Presented at IJCNN 2015, Killarney, Ireland, 12-17 July 2015.

A biologically inspired onset and offset speech segmentation approach

Andrew K. Abel, Dean Hunter, Leslie S. Smith
Computing Science and Mathematics
University of Stirling, Stirling FK9 4LA, Scotland
Email: {aka, dhu, lss}@cs.stir.ac.uk

***Abstract*—A key component in the processing of speech is the division of longer input sounds into a number of smaller sections. For speech interpretation it is generally easier to classify single sections. Similarly, when processing speech for other purposes (e.g. speech filtering), it can be easier and more relevant to process individual phonemes. Here, we propose a biologically inspired speech segmentation technique that filters the speech into multiple bandpassed channels using a Gammatone filterbank, and then uses an essentially energy-based spike coding technique in order to find the onsets and offsets present in an audio signal. These onsets and offsets are then processed using leaky integrate-and-fire neurons, and the spikes from these used to determine the speech segmentation. We evaluate this new system using a quantitative evaluation metric, and the promising results of segmentation of both clean speech and speech in noise demonstrate the effectiveness of this technique.**

I. Introduction

Sound is frequently a highly dynamic signal, and often the sound produced by a sound source has clearly defined sections within it. This is certainly true for speech (where one often thinks of syllables), and for many types of birdsong, as well as for melodies played on a single note instrument. Other sounds (the rushing of wind in trees, or the sound of flowing water, or indeed, many sounds produced by continuously running machinery) do not have this characteristic: yet often the sounds that need the most interpretation do have clearly defined relatively short segments. Segmentation is subjective as there can be debate over the appropriate length of a segment, depending both on the nature of the sound, and on the purpose of the segmentation. For example, there are scenarios where the appropriate segmentation is a complete utterance, and others where speech should be divided into phonemes for further processing. Here we consider speech segments to be closer to phonemes.

Speech segmentation can be a key enabler for further processing. For example, in speech filtering, the audiovisual work of Almajai and Milner [1] processes individual phonemes of speech with different models, requiring phoneme level segmentation. For interpretation, segmentation may be required at sentence, word, syllable [2], or phoneme level [3], [4] in order to correctly recognise speech: clearly word and sentence boundaries are unlikely to be findable directly from the signal. Music segmentation is also required for further processing [5]: for music, segmentation may produce individual notes, or entire musical phrases. It is not surprising, then, that sound segmentation has been of interest for a long time.

Many different segmentation techniques have been used, for example [2], [4], [6], [7]. Techniques applied to speech can be divided into two classes: blind techniques, which do not use a phonetic classification of the utterance, and non-blind techniques which do use this information. A similar distinction could be made for musical segmentation techniques, however, most of these are blind systems. Within the blind techniques, many different techniques have been applied, ranging from those based on onsets (like this one), to ones based on wavelets.

Here, we propose a biologically inspired speech segmentation, that uses onsets and offsets to identify the starts and ends of segments. It is a considerable extension of preliminary work developed by one of the authors [4], and takes into account the idea of using multiple resolutions suggested in many different works in image segmentation. In addition to the frequency bands of speech being considered in a logarithmic way (following convention, as well as the human auditory system [8]), we consider amplitude levels logarithmically. The onset (and offset) transform used (described below) is essentially a wavelet transform (but stopping half-way through the convolving function so that it remains applicable in real-time), and the actual decision about the location of the segments is partly implemented using a leaky integrate-and-fire neuron to combine onset (offset) information both at differing resolutions and across bands.

This speech segmentation system is designed to function in a causal manner, that is, rather than identifying onsets and offsets after the sound has been processed, it makes use only of data up to the current instant. This means that it does not require the whole utterance in order to process, and so has more potential for application to real time processing.

The system presented is evaluated with speech, and because speech segmentation here is close to the phoneme level, we can compare segmentation performance with annotated corpora. In this work, we have utilised the transcribed sentences of the TIMIT Corpus [9]. This provides the data needed for evaluation, and the evaluation process makes use of an approach first proposed by Hu and Wang [6] for speech segmentation. This considers how well a generated segment is considered to match a pre-defined, transcribed segment, and generates scores. This process has been used to investigate the segmentation of speech both with and without noise.

Results show that the system performs well both in clean speech, and with noise (from a variety of artificially generated and natural sounds) added. The system tends towards identifying more segments than identified with human transcription, suggesting that it is very sensitive to small changes. However, it

does not perform well in extremely noisy environments (SNRs of -20dB and below), as would be expected.

The remainder of this paper is divided as follows. Section II presents a full description of the proposed onset and offset based segmentation system, with the detailed evaluation metric discussed in Section III. The results of testing both clean speech and speech mixed with a variety of noises are given in Section IV, and finally, Section V concludes the paper.

II. Segmentation System Description

The onset/offset based speech segmentation system applies several layers of processing. Initially, the incoming monaural digitised sound signal is filtered using the Gammatone filterbank [10] into 100 bands, with centre frequencies ranging from 50Hz to 7000Hz. The output from each filterbank is transformed into a set of 16 spike trains, which represent biologically inspired auditory nerve threshold crossing, inspired by the high and low spontaneous rate nerve fibres of the auditory nerve, as described in depth in [11]. These spike trains perform a logarithmic coding of the signal amplitude, while maintaining the fine time structure.

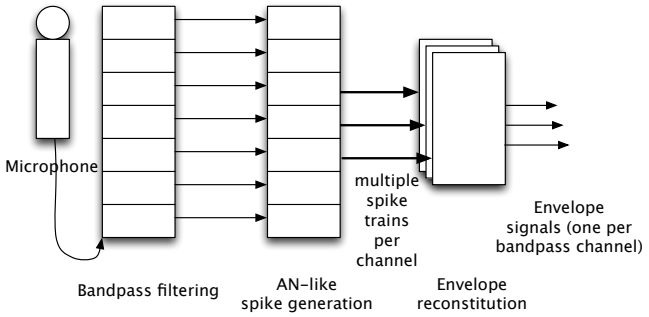

Fig. 1. Initial processing of sound. Each bandpass channel produces a number of spike trains (although they could be considered as a single height coded pulse train) providing a phase preserving amplitude coding of the bandpass signal output. The envelopes of each bandpass channel are reconstituted at a lower sampling rate from these.

These signals are then used to recode the signal at a lower sampling rate. The resulting envelope signals are all positive-going, with one per bandpass channel. They are an approximation to the logarithm of the envelope of each bandpass signal. Although they are low-pass filtered versions of the signal envelope, because of the way in which they are created, they do not incur the phase delay that is often associated with low-pass filtering. The low sampling rate makes processing them further relatively fast.

Following on from previous work [4], we applied a "half-difference of Gaussians" convolution to these signals. Logically, we should apply a difference of Gaussians to discover the beginning and end of the segment, but this would imply that one could not find them until some time after the segment had occurred. In order to be able to produce a causal system[1], we cut the transform in half, with the result that the transform output is available more rapidly. The convolving function used is

$$O_{i,j}(t,k,r_j) = \int_0^t (f(t-x,k_j) - f(t-x,k_j/r_j))b_i(x)dx \quad (1)$$

where $b_i(x)$ is the signal originating from bandpass channel i, and $f(x,y) = \sqrt{y}\exp(-x^2y)$. i indexes the channel number. In this work, a set of transforms are used (indexed by j in equation 1), resulting in a multi-scale set of convolved outputs: k was fixed at 1000 (so that the j index on the k can be dropped), with values of r_j varying geometrically from 1.1 to 3.5, and five transforms are used to allow for different duration resolutions in both the onsets and the offsets.

Each transform $O_{i,j}(t,k,r_j)$ is turned into two signals, an onset signal $s_{\mathrm{on}(i,j)}(t)$, and an offset signal $s_{\mathrm{off}(i,j)}(t)$ by

$$s_{\mathrm{on}(i,j)}(t) = \max(0, O_{i,j}(t,k,r_j)) \quad (2)$$
$$s_{\mathrm{off}(i,j)}(t) = \max(0, -O_{i,j}(t,k,r_j)) \quad (3)$$

The sum of the set of onset (offset) signals originating from each band is produced (summing over the different transforms, indexed by j), so that there is just one onset and one offset signal from each bandpass channel: see figure 2.

$$S_{\mathrm{on}(i)}(t) = \sum_j s_{\mathrm{on}(i,j)}(t) \quad (4)$$
$$S_{\mathrm{off}(i)}(t) = \sum_j s_{\mathrm{off}(i,j)}(t) \quad (5)$$

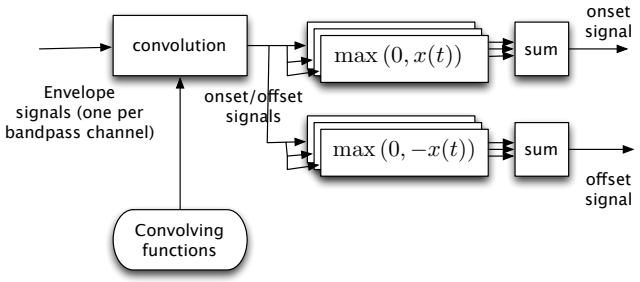

Fig. 2. Generation of the onset and offset signal. The convolving function used is a half-difference of Gaussians. $x(t)$ is $O_{i,j}(t,k,r_j)$. One onset signal and one offset signal is generated per bandpass channel.

The $S_{\mathrm{on}(i)}(t))$ and $S_{\mathrm{off}(i)}(t)$ are used as inputs to a pair of leaky integrate-and-fire (LIF) neurons ([12] Chapter 14) to create an onset and an offset spike train for each bandpass channel: see figure 3. To avoid issues where only a single channel spikes, a weighted sum across adjacent channels is used as the input to the LIF neuron:

$$A_i^{\mathrm{on}}(t) = \sum_{m=i-n}^{i+n} w_m S_{\mathrm{on}(i+m)}(t) \quad (6)$$
$$A_i^{\mathrm{off}}(t) = \sum_{m=i-n}^{i+n} w_m S_{\mathrm{off}(i+m)}(t) \quad (7)$$

The value of n defines the convergence: $2n+1$ bandpass channels affect each LIF neuron. If $n = 0$, then there is no convergence, and each channel is treated individually.

[1]Here, causal system implies one that can produce its output at a particular time using only the signals that have happened up to that time: this is a prerequisite for real-time operation.

Convergence allows summation of adjacent channels, reducing artefacts from noise.

The LIF neurons all have the same dissipation (time-constant), because the expectation is that segment lengths are independent of the location in the spectrum. All the LIF neurons also have a fixed refractory period (0.01 seconds), again because we disallow pairs of onsets (offsets) from occurring too close together. The output from all of these LIF

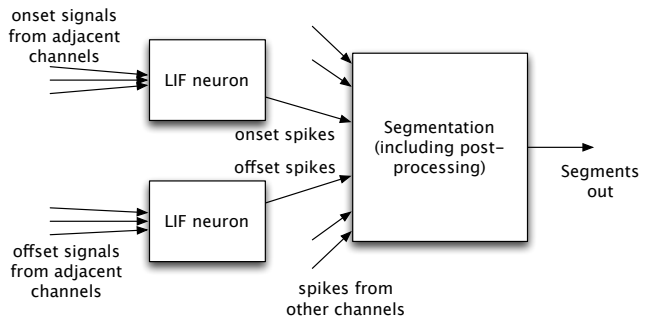

Fig. 3. Creation of the segments from the onset and offset signals. These are combined across a number of channels at each LIF neuron (there is one LIF neuron per channel), and then all the spikes are combined to produce the segmentation. Post-processing allows for removal of very short segments, separated from others by a very short time.

neurons is used as the input to a segmentation algorithm. This uses each onset spike to mark the beginning, and each offset spike to mark the end of a segment: segments are not allowed to overlap.

A segmentation is produced for each bandpass channel, and these are then combined to generate a final segmentation (alternatively, one might actually desire a 2-dimensional segmentation, across time and spectrum, rather than a 1-dimensional one: this would fit with Hu and Wang's concept of a tiling of the spectrotemporal space: figure 4 shows a spectrotemporal segmentation). The number of segments at each time point, $N_s(t)$, is calculated (this is between 0 and the number of bandpass channels). Each time that $N_s(t)$ starts to rise is marked as the beginning of a segment, and each time that $N_s(t)$ stops falling is marked as the end of a segment. This approach was taken because the different segments in each channel start and stop at about the same times, so that the centre of each segment is a maximum in $N_s(t)$. We use a post-processing stage that merges very short segments separated by very short intervals. Clearly more sophisticated methods for interpreting the spectrotemporal segmentation could be applied, and this is a topic of current research.

A. Discussion of methods used: choice of parameters

Because of the versatility and range of applications of this technique, there are a number of parameters that need to be set at each processing stage. Their precise optimal values are not always intuitive. Here, we discuss the parameters at each stage of the segmentation processing. We note that although divided into subsets for the purpose of reporting, the parameters interact with each other. For example, the values for the onset/offset convolution, and the parameters for the LIF neuron interact to determine the precise rates of change that matter for segmentation. The best values to use will depend on the application (speech from a fast talker, or one with a slow rate of speech, or single musical notes, or musical notes in a highly reverberant environment).

1) Digitisation and bandpass filtering: Digitisation depth was 16 bits. The bandpass filter used is the Gammatone filterbank [10]. The number of bands to use will depend on many factors: channel bandwidth, processing time available, frequency range, etc. Here, we have used 100 bands, from 50 to 7000 Hz. This gives a range of 7.13 octaves, i.e. about 14 bandpass channels/octave.

2) AN-like spike generation: As discussed in [11], a spike is produced for the most sensitive sensitivity level when the signal in the previous quarter cycle (as measured at the channel centre frequency) exceeds a specific value, K_0 prior to crossing 0. Signals at lesser sensitivity levels use a threshold of $R^i K_0$, for i from 0 (most sensitive level) to $N-1$ (least sensitive level), where $R > 1$. The parameters to adjust are the number of spike trains to produce per band (i.e. value of N), when the most sensitive spikes should be generated (value for K_0), and the difference in threshold between levels (value for R). These are not independent of each other, but define the level below which sound will be ignored, as well as the dynamic range of the system. We have used $K_0 = 0.0002$, $N = 16$, and $R = 1.414$. The value for K_0 will depend on the transduction system and coding used for sound; using $R = 1.414$, and considering the signal level as a voltage, means that the sensitivity levels are $10\log_{10}(R^2)$dB apart, which is approximately 3dB. Setting $N = 16$ provides approximately 48dB of dynamic range.

3) Logarithmic envelope generation: The spike train is turned into a level according to the index of the spike train. The primary parameter here is the resampling rate, we downsample in order to reduce the processing load during later stages of the segmentation system. Here, 400 samples/second have been used. This removes fine time structure, but for the purposes of identifying onsets and offsets, precise temporal structure information is not required (unlike the situation where one is considering computing the difference in time difference of arrival (TDOA), or even deciding exactly when to compute TDOA [13]).

4) Onset and offset convolution: As discussed, the values for k, r_j, and the number of convolving functions are important. Experimentation suggests that $k = 1000$, and r_j varying from 1.1 to 3.5, in 5 steps, works well. The exact values to use depend on the precise timing of the envelope variations that we take to constitute onsets and offsets. Given that onsets do tend to be faster than offsets (because of both the nature of sound production, and the effects of reverberation), there may be a case for using different values for the onsets than from the offsets.

The onset and offset signals are separated prior to being applied to their LIF neuron, so that different parameters can be associated with each. The parameters are:

The weight on the input to the neuron

This can either be a scalar (in which case the convolution outputs discussed above are all weighted equally), or a vector whose length is the same

as the number of convolving functions (in which case they are weighted differently). The value of this weight (or weight vector) directly affects the activation level in the LIF neuron. Note that the weight value will interact with the dissipation of the leaky integrate-and-fire neuron (see below).

The vector defining the convergence of the signal
This is used to allow signals from neighbouring channels to affect the LIF neuron's activity level.

To make these parameters independent, the vector used in this convolution across bands is normalised to sum to 1 (the vector elements are expected to be positive). Based on experimentation, we use weight values of 2.0 for the onset, and 2.75 for the offset. The vector for the convergence (or mixing) depends on a number of factors, including the nature of the sound of interest (is it very broadband, for example). Initial experiments identified that a short vector (length 7) provides better results, as it makes the initial segmentation in each channel more independent.

5) The leaky integrate-and-fire neurons: These have two parameters, dissipation, and refractory period. High values for dissipation mean that the activity leaks away rapidly. Thus the values used need to reflect the expected rate of change (or rather, the rate of change that matters in this application). Values of 8 for the onset LIF, and 9 for the offset LIF have been used. Higher values for the dissipation parameter require (generally) larger activation levels in order for the LIF neuron to produce spikes, and this interacts with the weight used at the input to the LIF neuron. The refractory period has been set to 0.01 seconds, although this has not been found to be critical.

III. Evaluation Metrics

Segmentation evaluation is complicated because one can argue that segmentation units should be words, or even sentences, though they are more usually syllables or phonemes. Assessment techniques are reviewed in [14]. Hu and Wang discuss this issue [6], and the limited range of evaluation utilised in related research. We adopt their technique, which focuses on spectral-temporal performance, based on evaluation metrics originally used in image segmentation [15]. Hu and Wang evaluate segmentation performance by considering overlapping regions between ideal and estimated segments, using spectral-temporal segmentation. We compare segments generated by the system to those identified by human transcription, that is, the phoneme based transcription data provided with the TIMIT corpus.

A set of ideal segments for an utterance is defined as $D_q[h], h = 1, \ldots, H$, and a set of estimated segments for the same sentence as $D_e[l], l = 1, \ldots, L$. An 'utterance' in this work is defined as one single sample of speech, (e.g. a single TIMIT spoken sentence). Each utterance has H segments identified by using hand transcribed information, and L segments identified automatically by the system presented in this paper. The overlapping region between an estimated segment $D_e[h]$ and an ideal segment $D_q[l]$ is defined as $D[h, l]$.

Each of the segments and regions has an equivalent energy $Z_e[h]$, $Z_q[l]$, and the overlapping energy region $Z[h, l]$. Given a threshold value β, an ideal segment $D_q[l]$ is considered to be well covered by an estimated segment $D_e[l]$ if $D[h, l]$ contains most of the energy of the ideal segment. Likewise, $D_e[l]$ is considered to be well covered by $D_q[l]$ if $D[h, l]$ contains most of the energy of $D_e[l]$. This is defined as:

$$Z[h, l] > \beta \cdot Z_q[l] \quad (8)$$

$$Z[h, l] > \beta \cdot Z_e[h] \quad (9)$$

The threshold value β is designed to ensure that an ideal segment is well covered by one estimated segment, and vice versa, when $\beta \in [0.5, 1]$. With the definition of well covered regions satisfied, any regions $D[h, l]$ can be labelled as either correct, under-segmented, over-segmented, or missing.

Correctly segmented regions are those regions $D[h, l]$ that meet both of the well covered criteria defined in equation 9, i.e. if the two regions $D_q[l]$ and $D_e[h]$ mutually cover each other well. This means that there is a close match between the ideal and estimated segments, depending on the threshold value β used.

Over-segmented regions are those where more estimated segments are found than ideal segments. In these regions, one ideal segment covers multiple estimated segments. For example, a TIMIT single phoneme may be covered by multiple estimated segments. To define over-segmentation, $\{D_e[h']\}, h' = h_1, h_2, \ldots, h'_{E_l}$, with $E_l > 0$ is defined as all the estimated segments that are well covered by one ideal segment $D_q[l]$. The overlapping regions $\{D[l, h']\}, h' = h_1, h_2, \ldots, h'_{E_l}$ are defined as being over-segmented if the combined regions contain most of the energy of $D_q[l]$. This is defined by:

$$\sum_{h'} D[l, h'] > \beta \cdot D_q[l], \quad h' = h_1, h_2, \ldots, h'_{E_l} \quad (10)$$

Under-segmentation is where there are more ideal segments than found segments, so that one estimated segment covers multiple ideal segments. This is a less serious issue as it produces larger groupings of segments (for example, covering multiple phonemes). The subset of ideal segments that are covered by one estimated segment $D_e[h]$ is defined as $\{D_q[l']\}, l' = l_1, l_2, \ldots, l'_{F_h}$, with $F_h > 0$. The overlapping regions $\{D[l', h]\}, l' = l_1, l_2, \ldots, l'_{F_h}$ are defined as being under-segmented if the combined regions contain most of the energy of $D_e[h]$. This is defined by:

$$\sum_{l'} D[l', h] > \beta \cdot D_e[h], \quad l' = l_1, l_2, \ldots, l'_{F_h} \quad (11)$$

Finally, if the region $D[l, h]$ is part of an ideal segment, but cannot be labelled as correct, over-segmented, or under-segmented, then it is considered to be an insufficient match, and so it is labelled as missing.

There are some ambiguities in these definitions, firstly, it is possible, depending on the threshold β, for a segment to be both correct and under-segmented. This is because a single region $D[l, h]$ may mutually cover ideal and estimated segments, but neighbouring estimated segments may also cover the ideal segment well (depending on the threshold), leading to the complete subset of regions (including the 'correct' region) being labelled as under-segmented. The same applies for over-segmentation. Both definitions are accurate, but for preference, we have assigned a 'correct' labelling a higher priority, so that any region that can be accurately identified as 'correct' is

defined as such, even if it can also be accurately defined as under-segmented or over-segmented. Again, this follows the precedent set by Hu and Wang [6].

Finally, to calculate the final output values, the energy in each labelled segment is summed. This provides the ideal energy of all ideal segments Z_q, estimated segments Z_e, and all regions labelled as correct Z_{corr}, missing Z_{miss}, over-segmented Z_{over}, and under-segmented Z_{undr}. Finally, we can use these to calculate the overall percentage values for each labelling. The correct percentage is defined as:

$$P_{\text{corr}} = Z_{\text{corr}}/Z_q \cdot 100\% \tag{12}$$

The under-segmentation percentage is defined as:

$$P_{\text{undr}} = Z_{\text{undr}}/Z_q \cdot 100\% \tag{13}$$

The over-segmentation percentage is defined as:

$$P_{\text{over}} = Z_{\text{over}}/Z_q \cdot 100\% \tag{14}$$

The missing percentage is more complex to calculate. This is because there are examples of segments that do not make up part of an overlapping region, and so would be missed by the above calculation. This can be calculated by identifying the difference between the total energy of all labelled regions $Z_{\text{tot}} = Z_{\text{corr}} + Z_{\text{undr}} + Z_{\text{over}} + Z_{\text{miss}}$, and then using the difference between this and the ideal match to determine the full percentage of missing data, as defined by:

$$P_{\text{miss}} = ((1 - (Z_{\text{tot}}/Z_q)) + (Z_{\text{miss}}/Z_q)) * 100; \tag{15}$$

This produces final percentages of each category that match the labels for over-segmentation, under-segmentation, correct segmentation, and missing data. It should be noted that in addition to the ambiguity over correct/over-segmented/under-segmented data, there is also the issue that the phonetic transcription carried out by humans may not be completely exact either.

IV. Results

We evaluate the system presented in Section II using the evaluation metric in Section III. We test: (i) the performance on clean speech segmentation using the TIMIT speech database [9], and (ii) the performance on speech plus noise mixtures, including white noise, sawtooth noise, and natural sounds (river noise), mixed at a number of different signal to noise ratios (SNRs), calculated as RMS SNRs.

Figure 4 demonstrates the segmentation output. The top image shows the spectrogram of a single TIMIT sentence (generated using Audacity: http://audacity.sourceforge.net), and the bottom image the segmentation for each channel (frequency band) output from the filterbank, showing how different segments are identified in different channels. Finally, the lines and crosses at the bottom of the figure show the identified segments subsequently used for evaluation. It can be seen that our multi-channel segmentation approach has identified a number of segments that represent a good match with the spectrogram, and this is then translated into a number of distinct segments.

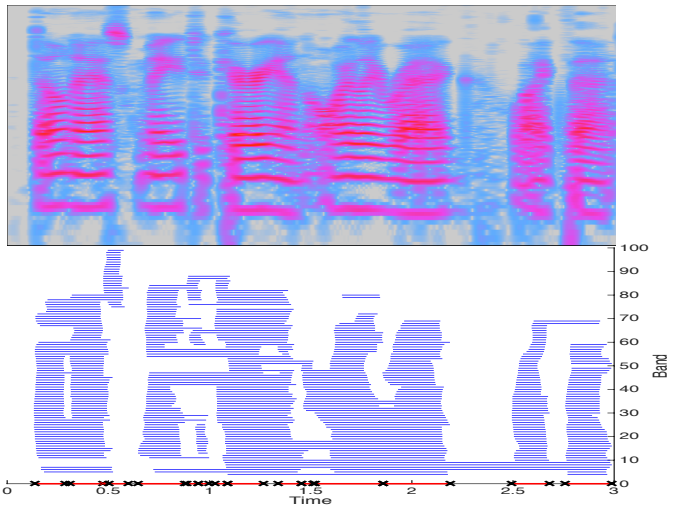

Fig. 4. Example of segmentation of a single TIMIT speech utterance, MAPV0SA2. The top is the spectrogram (50-7000Hz), where the Y axis is logarithmic. The bottom image shows the segmentation for each frequency channel (also 50-7000Hz, logarithmically spaced), with red lines with black x's at bottom marking the segmentation results.

A. Speech Segmentation Performance

To evaluate the system, we applied the segmentation system firstly to 4620 utterances from the well known TIMIT database. To assess the quality of these, we used the phonetic transcriptions provided with the dataset, discounting the closures associated with consonants, as well as pauses. The results are shown in figure 5.

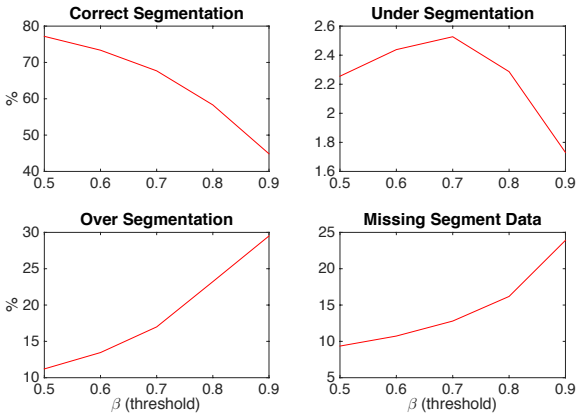

Fig. 5. Results from 4620 TIMIT utterances for clean speech. Top left shows the total percentage energy determined to be a good mutual match (correct), top right shows the percentage energy determined to be under-segmented (one estimated segment to multiple ideal segments), bottom left shows the percentage energy determined to be over-segmented (one ideal segment to multiple estimated segments), and bottom right shows the percentage energy mismatched (i.e. not considered to be a match over the minimum threshold or not overlapping at all). X axis is β (threshold) throughout.

Considering firstly correct segmentation, i.e. one estimated segment and one ideal (hand transcribed) segment cover each other mutually, positive results were found. This can be seen

in the top left graph of figure 5. When the threshold (β) of mutual covering was set to be 0.5, (at least 50% of an ideal and estimated segment cover each other), 77.2% of segments were correctly identified. As β increases, (a larger percentage of an ideal and estimated segment must mutually cover each other), the percentage of correctly identified segments dropped, from 73.4% ($\beta = 0.6$), to 44.9% ($\beta = 0.9$). This large drop is expected as a threshold of 0.9 requires a near perfect match, which is difficult to achieve. Our results for correct segmentation are an improvement on Hu and Wang's [6], although their work focuses on speech in noise, so it is not a full comparison. The ideal segments were transcribed by human listeners, so that start and end times may be slightly inaccurate. In addition, the system does not detect precise phonemes.

Over-segmented regions (see Section III), where one ideal segment covers multiple estimated segments, occur if the system detects multiple segments that are covered by a single ideal segment. The work by Hu and Wang finds very little over-segmentation (below 5% in all cases), however, as the graph at the bottom left of figure 5 shows, the system presented in this paper finds considerably more. With a threshold (β) of 0.5, 11.2% of the signal was found to be over-segmented. As β is increased, there must be more mutual segment coverage to be correctly estimated, meaning a more precise match is required. In fact, as β increases, over-segmentation increases far more than under-segmentation, rising to 29.5% of signal energy being classified as over-segmented. This means that there were 29.5% of incidences where multiple estimated segments were found to be covered by a single ideal segment. This suggests that the system is finding multiple segments per phoneme, and is therefore arguably very sensitive, as it is able to identify segments within phonemes.

Under-segmentation, where one estimated segment covers multiple ideal transcribed segments, is a much rarer occurrence with clean speech, as can be seen in the top right graph of figure 5. Unlike the work of Hu and Wang (and again, this comparison is limited due to their work not covering clean speech), no matter the threshold level, we find very few examples of under-segmentation, with the energy percentage defined as being under-segmented less than 3% in all instances. This confirms that the system is very sensitive to speech segments.

Finally, the bottom right graph of figure 5 shows the mismatched segment energy percentage. This covers segments that were not considered to be matched to an extent covering the threshold, and also those where there is no overlapping region at all. With a threshold (β) of 0.5, 9.3% of segment energy was found to be mismatched, which increased to 23.9% when $\beta = 0.9$. This follows a similar pattern to findings by Hu and Wang, who find mismatched percentages range between 7% and 20% as β increases.

Overall, with clean speech, very good results were found with a very high percentage of correct (mutually covering) segment energy when the system results were compared to the ideal transcribed segments. Although this reduces to 44.8% with a very high threshold, this is still a positive result. In addition, when considering the over-segmented percentage combined with the correctly segmented percentage, even with a threshold of 0.9, the combined percentage is 74.3%. This means that segments are identified, but the system is oversegmenting, and identifying sub-phoneme segments. This may be adjusted by further experimentation with system parameters. Although we compared our results to those reported by Hu and Wang, their work focused on speech in noise, rather than clean speech, and so are not directly comparable. We therefore examine segmentation of speech in noise.

B. Speech in Noise Segmentation

To assess performance on speech in noise, we processed a subset of the TIMIT dataset to which noise had been added. Four types of noise were used, two synthesised wideband sounds (white noise, and a 200Hz sawtooth wave), and two recorded wideband noises (fan noise from a server room, and a recording of the River Allan). The noise recordings were about 50 seconds long, varying little in level across this time: the actual noise added was a randomly chosen section of the same length as the utterance. Noise was added at RMS power levels of -40dB to 60dB SNR by keeping the original signal identical, and scaling the noise. The results are shown in figures 6 and 7.

Figures 6 and 7 show results from 15 selected TIMIT utterances for different types of noise: see figure legend. These results do not compare exactly to the clean speech results presented previously, as these use only 15 sentences selected from the TIMIT corpus rather than the 4620 sentences used for clean speech in figure 5.

A number of trends can be seen in the white noise results in figure 6a. At a very low SNR (-40dB and -20dB), the correctly segmented percentage is very low: further, the under-segmentation percentage is high, peaking at nearly 40% ($\beta = 0.6$) for -40dB SNR. The -20dB SNR mixture is only marginally better. Thus while the system is able to identify some segments, these are not correctly matched to the ideal segments, but each estimated segment is covered by multiple ideal segments. This is verified by the bottom right graph: at -40dB and -20dB there is very little over-segmentation, unlike the clean and reduced noise performance. Finally, at these low SNRs, it can be seen that at $\beta = 0.5$, 25% of segments were found to be mismatched, increasing to above 80% at $\beta = 0.9$. This is considerably higher than for clean speech. Overall, in white noise at very low SNRs, the system performs poorly, as might be expected.

At higher SNRs, the results are much improved. We consider SNRs of 0dB (speech and noise mixed equally), +20dB, +40dB, and +60dB. Firstly, at 0dB (yellow line), the system is finding a lower correctly segmented percentage, and generally increased over-segmentation in comparison to +20dB, +40dB, and +60dB. There is also a slightly increased incidence of under-segmentation in comparison to these SNRs, but a similar level of missing segment data. All scores represent a clear and visible improvement over the negative SNR results.

The results for the positive SNRs tend to be both very similar to each other, and to the clean speech results described in figure 5. All have a very similar level of both correct segmentation (from near 80% at a $\beta = 0.5$ to about 50% at a $\beta = 0.9$). These results compare well to the clean speech (some difference is expected considering the different number of sentences evaluated). Similarly, under-segmentation

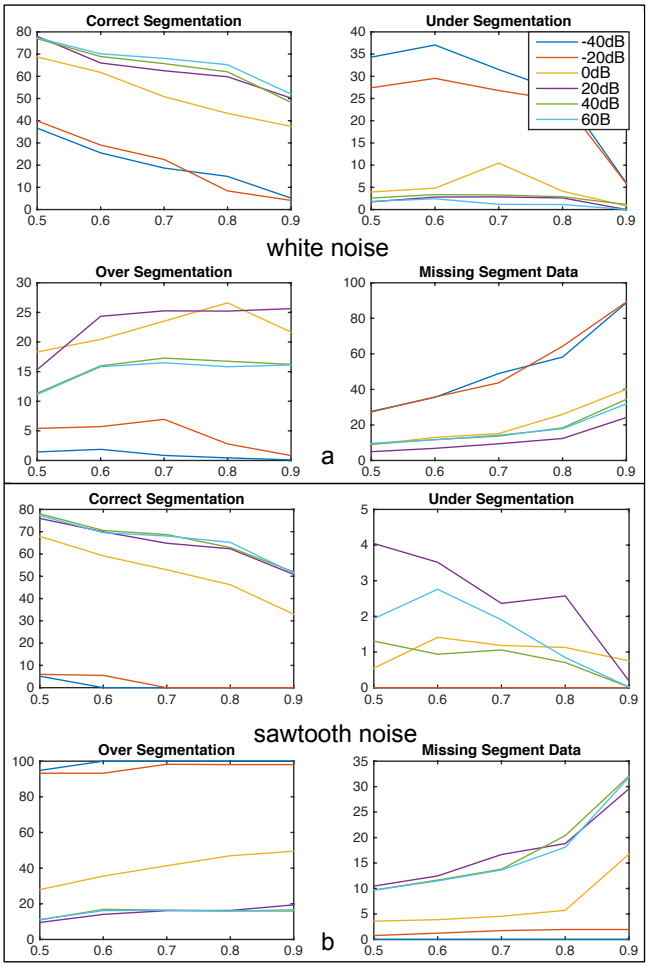

Fig. 6. Results from 15 selected TIMIT utterances for two different types of synthetic noise: (a) white noise, (b) sawtooth noise. For each type, top left shows the percentage energy determined to be a good mutual match (correct), top right shows the percentage energy determined to be under-segmented (one estimated segment to multiple ideal segments), bottom left shows the percentage energy determined to be over-segmented (one ideal segment to multiple estimated segments), and bottom right shows the percentage energy mismatched (i.e. not considered to be a match over the minimum threshold or not overlapping at all). Each figure shows the percentage energy for the β (threshold) value (0.5 to 0.9) for a good match, with each line representing a speech/noise mixture at a different SNR, from -40dB to +60dB. throughout.

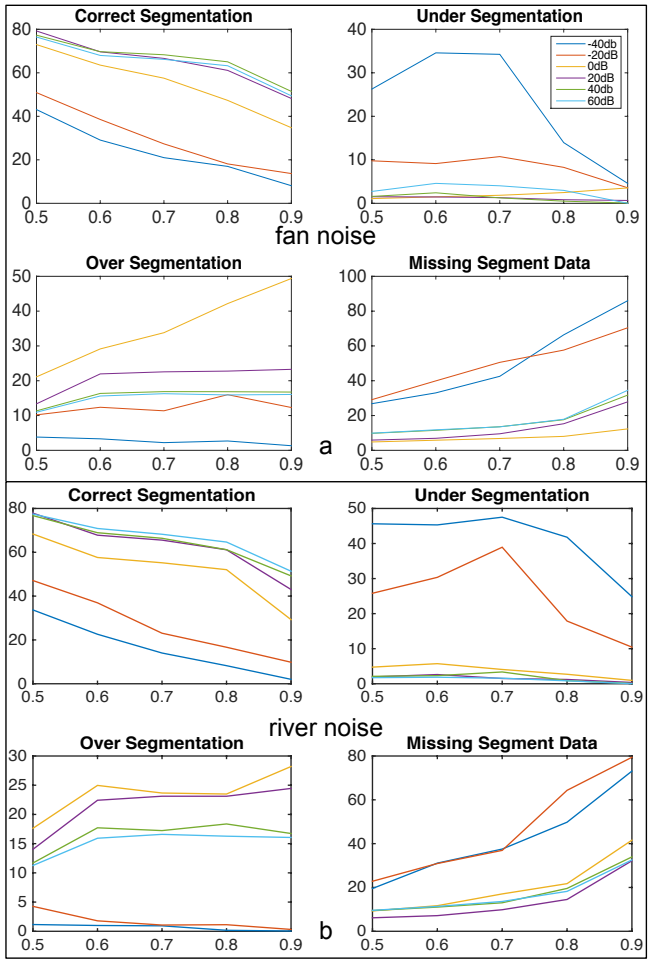

Fig. 7. Results from 15 selected TIMIT utterances for two different types of real noise: (a) fan noise, (b) river noise. Format is as for figure 6.

is generally below 5% as found for clean speech. Based on the clean speech results, we expected that the system would over-segment, and this is the case, with 0dB and +20dB producing more over-segmentation than +40dB and +60dB, demonstrating the effects of noise. However, these results are similar to the clean speech results, as is the mismatched data. This shows that at an SNR of +20dB, +40dB, and +60dB, the system performs similarly to how it does on clean speech, demonstrating a robustness to white noise. The results at 0dB were also found to compare well to the results presented by Hu and Wang.

In figure 6(b), (sentences mixed with sawtooth noise), at SNRs of 0dB, +20dB, +40dB, and +60db, the results are similar to the results found for the equivalent SNRs in (a), showing that the system is also robust to sawtooth noise. There are some differences from (a) at very low SNR (-20dB and -40dB). Rather than the under-segmentation identified previously, the system is reporting that 90%+ of the signal at all thresholds is over-segmented. This suggests that the changes in the sawtooth wave are generating additional segments that are being matched to speech segments incorrectly, and again demonstrates that while robust to noise, when the noise is extremely loud, with a SNR of -20dB and below, the system cannot function.

For the fan noise in figure 7(a), the results for correct segmentation are broadly similar to figure 6. The correct segmentation percentage for -20dB and -40dB is an improvement on white noise and sawtooth, while still being worse than for better SNRs. The other results continue to broadly match the results of figure 6, showing very low under-segmentation, except for SNRs below 0dB, where considerable under-segmentation can be seen. However rather than -40dB and -20dB being similar, -20dB is much closer to the positive SNR results for under-segmentation and over-segmentation, suggesting that the system is demonstrating more robustness to this type of noise than white noise and sawtooth. At positive SNRs (0dB and above), the results match the earlier ones, confirming robustness in this type of noise.

Figure 7(b) shows results for river noise, a realistic noise source. The results are very similar in all cases to the white noise case. In very noisy mixtures, the system is under-segmenting or not finding segments, whereas at 0dB SNR and above, the system is returning broadly similar results to clean speech, as shown in figure 5.

For all different noise types, the results for these 15 sentences follow the trend of clean speech results at SNRs of 0dB and above, with results at +20dB, +40dB, and +60dB being particularly well matched to clean speech results. It confirms that the system is robust to different types of noise, both natural and artificially generated. For very noisy speech mixtures the system performs poorly. The correct segment matches for all types of noise is very low in all cases, with data tending to be under-segmented or mismatched. This suggests that while segments are being found, they are not well matched. One exception is sawtooth noise, where we suspect that the system is finding segments in the sawtooth wave, resulting in false cases of over-segmentation. Overall, apart from in extremely noisy scenarios, the system performs well in both noisy and clean environments.

V. Conclusions and further work

This paper presented a biologically inspired onset and offset based approach to speech segmentation, designed to function in a causal manner. This work extends previous work, inspired by using multiple resolutions suggested in many different works in image segmentation. Here, both frequency and amplitude are considered in a logarithmic way. The onset (and offset) transform used is essentially a wavelet transform (but stopping half-way through the convolving function so that it becomes applicable in real-time), and the actual decision about the location of the segments uses a leaky integrate-and-fire neuron to combine onset (offset) information at differing resolutions, and across bands.

We utilised the TIMIT speech database to evaluate performance on both speech alone, and on speech with a variety of different noise types at varying SNRs. The results identified that with clean speech, the majority of segments were identified as being correctly matched with very little under-segmentation and mismatching, even with a high matching threshold ($\beta = 0.9$). When testing the system with four different noises, both synthetic and natural, at a range of SNRs, it was found that the system was very robust to noise mixtures with an SNR of greater than 0dB, with very similar results to the clean system. When the SNR was -40dB or -20dB, it was found that the system performed poorly, suggesting a limitation to the level of robustness, as might be expected in overwhelmingly noisy speech mixtures. The performance of the system varied slightly depending on the type of noise. Overall, it was found that the system was robust, but with a tendency to over-segment particularly at poor SNRs.

The results show the potential of this technique: future work aims to refine it further, present further detailed evaluation using more sounds (both speech and other) and noise types, and to make the source code and noise signals fully documented and available for other researchers to develop further. Its causal nature makes it suitable for potential real time application, and it is intended to use this system as part of other speech processing applications, for example as part of an audiovisual speech filtering system [16], [17]. This research has resulted in the development of an initial audiovisual speech filtering system, but identified a potential limitation being the single audiovisual model used, rather than the phoneme specific model developed in similar work by Almajai and Milner [1].

Acknowledgment

The authors would like to thank the UK EPSRC (award EP/G062609/1) for funding this work.

References

[1] I. Almajai and B. Milner, "Effective visually-derived Wiener filtering for audio-visual speech processing," in *Interspeech 2009, Brighton*, Aug. 2009, pp. 1–6.

[2] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[3] D. B. Grayden and M. S. Scordilis, "Phonemic segmentation of fluent speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1994, pp. I–73–I–76 vol. 1.

[4] L. S. Smith, "Onset-based sound segmentation," *Advances in neural information processing systems*, pp. 729–735, 1996.

[5] M. F. Caetano, J. J. Burred, and X. Rodet, "Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10), Graz, Austria, September 6-10, 2010*, 2010, pp. 11–21.

[6] G. Hu and D. Wang, "Auditory Segmentation Based on Onset and Offset Analysis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 396–405, 2007.

[7] N. Obin, F. Lamare, and A. Roebel, "Syll-O-Matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 6699–6703.

[8] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data ," *Hearing Research*, vol. 47, pp. 103–138, Nov. 1990.

[9] J. Garofolo, L. Lamel, W. Fisher, and J. Fiscus, "DARPA TIMIT," NIST, 1993.

[10] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Timedomain modeling of peripheral auditory processing: A modular architecture and a software platform," *Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.

[11] L. S. Smith and D. Fraser, "Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses," *Neural Networks, IEEE Transactions on*, vol. 15, no. 5, pp. 1125–1134, 2004.

[12] C. Koch, *Biophysics of Computation*. Oxford, 1999.

[13] L. S. Smith, "Determining ITDs Using Two Microphones on a Flat Panel During Onset Intervals With a Biologically Inspired Spike-Based Technique," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2278–2286, 2007.

[14] J. Gałka and B. Ziółko, "Study of Performance Evaluation Methods for Non-Uniform Speech Segmentation," *International journal of circuits, systems and signal processing*, vol. 1, no. 2, pp. 167–172, Jan. 2007.

[15] A. Hoover, G. JeanBaptiste, X. Y. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 673–689, Jul. 1996.

[16] A. Abel and A. Hussain, "Novel Two-Stage Audiovisual Speech Filtering in Noisy Environments," *Cognitive Computation*, vol. 6, no. 2, pp. 200–217, Oct. 2013.

[17] A. Abel, A. Hussain, and B. Luo, "Cognitively Inspired Speech Processing For Multimodal Hearing Technology," in *Proceedings of IEEE Symp. S. Comp. Intel*, Oct. 2014, pp. 1–8.