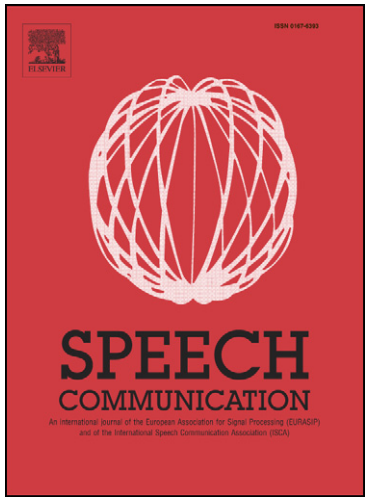
Extracting Amplitude Modulations from Speech in the Time Domain

Garreth Prendergast, Sam R Johnson, Gary G R Green



Extracting Amplitude Modulations from Speech in the Time Domain

Garreth Prendergast, Sam R Johnson, Gary G R Green

PII: S0167-6393(11)00036-7
DOI: 10.1016/j.specom.2011.03.002
Reference: SPECOM 1978

To appear in: *Speech Communication*

Received Date: 24 November 2010
Revised Date: 1 March 2011
Accepted Date: 2 March 2011

Please cite this article as: Prendergast, G., Johnson, S.R., R Green, G.G., Extracting Amplitude Modulations from Speech in the Time Domain, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.03.002

Extracting Amplitude Modulations from Speech in the Time Domain

Garreth Prendergast[,a], Sam R Johnson[a], Gary G R Green[a]

[a] *York Neuroimaging Centre, University of York, UK, YO10 5DG*

Abstract

Natural sounds can be characterised by patterns of changes in loudness (amplitude modulations), and human speech perception studies have focused on the low frequencies contained in the gross temporal structure of speech. Low-pass filtering the temporal envelopes of sub-band filtered speech maintains intelligibility, but it remains unclear how the human auditory system could perform such a modulation domain analysis or even if it does so at all. It is difficult to further manipulate amplitude modulations through frequency-domain filtering to investigate cues the system may use. The current work focuses on a time-domain decomposition of filter output envelopes into pulses of amplitude modulation. The technique demonstrates that signals low-pass filtered in the modulation domain maintain bursts of energy which are comparable to those that can be extracted entirely within the time-domain. This paper presents preliminary work that suggests a time-domain approach, which focuses on the instantaneous features of transient changes in loudness, can be used to study the content of human speech. This approach should be pursued as it allows human speech intelligibility mechanisms to be investigated from a new perspective.

1. Introduction

The temporal features of the amplitude envelope of natural sounds are known to provide key cues for many aspects of fundamental auditory processing including speech intelligibility, sound localisation and acoustic grouping.

This article revisits the analysis of natural speech sounds and describes the modulations present within these signals. The framework described provides a new way of characterising the amplitude modulations found in natural speech and performs the analysis in the time domain. Rather than describing the temporal information in the modulation domain using a band-pass filter, speech envelopes are decomposed into a series of pulses, each of which is defined by its amplitude, half-duration and position in time. This process allows the types of modulations found in natural speech to be described in more detail than if the information is only considered in the modulation domain. The technique outlined leads toward an alternative method for the analysis and synthesis of speech and a model of auditory processing that is designed to focus on the instantaneous features of the temporal structure of sounds.

Vocoded speech experiments typically filter speech signals into a series of sub-bands in a manner akin to the processing performed by the inner ear and cochlear nerve. The temporal envelopes are then extracted from the sub-band filtered speech signals and these envelopes used to modulate narrowband noise or pure-tone carriers. It is known that vocoded speech produced in this way remains highly intelligible (et al., 1939(@) and also that the low-frequency information in speech temporal envelopes, below 16 Hz, is important for maintaining accurate speech intelligibility (et al., 1994a(@,b, 1995(@).

One theory of amplitude modulation processing within the human auditory system is that of a modulation filterbank (et al., 1982(@). A modulation filterbank theory of processing postulates that there is a series of channels in

the auditory system, beyond the filtering of the auditory periphery, which are tuned to specific frequencies of modulation. Such a model has received empirical support in the form of modulation masking experiments (et al., 1989(@) and in recent years has become the main theory of how temporal changes in loudness of signals are processed in the auditory pathway (et al., 2002(@, 1996a(@,b). The principles of linear superposition suggest that the response of a system to complex acoustic patterns which contain amplitude modulations can be predicted by the response to basic, well defined modulations. Although many models of modulation processing include an element of non-linear processing, the foundations of the models are based on linear assumptions. Such a model of processing is related closely to speech vocoding experiments which use band-pass filters to limit the frequency content of speech temporal envelopes. Both the behavioural experiments and modulation filterbank model approach the issue of modulation processing and speech intelligibility mechanisms from the modulation domain, and as a result do not focus on the shape of the modulations or their location in time. Although there are exceptions to this, for example **?** demonstrate that under some circumstances speech intelligibility is sensitive to the phase of the amplitude modulations, in general modulation domain mechanisms are thought to be crucial to accurate speech decomposition and intelligibility (**???**).

Low-frequency amplitude modulations in the temporal envelopes of speech signals have also been used to provide a representation of speech unaffected by reverberation and spectral shaping in the so-called "modulation spectrogram" (et al., 1998(@). The modulation spectrogram focuses on the robust features of a speech signal that are thought to be invariant across different speakers and acoustic conditions by using modulation frequencies between 0-8 Hz with a peak sensitivity at 4 Hz. This peak sensitivity matches the typical peak between 3 and 5 Hz in the long-term modulation spectrum of speech and corresponds to the average frequency of a syllable (et al., 1985(@, 2004(@). The modulation spectrogram is a method of visualising low-frequency modulations contained within a speech utterance and can also be used as a front-end to an automatic speech recognition system to produce more accurate recognition of reverberant speech. Although the modulation spectrum of speech is dominated by the low-frequency terms in a 1/f function, a normalisation of this information with respect to the auditory filter shows a clear peak in the spectrum at around 4 Hz, which is why these frequencies form the focus of much work in the modulation domain.

Experiments which use low-pass filtered envelopes remove all information above a given frequency, but this process allows no insight into which cues in the remaining modulations are crucial for speech processing. It is not necessarily the case that this method of processing accurately mimics the type of processing performed by the auditory cortex. When using stimuli constructed by filtering in the modulation domain, parameters such as the duty cycle, the rate of change and the modulation frequency all co-vary. However, et al. (2008(@) recently demonstrated that if these parameters are dissociated there are neurons in the inferior colliculus of the gerbil which are sensitive to aspects of the waveform in a manner independent of the frequency of modulation. Such studies suggest that considering the problem in the modulation domain may not, necessarily, be the correct approach, as this assumes sinusoidal, periodic and linear processes are responsible for producing an encoding of speech that leads to intelligibility. It may be important to maintain one or more aspects of the amplitude modulations which low-pass filtering does not degrade, meaning that these cues (and therefore intelligibility) remain unaffected. Such features could be the timing of the activity, obtaining an accurate representation of the magnitude of the modulation, preserving information regarding the shape of the acitivty or a combination of these and other features. Few studies investigate the en-

coding of amplitude modulation using nonsinusoidal modulation, despite the fact that the advantages of such an approach have been documented (et al., 1972(@, 2004(@). It is known that both frequency and amplitude modulation detection mechanisms are sensitive to the shape of the waveform used to stimulate the system (et al., 1974(@), however many approaches in the modulation domain focus solely on the amplitude of the modulations and discard any phase information. Recently **?** performed an experiment in which the auditory steady-state response was measured neuromagnetically in response to sinusoidal and non-sinusoidal sounds modulated at a frequency of 4 Hz. The study concluded that the magnitude of the evoked response was non-linearly related to the waveform shape, with non-sinusoidal pulsatile bursts of loudness producing the largest 4 Hz evoked component. However, the timing of the response was found to vary linearly with the shape of the modulation waveform. Therefore it is clear that the human auditory cortex is able to phase-lock to the frequency of interest of a sound, but the mechanism by which this process works is affected by the waveform shape of the sound.

In this article we introduce a novel method for extracting patterns in speech that parameterise the sounds in a manner that allows a flexible examination of the cues associated with speech intelligibility. The methodology described focuses on a time-domain analysis rather than a purely frequency or modulation domain approach. It is important to stress that the work is aimed at generating experimental manipulations of speech and also provides a model of amplitude modulation extraction. It is not concerned with issues of computational efficiency or the compression of speech signals for transmission or storage. Our proposed approach has similarities with work by et al. (2002(@), however this previous work approaches the decomposition of sounds from the perspective of efficient coding theory. That is, investigating how a system can maximise the information stored and transmitted using statistically independent features. et al. (2002(@) shows that different types of environmental sounds are better characterised by different filters and so when considering the problem of how the human auditory system encodes the modulations contained in speech it may be necessary to carefully consider the signal processing mechanisms chosen. Although some of the problems and conclusions are shared by these two approaches, the two implementations are very different and they set out to achieve different goals. The driving force behind the current work is to allow future behavioural and neuromagnetic experiments to use a framework such as the one outlined to develop manipulations of speech which are difficult to obtain using standard filtering techniques in the modulation domain. This will allow further investigation of the specific cues involved in transmitting intelligible speech.

2. Natural modulations in speech

The first aim was to quantify the nature of amplitude modulations that are present in natural speech. This process maintains the phase and shape information of the modulations by decomposing the signals in the time domain and is outlined schematically in figure 1.

Figure 1: Schematic describing the stages of the fitting process. The modulations in a filter envelope are decomposed into pulses of modulation. These pulses can be used to describe the amplitude modulations found in the signal and also to modulate sinusoidal carriers to produce synthesised speech.

1500 low-predictability IEEE sentences were used as stimuli (Advanced Bionics UK Ltd), consisting of 750 unique sentences spoken both by a male and female. Sounds were sampled at 44.1 kHz using a 16 bit integer representation. Sentences were first passed through a Gammatone filterbank (et al., 1988(@) containing 128 logarithmically spaced filters centred at frequencies between 100-21328 Hz. The equivalent rectangular bandwidth of the filters was calculated using the parameters described by et al.

(1990(@).These parameters specify an asymptotic filter quality (Q) for large frequencies of 9.26 and a minimum bandwidth for low frequency filters of 24.7 Hz. The specific implementation produced an equivalent filter quality, Q, as a function of filter position that was flat, except for at low frequencies where a wider relative bandwidth results in a smaller Q (et al., 1994(@). The filters within the filterbank are uniformly spaced on an ERB scale as specified by et al. (1993(@). This involves evenly spacing the 128 centre frequencies between a specified minimum of 100 Hz and the Nyquist frequency. The implementation also ensures a consistent overlap between the bandwidths of neighbouring filters. The result is filters that are logarithmically spaced, apart from at lower frequencies where the filters are positioned more closely.

The temporal envelope of each auditory filter output was extracted using the Hilbert transform and this signal was low-pass filtered (cut-off frequency of 60 Hz). The method of extracting the initial amplitude modulations from the envelope used incoherent detection. **?** highlight the fact that this is a far from optimal method of extracting modulations as less stop-band attenuation than desired is achieved. They instead propose a coherent approach to envelope filtering in which the instantaneous phase is related to the phase of the sub-band. The most commonly implemented method when decomposing and synthesising speech in the modulation domain is incoherent detection, and so this is the approach which we employ. Many of the parameters described are arbitrary starting points from which to outline the technique and demonstrate its applications, it is not the focus of this paper to determine the combination of parameters which gives the most accurate modelling of the amplitude modulations.

Inspection of a single auditory filter output (shown in panel A of figure 2) demonstrates that single-frequency sinusoidal amplitude modulation is not apparent in speech signals and the signal is made up of pulsatile bursts of energy. Due to this transient nature of the extracted envelopes, raised cosine pulses were used as a basis function for the modelling. Each fitted pulse was defined by 3 parameters; the amplitude, the half-duration (time taken from the start of the pulse to the peak) and the centre position of the pulse in time.

The third step was to define a window of modulation. The time point with the peak amplitude in the temporal envelope was found and the signal analysed sample-by-sample to determine the start and stop points of the peak. A maximum window half-duration of 6000 samples (136 ms) was allowed and these default start/stop points were modified if either of two terminating conditions was met. The first condition concerned amplitude and was met if a sample was lower than the termination value, defined as; $\alpha A j$. α was a threshold variable, A was the amplitude of the peak and j was the pulse number. The amplitude of the peak was scaled as a function of pulse number; as more pulses are identified the amplitude of the peak decreases. The second condition was concerned with identifying whether a decrease and then an increase in amplitude was a separate modulation and was met if the amplitude of the time point currently being analysed exceeded the termination value, defined as; $m_n + (\beta m_x)$. β was a tuning variable, m_n the lowest amplitude in the window and m_x the amplitude of the current sample. If the second condition was met, the default start/stop point was shortened to the position at which m_n was measured. If a window was less than 100 samples (around 2.3 ms) in half-duration, the modulations within the window were not modelled. α and β are parameters of the fitting technique and for the results described in this paper they were set to 0.02 and $\frac{1}{30}$ respectively. It must be noted that these parameters were chosen as they gave acceptable representations of the original waveforms. The specific parameters that provide the most accurate representation is currently unknown.

An example of the windowing process can be seen in panel B of figure 2. The initial start and stop point were found 136 ms before and after the peak (shown by a dia-

mond). The circle on the onset of the peak identifies the new start point which was set due to the second terminating condition being met. The stop point was re-positioned due to the amplitude terminating condition being met.

Figure 2: The fitting process. Panel A shows the envelope extracted from the auditory filter centred at 500 Hz after low-pass filtering (cut-off frequency of 60 Hz). Panel B shows where the algorithm has identified the peak (diamond), and the start and stop point of the modulation (circles). The amplitude termination value is plotted as a square. Panel C shows the raised-cosine pulse fitted to the window of modulation and panel D shows the waveform generated by adding the raised-cosine pulses together.

As each window of modulation was identified, a non-linear least squares fitting algorithm was applied to determine the best fit parameters of a raised cosine, the initial estimates of which were the amplitude of the peak and the shortest distance between the peak and the window edge. Panel C of figure 2 shows the cosine pulse fitted to the window defined in panel B. The fitted cosine was then subtracted from the original envelope and the next peak identified and a new window defined. The decomposition process used a maximum of 35 raised cosine pulses per filter envelope and was set to account for no more than 99% of the variance. Only peaks in the envelope greater than zero were modelled, although in practice the first 35 peaks always satisfied this constraint. These parameters resulted in an accurate representation of the waveform whilst remaining computationally manageable. A visual comparison of the waveforms shown in panel A and D of figure 2 confirms that, at the global level, the signals are similar; both waveforms showing five clear bursts of energy and the timing of the bursts is consistent in the two representations. Where the waveforms differ is in the lower amplitude, fine-detail of the envelopes. Although the first two bursts of energy appear very similar in shape in the two representations, the other sections of activation have subtly different structure in the original waveform when compared to the modelled waveform.

Figure 3 shows the distribution of pulse half-durations averaged across the processing of 1500 sentences. The x-axis plots pulse half-duration in 1 ms bins and the y-axis shows the 128 centre frequencies of the Gammatone filterbank used. Although longer pulse durations were found, these were few in number for any specific sentence and so the figures and subsequent discussions focus on half-durations within the range 1-75 ms. The colour map represents the average number of pulses found at any particular grid location and therefore the figure can also be thought of as a probability distribution for the pulse distribution of any given sentence. Figure 3a shows that the highest concentration of pulses is in the 500-2000 Hz range at a half-duration of around 8 ms. The lower frequencies (i.e. below 500 Hz) tend to show the most concentration of pulses at a half-duration which increases as the centre frequency of the filter decreases. This may be related to the bandwidths being narrower for these low-frequency filters. In addition, at higher frequencies (5-10 kHz) there is a small concentation of pulses with half-durations around 5 ms and also a broader peak in the distribution than that seen for the middle frequencies (500-2000 Hz). The most common half-duration in these higher frequencies is around 11 ms, which is longer than that seen at the middle frequencies. It is difficult to determine if these subtle features in the distribution which vary across filter centre frequencies are related to the bandwidths used for the decompositon, or if they represent features inherent to the modulations contained in speech.

The analysis described led to an average of 4473 pulses in 128 auditory filters which cover the entire spectrum for each sentence. The most common pulse half-duration in natural speech sentences was around 12 ms. As the average sentence duration of the corpus was around 2.5 seconds this equates to just 14 pulses per second in each channel. Across all 1500 utterances and 128 filter outputs, the average variance accounted for by the modelled temporal envelope was 92%. The maximum of 35 pulses per chan-

nel was chosen as pilot work indicated this gave a good representation of the original waveform. In future work it is anticipated that further investigation on intelligibility measures would yield some optimal rate of pulses per second and the maximum number of pulses could be adaptive. Describing all the pulses is an appropriate starting point and analysis of the results confirms that as fewer pulses were modelled per envelope, the structure of the histogram remained the same with fewer pulses at each location.

Figure 3: The top panel shows the number of pulses found in each auditory channel at each half-duration for a typical sentence, generated by averaging the decomposition of 1500 utterances. The lower panel is a histogram showing the half-duration of all the pulses for an average utterance (generated by summing the number of pulses across channels).

The analysis method described provides a time-domain decomposition of transient changes in loudness found in speech temporal envelopes. The pulses can be used to generate a modulation waveform that can be imposed upon a pure tone with a frequency that matches the centre frequency of the Gammatone filter that produced the original envelope. These signals can be generated for all 128 filters and then summed to generate a vocoded speech utterance. As the envelopes of the synthesised speech are generated from the pulses, and each pulse is described by its own unique parameters, there are a number of manipulations possible. The degree of flexibility that the parametrisation provides would be difficult to achieve by using filtering techniques alone. The method outlined also presents a model of how the auditory system extracts and encodes transient changes in loudness. The model proposes that rather than a bank of filters existing, each of which is sensitive to a specific frequency of modulation, the system is concerned with detecting transient changes in loudness and encoding the duration and position of these changes.

3. Comparison with low-pass filtered envelopes

As outlined in the earlier discussion, the temporal envelopes of speech can be low-pass filtered below 16 Hz and speech remains highly intelligible when tested in quiet (et al., 1994a(@,b, 1995(@). These experiments demonstrate that accurately representing the original temporal envelopes of speech is not essential. It is important to preserve the gross temporal structure, but the detail provided by high-frequency modulations is not critical to maintaining speech intelligibility. Therefore, although the decomposition technique outlined is able to accurately model the original speech temporal envelope this is not necessary to allow accurate speech intelligibility. Experiments that low-pass filter speech envelopes are often taken as evidence that it is the low-frequency modulations that are important and the system processes this information in the modulation domain. However, it is possible that the system relies on other cues which are preserved by the process of low-pass filtering and it is therefore of interest to investigate the types of transients maintained by filtering in this way. In order to investigate this the same decomposition technique was used with the extracted envelopes low-pass filtered with a cut-off frequency of 16 Hz rather than 60 Hz. The pulses extracted from the low-pass filtered envelopes are shown in figure 4.

Figure 4: Pulse distribution for low-pass filtered temporal envelopes. The top panel shows the number of pulses found in each auditory channel at each half-duration for a typical sentence, generated by averaging the decomposition of 1500 utterances. The lower panel is a histogram showing the half-duration of all the pulses for an average utterance.

Inspection of figure 4 confirms that the process of low-pass filtering the temporal envelopes at 16 Hz alters the distribution of pulse half-durations for an average sentence. 48% of all pulses now fall in the 15-35 ms region. One possible hypothesis is that transient changes in loudness within a specific range (for example 15-35 ms) are

important for speech intelligibility and they are extracted and encoded in the time domain rather than the modulation domain. As low-pass filtering sub-band filtered speech envelopes does not degrade these cues, experiments using low-pass filtered speech envelopes show high levels of intelligibility and attribute the key features of this process to modulation domain processing. Testing such a hypothesis using the method outlined in this paper may allow an investigation of why low-pass filtered speech remains intelligible and perhaps uncover information regarding the mechansisms of how these modulations are extracted from an acoustic input and encoded in the auditory system. Using the framework outlined it is possible to extract specific types of pulses, or to experimentally manipulate these modulations in order to test specific and subtle hypotheses which may allow us to uncover the mechanisms used by the auditory system to perceive intelligible speech.

4. Refined fit

The initial fitting process achieved an accurate fit of the waveform (on average 92% of the variance in an envelope was explained). However, as discussed, it is not necessary to accurately maintain the temporal envelope as low-pass filtered speech remains intelligible. Therefore, rather than using a time-domain decomposition to accurately represent temporal envelopes, it may only be necessary to maintain modulations equivalent to those which are retained by the process of low-pass filtering the envelopes. The next step in the analysis of the extracted pulses was to investigate whether a system could extract pulses in the time domain which matched the types of signals generated by low-pass filtering the original temporal envelopes. This process "refines" the pulses initially extracted in order to focus less on the fine detail of the original envelope and to maintain the gross temporal structure of the signal. The refining process essentially defines a region of a filter output where a number of pulses overlap and then uses a *single* pulse to model this region of modulation.

Figure 5 (panel A) shows the lengths and positions of the pulses generated by the initial fit for a single auditory filter. Each pulse is plotted as a horizontal line, with the position along the abscissa signalling the pulse number (and therefore the relative amplitude of the pulse as the amplitudes decrease with increasing pulse number) and the horizontal line the duration of the pulse. The process of refining the pulses identified those which have sufficient overlap to be identified as part of the same modulation complex. The first window was defined as being the time points which the first pulse spanned. All the pulses were iterated over and if a pulse fell entirely within this window it was discarded. If 30% of a pulse overlapped with the current window, the window was reset to encompass both pulses. When the window was re-sized, all the pulses were iterated over again with the new window length. Panel B of figure 5 shows the pulses identified as falling within a common window as black lines and all pulses outside this window as grey lines. The window over which a single pulse was fitted to account for all the pulses identified is shown by the dashed line above the final pulse. The process was repeated until no pulses overlapped with the current window. A new window was then defined as being the duration of the next pulse not already accounted for. These newly defined windows (shown in panel C) were taken in turn and the same non-linear least squares fitting algorithm returned the parameters describing a single best fit raised-cosine pulse. Panel D of figure 5 shows the original envelope of an auditory filter, this signal low-pass filtered (the cut-off point being 16 Hz), and the waveforms generated by using both the original pulses and the refined pulses. Inspection of this figure confirms that the low-pass filtered waveform is similar to the waveform created using the refined pulses. For all 1500 utterances and 128 filters, the correlation between the original envelope low-pass filtered at 16 Hz and the envelope generated using the refined pulses was calculated. The average correlation coefficient for the envelopes was 0.83, with one standard deviation of

0.07. This analysis confirms that the signals generated by the refining process contain similar information to a low-pass filtered version of the original filter output envelope.

Figure 5: Figure showing the process of refining the pulses. Panel A shows the duration and position of the original pulses extracted from a typical channel for one sentence. Panel B shows the pulses that are identified as falling within a single window (heavy lines), the final window over which one pulse will be fitted (the dashed line) and all remaining pulses (light lines). Panel C shows the position and duration of the 10 refined pulses. Panel D allows a comparison of the original modulation signal, the modelled signal, the refined signal, the signal generated by low-pass filtering (cut-off frequency of 16 Hz).

The upper panel of figure 6 shows the pulse distribution across all channels of the original modulation extraction (shown previously in the lower panel of figure 3). The middle panel shows the same information for waveforms low-pass filtered at 16 Hz (shown previously in figure 4) whilst the lower panel shows the total pulse distribution for the refined pulse extraction described above. The first point to note is that the low-pass filtered distribution shows half-durations predominantly in the 20-30 ms range, with only a small number of short pulses. Low-frequency temporal modulations are often cited as being crucial for maintaining speech intelligibility but it may have more to do with the shape of the modulations maintained by such a filter.

The refining process described uses fewer pulses than the initial envelope decomposition (as indicated by the y-axis scale of figure 6) and rather than the distribution continually decreasing after reaching the peak, there is a plateau at half-durations of around 20-35 ms before the distribution begins to form its tail. The initial peak in the histogram of refined pulses (at around 10 ms) can be attributed to the brief duration pulses that were not discarded or found to overlap with a window during the refining process. Whereas the initial pulse extraction saw a constant decline in the number of pulses at half-durations after the peak, the refined distribution sees a sustained concentration of pulses from 20-35 ms. The refined fit of the modulations can be thought of as signalling the position in time of the modulations; encoding when a modulation occurs, but only giving gross information as to the duration and shape of this modulation. An average sentence has 1600 refined pulses, which corresponds to an average of 5 pulses per second in each channel. The plateau seen after the peak in the refined pulses is similar to the half-durations most dominant in the low-pass filtered pulse extraction. Therefore the information preserved by low-pass filtering speech envelopes can potentially be retained by a time-domain approach. So, although low-pass filtered speech remains intelligible, the actual method of extraction may not be based in the modulation frequency domain. Such a hypothesis of course requires more explicit and rigorous testing, but what we outline in this paper is a framework within which such experiments can be conducted.

Figure 6: The upper panel shows the histogram of pulse half-durations for an average sentence from the initial decomposition process. The middle panel shows the distribution of these pulses extracted from low-pass filtered envelopes and the lower panel shows the distribution of the refined pulses.

5. Real-time decomposition

The method of amplitude extraction outlined has the primary goal of being used to synthesise vocoded speech sentences which may then allow researchers to consider the role of cues outside of the modulation domain that are typically ignored by many synthesis methods. Thus far we have demonstrated that the types of amplitude modulations maintained by filtering in the sub-16 Hz range are similar to the types of pulses that can be extracted by a time-domain analysis of the incoming signal. The current assumption in the field of speech perception is that incoming auditory signals are analysed in the modulation

domain. If the approach proposed in this paper is more appropriate it is necessary to determine if such a method could work on the time-scale required for intelligible speech perception. The method outlined requires the whole utterance to be present in memory so the peaks can be extracted. The final analysis step was to determine if the pulses could be extracted in real-time.

The refined pulse representation described encodes the precise position of the modulation in time and an estimate of the duration and shape of the transient change in loudness. The low-pass filter analysis demonstrated that pulses of 20-30 ms are typically maintained by filtering in this range and this matched the secondary peaks in the distribution of refined pulses. The aim of the real-time analysis was to investigate whether comparable pulses could be extracted by analysing the waveform in the temporal domain. The temporal envelopes were extracted from the sub-band filtered speech using the Hilbert envelope as previously described. Envelopes were then analysed sample-by-sample and a critical modulation was signalled when a predetermined threshold was crossed. The peak of this critical modulation was encoded and a pulse was used to model the activity which had a fixed half-duration of 30ms and an amplitude to match the amplitude of the identified peak. To calculate the threshold the smallest amplitude of pulse from the average distribution was extracted for each filter and the largest of these values was used. This value also fell within the average pulse amplitude for the original and refined pulses. The threshold was fixed across all filter outputs and utterances. One point to note however, is that the implementation described was used in order to determine if such an approach can be effective in principle. If such an approach was to accurately describe how the auditory system performs decomposition of acoustic inputs, or was to be used to truly analyse speech in real-time it would be necessary to develop a method of determining an appropriate threshold in real-time.

The fitting process outlined extracted an average of 1560 pulses per utterance, which is comparable to the number of pulses extracted for the refined pulse representation. Figure 7 shows two examples of the pulses used to model filter envelopes in real-time. If such a decomposition is able to produce intelligible speech this would add weight to the notion that human mechanisms of amplitude modulation extraction and encoding could function outside of the modulation domain. The issue of intelligibility is the focus of the final section of this paper.

Figure 7: Plotted in blue are the temporal envelopes extracted from two different auditory filter outputs. The pulses fitted by the real-time analysis are shown in red, and the pre-determined threshold level is plotted as a horizontal green line.

6. Intelligibility

The second aim was to assess how effective this technique was for creating vocoded speech sentences. Once the pulses had been extracted from an auditory filter envelope they were added together to create a modulation waveform. This waveform was used to modulate a pure-tone with a frequency that matched the centre frequency of each auditory filter. This was performed on all 128 filter outputs and the signals were then summed across auditory channels to produce synthesised speech. To measure intelligibility, 30 sentences were selected from the corpus at random and intelligibility scores were collected from 10 individuals selected from the student population of the University of York (5 males and 5 females with a mean age of 19 years). All participants confirmed they had no known hearing impairment and were asked to listen to each of the sentences and write down what they heard. Participants were given no training on the task and were allowed to listen to only 10 sentences to familiarise them with the types of sounds being used. These pre-test sounds were not used during the experiment. For each utterance the number of words correctly identified was divided by the total number of words to yield an intelligibility score. Marking was

stringent, with any grammatical errors resulting in a word being recorded as incorrect. A single value relating to the percentage of words correctly identified was calculated for each individual, averaging scores across all 30 sentences. The group average confirmed that this produced vocoded speech which was 93% intelligible (S.E. 1%).

The intelligibility of the refined pulses was also measured using the same method and participants as described previously, and vocoded speech sentences using the refined pulses yielded an intelligibility measure of 88% (S.E. 1%). This confirms that it is not necessary to accurately maintain the shape of the modulation and that in 128 auditory filters which cover the entire spectrum, an average of 1600 pulses are sufficient to provide a high level of intelligibility. This corresponds to an average of only 5 pulses per second in each channel, which is similar to the average syllable rate in speech (et al., 2004(@). The refined fit of the modulation can be thought of as signalling the position in time of the modulations; encoding when a modulation occurs, but only giving gross information as to the duration and shape of this modulation. Figure 5 (panel D) clearly shows that the waveform generated from the refined pulses is very similar to a 16 Hz low-pass filtered version of the original envelope. Therefore, it may not be that it is crucial to maintain the low-frequency envelope information, it may simply be that low-pass filtering the signals does not eliminate the crucial cues which allow the position in time of a modulation to be encoded.

The intelligibility of the real-time decomposition was also evaluated. The pulses in this synthesis were fixed to have a half-duration of 30 ms and produced an intelligibility score of 69% (S.E. of 5%). However, it was noted that in this experimental manipulation, there was not only greater variability across the intelligibility score for each individual, but also across the 30 utterances heard by each participant. It appears as though the decomposition used either provides speech which is as intelligible as other conditions, or speech that is highly unintelligible. This suggests that the parameters used in this experiment, whilst suitable for some utterances, are not able to efficiently decompose all the speech utterances. For the original synthesised speech, using all 35 pulses per channel, from the 300 sentences heard across all 10 participants, over two thirds were reported with no errors (202 utterances). For the real-time decomposition this number dropped to around one third (91 utterances). For the original reconstructions using all the pulses, over 86% had an intelligibility score greater than 80%, whereas only 53% of the real-time decompositions resulted in intelligibility over 75%. Therefore, the real-time decomposition described, using a fixed threshold for each channel and a fixed pulse-duration creates speech which is, on average, 70% intelligible. However, further analyses suggest that this approach may be a plausible way to analyse speech, but the thresholds used to determine the presence of a critical modulation must be adaptive, or at the very least specific to individual sentences.

7. General Discussion

The current paper describes a method which extracts the amplitude modulations found in natural speech as a series of independent pulses with an amplitude, half-duration and position in time. This process accurately models the temporal envelope of the auditory filter outputs and the distribution of these pulses confirms that speech is made up of short duration bursts of amplitude modulation, which are rapid and transient in nature. However, this representation of speech signals allows a more flexible parameter space within which to investigate speech intelligibility mechanisms.

The extracted pulses can then be refined to produce a less accurate representation of the waveform. The refined temporal envelope has, on average, a high correlation with signals generated by low-pass filtering the original temporal envelope. This suggests that the crucial speech intelligibility information preserved by low-pass filtering the sig-

nal can be obtained in the time domain by focusing on the transient changes in loudness. This model of modulation processing in the time domain could potentially describe how the human auditory system performs modulation extraction and can be used to further investigate the role of this information in speech intelligibility mechanisms. Experiments reporting the intelligibility of low-pass filtered speech are typically conducted in quiet, which raises the question of how noise affects the decomposition of speech signals and intelligibility mechanisms. One next step using the technique outlined would be to assess the ability of the algorithm to extract crucial amplitude modulations in the presence of noise and to examine the intelligibility of these reconstructions.

The decomposition allows a flexible parameter space within which to focus on the cues that lead to speech intelligibility. Each pulse has its own parameters and any of these can be manipulated ad-hoc. Therefore, it is possible to understand in more detail which cues are crucial for speech intelligibility. Although the current implementation performs the decomposition and assesses intelligibility as a whole, the approach could be expanded to focus on specific units of speech, such as phonemes or syllables. In contrast, when the temporal envelope has been filtered it is more difficult to further manipulate this information to investigate the relative importance of the remaining cues. For example, the refined pulses and the real time pulses potentially identify the timing of crucial modulations. These timings could be analysed across neighbouring channels in order to extract the shape of an auditory feature These features could then be manipulated in order to create a continuum of sounds across which the variations in the envelopes is subtle. Such manipulations may be able to identify whether the key component of an envelope modulation is its duration, the timing of its onset, the timing of the peak or an interaction of these parameters. The decomposition model proposed could then be further refined and used to better predict the intelligibility level of speech manipulations and experiments which aim to understand the neural representation of these changes in loudness. For example, it may transpire, that the onset of the modulations is more important than the offset, or that an accurate representation of waveform shape is paramount. In such a model, it could therefore be sub-optimal to use a symmetric basis function and future work could also explore if a different, non-symmetric basis function could more accurately represent certain acoustic features that are key to intelligibility mechanisms.

The current work describes a framework within which to decompose speech signals in a highly parametrised format. This process was performed in the time domain and shows that speech is predominantly made up of short pulses of energy (around 10 ms in half-duration). This representation of speech can be further reduced to focus on the timing of the modulation rather than the duration of specific modulations and the information retained is found to show similarities to pulse half-durations extracted from speech low-pass filtered in the modulation domain. Both the original pulses and the refined pulses can be used to create flexible vocoded speech manipulations to investigate the role of instantaneous envelope features in speech perception.

References

Dudley, H, 1939. Remaking speech. Journal of the Acoustical Society of America 11, 169–177.

Capranica, R R, 1972. Why auditory neurophysiologists should be more interested in animal sound communication. The Physiologist 15 (2), 55–60.

Green, G G and Kay, R H, 1974. Proceedings: Channels in the human auditory system concerned with the wave form of the modulation present in amplitude and frequency-modulated tones. J Physiol 241 (1), 29P–30P.

Kay, R H, 1982. Hearing of modulation in sounds. Physiol Rev 62 (3), 894–975.

Houtgast, T and Steeneken, H J, 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. Journal of the Acoustical Society of America 77 (3), 1069–1077.

Patterson, R and Holdsworth, J and Nimo-Smith, I and Rice, P, 1988. Tech. Rep. 2341, Cambridge, UK: MRC Applied Psychology Unit.

Bacon, S P and Grantham, D W, 1989. Modulation masking: effects of modulation frequency, depth, and phase. J Acoust Soc Am 85 (6), 2575–80.

Glasberg, B R and Moore, B C J, 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Research 47, 103–108.

Slaney, M, 1993. Perception group - Advanced technology group. 35, Apple Computer Technical Report, Apple Computer Inc.

Slaney, M, 1994. Tech. Rep. 45, Apple Tenchical Report, Apple Computer Inc.

Drullman, R and Festen, J M and Plomp, R, 1994a. Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95 (5 Pt 1), 2670–80.

Drullman, R and Festen, J M and Plomp, R, 1994b. Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95 (2), 1053–64.

Shannon, R V and Zeng, F G and Kamath, V and Wygonski, J and Ekelid, M, 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–4.

Dau, T and Puschel, D and Kohlrausch, A, 1996a. A quantitative model of the "effective" signal processing in the auditory system. i. model structure. J Acoust Soc Am 99 (6), 3615–22.

Dau, T and Puschel, D and Kohlrausch, A, 1996b. A quantitative model of the "effective" signal processing in the auditory system. ii. simulations and measurements. J Acoust Soc Am 99 (6), 3623–31.

Kingsbury, B. E. D. and Morgan, N. and Greenberg, S., Aug 1998. Robust speech recognition using the modulation spectrogram. Speech Communication 25 (1-3), 117–132.

Lewicki, Michael S, 2002. Efficient coding of natural sounds. Nat Neurosci 5 (4), 356–63.

Viemeister, N F and Rickert, M and Law, M and Stellmack, M A, 2002. In: . , Genetics and the function of the auditory system. pp., 273–291.

Joris, P X and Schreiner, C E and Rees, A, 2004. Neural processing of amplitude-modulated sounds. Physiol Rev 84 (2), 541–77.

Greenberg, S and Arai,T, 2004. What are the essential cues for understanding spoken language? IEICE transactions on information and systems 87 (5), 1059–1070.

Krebs, B and Lesica, N A and Grothe, B, 2008. The representation of amplitude modulations in the mammalian auditory midbrain. Journal of Neurophysiology 100, 1602–1609.