan, USA, May 1995

Petek, B., Andersen, O., & Dalsgaard, P. (1996). On the Robust Automatic Segmentation of Spontaneous Speech, *Proceedings of International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, October, 1996

Räsänen, O., Laine, U.K., & Altosaar, T. (2008). Computational language acquisition by statistical bottom-up processing, *In Proceedings of 9th Annual Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, September, 2008

Räsänen, O., Laine, U.K., & Altosaar, T. (2009). An Improved Speech Segmentation Quality Measure: the R-value, *Proceedings of 10th Annual Conference of the International Speech Communication Association (Interspeech '09)*, Brighton, England, September, 2009

Räsänen, O., & Driesen, J. (2009). A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition, *Proceedings of 17th Nordic Conference on Computational Linguistics (NODALIDA)*, Odense, Denmark, May, 2009

Sarkar, A., & Sreenivas, T.V. (2005). Automatic speech segmentation using average level crossing rate information, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, March, 2005

Scharenborg, O., Ernestus, M., & Wan, V. (2007). Segmentation of speech: Child's play? *Proceedings of 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, Antwerp, Belgium, August, 2007

Sharma, M., & Mammone, R. (1996). 'Blind' speech segmentation: automatic segmentation of speech without linguistic knowledge, *Proceedings of International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, October, 1996

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech, *Proceedings of Fonetik 2003, the XVI Swedish Phonetics Conference 9*, Lövånger, Sweden, June, 2003

Wesenick, M.-B., & Kipp, A. (1996). Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals, *Proceedings of International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, October, 1996

Zhang, T., & Kuo, C.-C.J. (1999). Hierarchical classification of audio data for archiving and retrieving, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, Phoenix, Arizona, March, 1999

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and Models* (2nd ed.), Springer Series in Information Sciences, Springer, Berlin