Accepted Manuscript

Enhancement of speech perception in noise by periodicity processing: A neurobiological model and signal processing algorithm

Neil McLachlan, David B. Grayden

PII: S0167-6393(13)00128-3
DOI: <http://dx.doi.org/10.1016/j.specom.2013.09.007>
Reference: SPECOM 2188

To appear in: *Speech Communication*

Received Date: 31 July 2012
Revised Date: 16 September 2013
Accepted Date: 16 September 2013

Please cite this article as: McLachlan, N., Grayden, D.B., Enhancement of speech perception in noise by periodicity processing: A neurobiological model and signal processing algorithm, *Speech Communication* (2013), doi: <http://dx.doi.org/10.1016/j.specom.2013.09.007>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Enhancement of speech perception in noise by periodicity processing:
A neurobiological model and signal processing algorithm**

Neil McLachlan ^{a,*}, David B. Grayden ^{b,c}

^a Melbourne School of Psychological Sciences, University of Melbourne, Victoria 3010, Australia

^b NeuroEngineering Laboratory, Department of Electrical and Electronic Engineering, University of Melbourne, Victoria 3010, Australia

^c Centre for Neural Engineering, University of Melbourne, Victoria 3010, Australia

* Corresponding author. Tel.: +613 8344 6290.

E-mail addresses: mcln@unimelb.edu.au (N. McLachlan), grayden@unimelb.edu.au (D. B.

Grayden)

Abstract

The perceived loudness of sound increases with its tonality or periodicity, and the pitch strength of tones are linearly proportional to their sound pressure level. These observations suggest a fundamental relationship between pitch strength and loudness. This relationship may be explained by the superimposition of inputs to inferior colliculus neurons from cochlear nucleus chopper cells and phase locked spike trains from the lateral lemniscus. The regularity of chopper cell outputs increases for stimuli with periodicity at the same frequency as their intrinsic chopping rate. So inputs to inferior colliculus cells become synchronized for periodic stimuli, leading to increased likelihood that they will fire and increased salience of periodic signal components at the characteristic frequency of the inferior colliculus cell. A computer algorithm to enhance speech in noise was based on this model. The periodicity of the outputs of a Gammatone filter bank after each sound onset was determined by first sampling each filter channel at a range of typical chopper cell frequencies and then passing these amplitudes through a step function to simulate the firing of coincidence detecting neurons in the inferior colliculus. Filter channel amplification was based on the maximum accumulated spike count after each onset, resulting in increased amplitudes for filter channels with greater periodicity. The speech intelligibility of stimuli in noise was not changed when the algorithm was used to remove around 14 dB of noise from stimuli with signal-noise ratios of around 0 dB. This mechanism is a likely candidate for enhancing speech recognition in noise, and raises the proposition that pitch itself is an epiphenomenon that evolved from neural mechanisms that boost the hearing sensitivity of animals to vocalizations.

Keywords

Neurocognitive; model; periodicity; segregation; algorithm; speech

1. Introduction

Pitch height is a psychological dimension related to the frequency or period of sounds. Pitch strength may be defined as the certainty of a pitch height (Beerends and Houtsma, 1989; Rakowski, 1996), and has been measured by pairwise comparisons between stimuli of various tonal characteristics where loudness was held constant (Fastl and Stoll, 1979). Pitch strength and loudness judgments are not independent. For pure tones, pitch strength increases linearly with sound pressure level over a range of 60 dB (Fastl, 1989), and both loudness and pitch strength increase with stimulus duration up to around 200 ms (Fastl, 1989; Viemeister and Wakefield, 1991). In contrast to the variability of pitch strength with loudness, the pitch height of a 1 kHz pure tone varies by less than 1% for sound pressure levels between 40 and 80 dB SPL (Zwicker and Fastl, 1999).

Further evidence for a relationship between pitch strength and loudness include the findings that pure tones produce much greater loudness than noise bands (Fastl and Stoll, 1979), such that the sound pressure level of octave band noise with an embedded pure tone is 10 dB less than octave band noise alone with matched loudness (Kryter and Pearsons, 1965). Soeta et al. (2004) and Kidd et al. (1989) showed that this effect was preserved even if the width of the noise band was less than one critical bandwidth. Furthermore, Soeta et al. (2007) showed that increases in the pitch strength of iterated rippled noise increased perceived loudness, thereby suggesting that waveform periodicity subserves the relationship between loudness and pitch strength. Furthermore, in studies of auditory evoked potentials, an N1 peak and sustained negative potentials at latencies between 400 and 800 ms were associated with periodicity processing by appending iterated rippled noise to bandpass noise with the same spectral profile (Krumbholz et al., 2003; Seither-Preisler et al., 2006). Increases in the amplitudes of the N1 peak and the longer latency sustained negative potentials are likely associated with enhanced spike rates, and were proportional to the iterated rippled noise pitch strength (Krumbholz et al., 2003; Seither-Preisler et al., 2006).

Most research on the neurobiological function of periodicity processing has focused on the importance of periodicity processing for pitch ([Cariani, 2001](#); [Dau et al., 1997](#); [Dicke et al., 2007](#); [Guérin et al., 2006](#); [McLachlan, 2009](#); [Meddis and O'Mard, 2006](#); [Nelson and Carney 2004](#)). However, the ability to hear tones in noise has important implications for speech intelligibility under noisy conditions ([Plomp, 1994](#)). In support of this, the detection of common periodicity has been successfully used to segment parts of noisy speech signals that contain vowel sounds ([Hu and Wang, 2008](#)), and the use of periodicity-based features was found to enhance the automatic speech recognition rates of voiced components of speech in the presence of noise ([Ishizuka & Nakatani, 2006](#)). Other researchers have considered the role of periodicity processing in segregating vowels by autocorrelation and similar algorithms such as recurrent neural networks ([Assmann and Summerfield, 1990](#); [Cariani, 2001](#); [de Cheveigné and Kawahara, 1999](#); [Meddis and Hewitt, 1992](#)), and in the formation of stable auditory images of periodic sounds through the cross-channel temporal alignment of the maxima of auditory nerve spike rates ([Patterson et al., 1995](#)).

While most techniques that attempt to improve speech intelligibility in noise seek to remove the energy associated with noise by estimating its spectrum ([Yoo et al., 2007](#)), a small number of studies have sought to enhance spectral components associated with vowels ([Cheng and O'Shaughnessy, 1991](#); [Turicchia and Sarpeshkar, 2005](#)). However, many studies have shown that transient voiced information is very important for speech intelligibility ([Hazen and Simpson, 1998](#); [Howell and Rosen, 1983](#); [Richardson et al., 2004](#); [Strange et al., 1983](#)), and algorithms that enhance speech transients given prior knowledge of the speech content have been successful in increasing recognition rates for speech in noise ([Yoo et al., 2007](#)). Such algorithms cannot recover non-voiced components of speech such as fricatives, and these components must be deduced from the voiced information when masked by noise.

This paper introduces a neurobiologically inspired model of periodicity processing (Periodic Sampling Model, PSM) that may enhance the loudness of transient formants in speech relative to noise without the long signal durations required to extract fundamental frequency information. Overall, the PSM proposes that inferior colliculus spike rates integrated over multiple waveform periods after each sound onset will be greater at specific sub-bands (or best modulation frequencies, BMF) within auditory frequency channels that contain periodic auditory nerve (AN) spike rate modulations. The specific loudness associated with amplitude modulated filter channels will be greater, in keeping with the relationship between pitch strength and loudness observed by Fastl (1989). The second part of the paper reports a pilot study in which an algorithm based on the PSM is used to enhance periodic components of speech signals embedded in pink noise and speech babble, and speech intelligibility of processed speech stimuli were compared with unprocessed stimuli for normal hearing participants.

2. The Periodic Sampling Model (PSM)

2.1 Neurobiological Motivation

This model focusses on brainstem networks that involve the cochlear nucleus (CN), inferior colliculus (IC), and the ventral nucleus of the lateral lemniscus (VLL). The cell types that project to the IC from the CN comprise multipolar (otherwise known as stellate or chopper), fusiform, octopus, and giant cells (Cant and Benson, 2003). Sustained chopper cells exhibit highly regular firing patterns when excited near their characteristic frequency (CF), with a frequency of firing known as the cell's best modulation frequency (BMF; Wiegand and Winter, 2001) that is independent of the stimulus frequency (Blackburn and Sachs, 1991).

Octopus, bushy and chopper cells in the CN often synapse to the VLL, which then projects to the IC. Octopus cells are innervated by a wide range of AN fibers and so integrate waveform information from many critical bands (around 2-3 octaves in mice, Oertel et al., 2000). They have

very short latencies, and a single cell can phase-lock to acoustic waveforms at frequencies up to around 800 Hz (Ferragamo and Oertel, 2002). Bushy cells have similar temporal response properties to Octopus cells, but are innervated by only a few AN fibers (Oertel, 1997). Neurons in the VLL generally have short temporal integration times (around 5 ms), broad tuning, and exhibit a wide range of response types (Covey and Casseday, 1991). The combined excitatory and inhibitory input to the central nucleus of the IC from the VLL is likely to contain precise waveform timing information that is superimposed on direct inputs from sustained chopper cells in the CN (see Fig. 1; Riquelme et al., 2001; Schofield, 2005; Pickles, 2008).

Neurons in the central nucleus of the IC are arranged in a three-dimensional map with fine and coarse frequency resolution in two dimensions and minimum intensity thresholds in the third dimension (Ehret and Schreiner, 2005). Frequency-band laminae contain neurons that are tuned to characteristic frequencies within a critical bandwidth, and so the array of laminae comprises the coarse, tonotopic frequency dimension. Around the circumference of each lamina, the neurons are arranged according to their best modulation frequencies (BMF), which range from around 10 Hz to 600 Hz and represent a place-rate code for periodicity (Langner et al., 2002). Each lamina also exhibits concentrically arranged regions of neurons with similar minimum intensity thresholds that decrease toward the center of the lamina by around 60 dB. This may represent a place code for intensity within each critical band (or specific loudness) that reflects a similar distribution of minimum thresholds for auditory nerves (Stiebler, 1986). Each auditory nerve ganglion cell has a monotonic rate level function above its minimum threshold over about 30-70 dB (Smith and Zwislocki, 1975), suggesting that stimulus intensity may be encoded using both rate and place information in the IC to cover the full 130 dB range of human hearing.

<INSERT FIG 1 ABOUT HERE>

The PSM proposes a mechanism that could generate a rate-place code for signal amplitude in the IC that is enhanced by the periodicity of AN responses to the stimulus (Fig. 1). In this model, the AN innervates both sustained chopper cells (shown on the left side of Fig. 1) and bushy cells (shown on the right side of Fig. 1). Phase locked spike rate representations of the AN response at each CF generated by bushy cells are relayed through the VLL before being superimposed with the chopper responses in the inputs to IC neurons. Both the bushy and chopper cell responses commence at the stimulus onset, but only chopper cells with BMF at unitary or harmonic frequency relationships to the AN spike rate modulations will remain in phase with the bushy cell responses. Integrate-and-fire cells in the tonotopic laminae in the central nucleus of the IC (shown at centre top of Fig.1) are more likely to generate action potentials when spikes from these two sources coincide.

Sustained chopper cells in the CN have well defined BMFs over a range of modulation frequencies up to 500 Hz over a wide range of CFs ([Blackburn and Sachs, 1991](#); [Blackburn and Sachs, 1992](#); [Frisina et al., 1990](#); [Oertel, 1985](#); [Wiegube and Winter, 2001](#)). For simplicity, the chopper cells shown in Fig. 1 all have a CF of 250 Hz, but it is proposed that many arrays of chopper cells with varying BMF occur at CFs similar to those of the tonotopic laminae of the IC. The spike rates of IC neurons will be consistently higher in IC laminae in which the chopper BMF matches the AN modulation frequency or its sub-harmonics due to the arrival of coincident spikes from both pathways. This effect will be further enhanced by the tendency for chopper cells to fire more regularly when innervated by spike trains that are amplitude modulated at frequencies close to their BMF ([Blackburn and Sachs, 1992](#); [Wiegube and Meddis, 2004](#)). Competitive lateral inhibition within iso-frequency laminae can lead to a single dominant BMF response and a sharpened frequency representation in the IC ([McLachlan, 2009](#)). Sound onsets are detected by Octopus cells that are enervated by a broad range of AN fibers. These cells project largely inhibitory enervation to the VLL ([Cant and Benson, 2003](#)), possibly resulting in the resetting of spike rate integration in the IC after each new sound onset.

McLachlan and Wilson (2010) postulated that sound recognition initiates early in auditory processing. Temporal sequences of IC neural activation patterns could capture both the spectral and temporal dynamics of sound over the first approximately 300 ms after onset (McLachlan and Wilson, 2010). Hebbian learning would then enable the evolution of knowledge representations as long-term memory templates of commonly occurring sounds (spectrotemporal patterns of neural activity) that are associated with verbal labels and multi-modal representations of sound identities (McLachlan and Wilson, 2010). In keeping with Patterson *et al.* (1995), the comparison of these templates with signal representations must be closely time-locked to the detection of onsets in the auditory nerve spike rates. Since recognition mechanisms are independent of loudness, they are likely to use saturated, non-monotonic spike rate information that has been observed for neurons in the IC of cats and mice (Ehret and Schreiner, 2005; McLachlan, 2011). Auditory nerve spike rates exhibit strong phase locking to frequencies up to 2 kHz which declines up to 5 kHz (Hu and Wang, 2008), so periodicity could be used to enhance the salience of speech formants. This would allow attention to shift to neural outputs at higher activation thresholds that are less likely to be contaminated by environmental noise, which may lead to better speech recognition.

2.2 Algorithmic implementation of the PSM

Inspired by the PSM (Fig. 1), the proposed algorithm determines the dominant periodicity in each filter channel after each speech onset and applies a gain to the filter amplitude proportional to the computed degree of periodicity over time. An output signal is then generated by combining the channel outputs at their relative amplitudes. This enables the algorithm to respond selectively to components of signals with strong and stable periodic waveforms. The algorithm is outlined in Fig. 2 and the text below, and described in detail in Appendix A.

<INSERT FIG 2 ABOUT HERE>

2.2.1. Filter and rectify

A 100-channel Gammatone filter bank with center frequencies between 80 Hz and 5,500 Hz was applied to speech examples to approximate the frequency transform performed in the cochlea (Moore and Glasberg, 1983). The outputs of the filter channels were full-wave rectified. Full-wave rectification enables periodic information to be accumulated more rapidly than is possible in the human auditory system, and will reduce processing delays in future real-time implementations.

2.2.2. Detect onsets

Speech onsets were used to lock the algorithm to the phase with the signal temporal fine structure. They were detected in the filter bank outputs by approximating the behavior of CN octopus cells, which are very sensitive onset detectors. Since octopus cells have very short latencies and are innervated over a wide range of AN fibers (Ferragamo and Oertel, 2002), the rectified filter outputs were differentiated through time and then integrated over consecutive 1 ms time windows and across 10 filter channels (approximately corresponding to an octave band filter).

A new speech onset was defined when the magnitude of any of the modeled octopus cell activation values reached a user-defined threshold, and the filter outputs were set to zero for the next 30 ms to prevent further onsets associated with the signal's temporal fine structure. The threshold was dynamically set to 10% above the average background noise level in each speech trial using the noise level before speech commenced. Each new onset event initiated an analysis frame for periodicity processing, and new onsets were artificially created if the duration between detected onsets was greater than 200 ms so that analysis frames didn't become too long for the available computer memory.

2.2.3. *Re-sampling*

Coincidence of chopper and waveform spike trains will greatly increase the probability that IC neurons will fire. This can be likened to chopper neurons inducing IC neurons to sample the waveform amplitude at the chopper neuron's BMF. In this algorithm, the rectified filter channel outputs after each speech onset were re-sampled at a range of sampling periods consistent with the BMF range of sustained chopper cells. The number of samples in each re-sampling period was varied from 28 to 55 to create an array of re-sampling frequencies between 300 and 580 Hz. Since the filter outputs have been full-wave rectified, this frequency range corresponds to stimulus frequencies of 150 to 290 Hz. The filter outputs were re-sampled for up to 10 periods after each onset. This is likely to be the number of periods that can be integrated by the auditory system since just noticeable differences of pitch reach their minimum in people without absolute pitch after 10 periods (Hsieh & Saberi, 2007).

2.2.4. *Compute periodicity*

The amplitudes of the re-sampled filter outputs were then passed through a Heaviside step function with a threshold set to 0.15 of the maximum amplitude of the rectified filter outputs, and the number of spikes was counted in each filter channel. The value of this threshold was found to provide good discrimination between periodic and noise inputs without loss of signal information; the value was found by iteration over the range of 0.05 to 0.5 with a small subset of the speech in noise example sound files. Spike counts accumulated more quickly for signals with greater periodicity since the re-sampling period remained in phase with the rectified filter output.

2.2.5. *Compute filter gains*

The gain applied to each filter channel over time was equal to the spike count at the re-sampling frequency with the maximum spike count in each analysis frame after each speech onset. Filter

gains were set to zero at each onset and then increased with the accumulation of spikes until the maximum number of 10 periods was reached. Linear offset ramps of 30 ms length were then applied to the filter gains at each new onset to smooth filter transitions and prevent click artifacts. Note that this approach is only suitable for off-line processing due to the time required to determine the re-sampling frequencies with the maximum spike counts in each channel. For real-time applications, the first re-sampling frequency to produce a spike count greater than all other frequencies could be used to minimize this temporal delay.

2.2.6. Filter signal and output

The original Gammatone filter outputs were then multiplied by the filter gains for each channel and the signal was reconstructed by summing the filter outputs. DC offsets caused by the signal reconstruction were subtracted and then the computed signal was normalized relative to the peak amplitude. Since the PSM algorithm removed fricative and impulsive speech components, the filtered signal could not simply replace the original signal without substantially reducing speech intelligibility. The PSM proposes that the neural mechanisms approximated by this algorithm enhance the salience of the periodic components of the signal, rather than remove all non-periodic components. Therefore, the original unfiltered signal at 20% amplitude was added to the filter output, normalized again relative to the peak amplitude, and then written in the *wav* file format for reproduction. Normalization was undertaken relative to peak amplitudes because they were due to the periodic components of both the filtered and unfiltered signals, and so was less affected by the filter. Reduction of the peak amplitude of the original noisy signal to 20% reduced the SNR by about 14 dB (since $20 \log_{10}(0.2) = 13.98$). This is greater than the 10 dB enhancement of loudness due to pure tones embedded in noise reported by Kryter and Pearsons (1965). Normalization would not be required (or indeed possible) in a real-time implementation of the algorithm in which the filter gains would be added to the input gain and scaled to prevent clipping of the output device.

2.3. Method

The algorithm was implemented using the mathematical software package Matlab on a desktop computer. The algorithm was tested using pre-recorded spoken sentences from the BKB library (Bench and Bamford, 1979) that were mixed with varying levels of pink noise and speech babble and normalized relative to the peak amplitude. The user-defined variables used to generate the data presented in this paper are shown in each processing step in Appendix A. For simplicity, the algorithm is described for a sampling rate of 11,025 Hz. Other sampling rates could be used by adjusting the variables shown in Appendix A accordingly, although lower sample rates reduce the frequency range of the Gammatone filterbank and the minimum possible resolution of re-sampling frequencies. Increasing the frequency resolution of the Gammatone filter bank and the re-sampling step resulted in improved signal quality but at the expense of computer memory. Increasing the values of the onset detection threshold and the re-sampling threshold made the filters less sensitive to noise, but this was at the expense of the occasional loss of voiced information.

2.4. Evaluation of the PSM algorithm

In this section, the performance of the algorithm will be explained for a specific example of male speech in pink noise. Panel A in Fig. 3 shows a spectrogram of the sentence “A cat sits on the bed” spoken by a male speaker in the presence of pink noise at -2 dB SNR, compared to the outputs of the algorithm for the same signal in panel B. The outputs of the onset detection algorithm are shown as a series of vertical lines in panel C. Onsets can be observed at the beginning of each word, when the spectral content changes in the word, and at the end of each word. One interpolated onset can also be observed at around 0.65 seconds since no new voice onsets were present in the signal for 0.2 seconds after the last offset at around 0.45 seconds. In many instances near the end of the word, the

onset occurs before all the spectral amplitudes have decayed due to the offset ramps applied to the filter gains corresponding to each analysis frame.

<INSERT FIG 3 ABOUT HERE>

Comparison of the top and lower panels of Fig. 3 indicates that almost all of the broadband noise has been removed from the signal without removing the prominent speech components that are visible in the top panel. Even more promising is the preservation of spectral glides (speech transients) that are present in the voiced signal between 1 and 1.4 seconds. This is achieved by the generation of new analysis frames by detection of any changes of spectral content since either positive or negative change in individual channels contributes to onset detection. The glides are then interpolated by the addition of the 30 ms offset ramp to the spectral filters. This is particularly evident in the large number of analysis frames created between 1.3 and 1.5 seconds when the spectral content is rapidly changing.

<INSERT FIG 4 ABOUT HERE>

Fig. 4 shows the amplitudes of the re-sampled signal in the 581 Hz filter channel after the onset of the word “sits” that occurs at about 0.75 seconds in Fig. 3. The re-sampling frequencies ranged from 300 to 580 Hz and were closer together at the low end of the frequency range since the sampling periods were equally spaced integer factors of the initial sampling rate. The best frequency response at 394 Hz is evident from consistently high amplitudes at each sampling period. The accumulated spike count at this frequency exceeded all other re-sampling frequencies by the 5th spike, at around 18 ms after onset. This was typical of other vowel onsets, and indicates that the optimal re-sampling frequency could be determined within 20 ms of the voice onset, fast enough to boost the auditory systems sensitivity to the onset transients of voiced speech that are important for

phoneme recognition ([Richardson et al., 2004](#); [Howell and Rosen, 1983](#)). Since the filter outputs were full-wave rectified, the best frequency of 394 Hz represents periodic sampling at the 3rd sub-harmonic (197 Hz) of a 591 Hz signal. This was the 6th harmonic of the voiced signal.

Fig. 5 shows the filter gains computed by counting the number of times the re-sampled filter output exceeded the spiking threshold at the best re-sampling frequency. High gains are observed in the spectro-temporal regions corresponding to voiced speech. However, the maximum gain is not always at the same frequency as the maximum output of the algorithm at any given time (lower panel of Fig. 3) because the filter gain simply reflects the presence of a periodic component at amplitudes above the spike threshold. The final amplitude of the algorithm is the product of the original signal amplitude and the filter gain and so more closely resembles the original speech input.

<INSERT FIG 5 ABOUT HERE>

2.5 Discussion

The algorithm described in this paper differs from previous research in which periodicity information was used to segregate vowels. Autocorrelation-like pitch models have been optimized to estimate concurrent pitch heights at close to 100% accuracy ([Assmann and Summerfield, 1990](#); [de Cheveigné and Kawahara, 1999](#); [Meddis and Hewitt, 1992](#)), despite behavioral research in which correct identification of two vowels has not been found to exceed 75% ([Assmann and Summerfield, 1990](#); [Alain, 2007](#)). [de Cheveigné and Kawahara \(1999\)](#) showed that in principle, with sufficient time and computational resources, all concurrent pitches with distinct periodicities could be determined by iterative cancellation of periodicity information using autocorrelation-like algorithms; however, [Robinson and Patterson \(1995\)](#) suggested that these approaches lack ecological validity because of the high number of waveform cycles and the associated signal durations required to refine pitch judgments.

McLachlan (2011) showed that spectral template matching of IC neural activation patterns after lateral inhibition more closely replicated the Assmann and Summerfield (1990) data for vowel segregation than autocorrelation based algorithms (Assmann and Summerfield, 1990; de Cheveigné and Kawahara, 1999; Meddis and Hewitt, 1992). Furthermore spectral template matching could be achieved within 20-30 ms of a speech onset, which is within the time required to segregate speech phonemes. The current algorithm is consistent with the McLachlan (2011) model of sound recognition, and the sensitivity of speech recognition to transient voiced information (Hazen and Simpson, 1998; Richardson et al., 2004; Strange et al., 1983; Yoo et al., 2007), since periodicity based enhancement of IC spike rates could occur within 20 ms of a vowel onset, and could follow frequency variations of formants associated with speech transients over the width of the respective filter channel (e.g. over about 130 Hz for a 1 kHz component) that occur over 30 ms, the minimum duration of one analysis frame (e.g. the frequency glide at around 1.4 s in Fig. 2).

Full wave rectification allowed the algorithm to respond faster than would be possible in the auditory system, and ensured that the maxima of the differentiated filter outputs used in onset detection were most likely to occur after zero crossings in the filter outputs. The performance of the algorithm indicates that onset detection was adequate to ensure approximate alignment of subsequent re-sampling with waveform maxima across the filter array. Onset detection was likely to be triggered by higher frequency bands in which the waveform maxima occurs earlier than in lower frequency bands, and within the 1 ms integration window used to compute onset detector values. Since filter gain was defined by the number of re-sampled amplitudes that exceeded a threshold, rather than actual re-sampled amplitude, the algorithm appears to be quite robust to phase differences that may occur between the time of sampling and waveform maxima in lower frequency filter channels due to longer lag times and waveform periods. More accurate alignment of re-sampling with waveform peaks would enable greater discrimination of periodic signals from noise

and might be achieved in future implementations by lagging higher frequency channels to align the waveform maxima (Herzke and Hohmann, 2007; Patterson *et al.*, 1995).

3. Assessment of the effect of the PSM algorithm on speech intelligibility

The PSM algorithm was used to reduce the amplitude of non-tonal signal components by 14 dB, which is greater than the 10 dB SPL increase in the perceived loudness of tonal signals reported by Kryter and Pearsons (1965). Since the algorithm replicates normal hearing mechanisms it is unlikely to improve speech intelligibility for healthy listeners; so it is hypothesized that there will be no loss of intelligibility in the processed examples, despite the reduction of the amplitude of non-periodic components of the signal.

3.1 Method

To measure the performance of the PSM algorithm four to six word sentences were randomly selected from the BKB library (Bench and Bamford, 1979). Two noise conditions (pink noise and speech babble) at three signal-to-noise ratios (SNRs) were added to equal numbers of the male and female speaker sentences (Table 1). Pink noise has a power spectral density that is inversely proportional to frequency, reducing by 3dB per octave, and is commonly experienced in environmental conditions. The speech babble sound was randomly sampled from a recording of 4-speaker babble (Auditec, St. Louis, MO, USA). One half of these stimuli were processed by the algorithm and the processed sound files were saved to computer memory. Twelve stimuli by male and female speakers, in each of the six noise conditions, and with or without processing by the algorithm, were presented to each participant (12 sentences x 2 genders x 6 noise condition x 2 processing conditions = 288 stimuli in all). In each condition, two of these stimuli were used as practice stimuli, leaving 240 stimuli that were used in the analysis.

<INSERT TABLE 1 ABOUT HERE>

Seven participants (3 male, 20-50 years old) listened to all 288 stimuli produced by an *Alesis M1* audio monitor placed 1 m directly in front of them at head height when seated in an anechoic chamber. The stimuli were presented at 70 ± 1 dB A-weighted average SPL. All participants were native English speakers and none reported having any hearing difficulty or neurological conditions. The stimuli were presented in short blocks of 12 stimuli. Each block comprised all the stimuli for each experimental condition and the first two stimuli of each block were used for practice with feedback. The blocks were presented in random order and each block required approximately 3 minutes to complete, with a short break of 1-2 minutes between each block. Overall, each experimental session lasted approximately 60 minutes. The number of correct, pre-defined key words that the participant was able to verbally repeat from each sentence was scored by the researcher (3-4 key words per sentence), and the percentage of correct key words was computed for each experimental condition across all participants. The study was conducted in accordance with ethical guidelines established by the University of Melbourne and written informed consent was obtained from all participants.

3.2 Results

Fig. 6 compares word recognition rates for male and female speakers at various SNR with pink noise and 4-speaker speech babble, with and without processing by the PSM algorithm. The data for pink noise show similar loss of word intelligibility as the SNR decreases for all experimental conditions. However, the word recognition rates for the female speaker at SNR of 0 and -2 dB speech babble are lower for the processed stimuli than the unprocessed stimuli. Word recognition rates were close to 100% at the highest SNR in both conditions.

<INSERT FIG 6 ABOUT HERE>

To investigate the differences between processed and unprocessed signals, 3-way repeated measures Analysis of Variance (ANOVA) were run on the word recognition rates for pink noise and speech babble with independent variables of SNR (3 levels), processing (2 levels) and speaker gender (2 levels). All the data met tests of normality and equality of variance. No statistical differences were observed between the word recognition rates for processed and unprocessed signals in any of the conditions for pink noise. However, a 3-way interaction between processing, SNR, and gender ($F(2) = 6.56, p < .05$), and main effects of SNR ($F(2) = 213.33, p < .001$) and processing ($F(2) = 28.53, p < .01$) were observed for the speech babble conditions, indicating that word recognition rates for the female speaker at SNR of 0 and -2 dB speech babble were significantly lower.

3.3 Discussion

Increasing the amplitude of speech by around 14 dB with the re-synthesized voiced components extracted by the PSM algorithm did not affect word intelligibility of the male speaker over a range of SNRs down to -4 dB for normal hearing participants. The results show that the algorithm did not decrease word intelligibility of male or female speakers for normal listeners in the pink noise condition, and provide only partial support for the same in speech babble at higher SNRs due to poorer performance for the female speaker.

The PSM algorithm was able to remove 14 dB of speech babble without affecting the word intelligibility of a single male speaker at SNRs as low as -2 dB. This is due to the sensitivity of the algorithm to slower rates of change of the frequency components of single speaker speech compared to speech babble. However, it should be noted that even at -2 dB SNR the target speaker was always louder than any of the individual speakers in the babble. This allowed the algorithm to detect and amplify the target speech components when they occurred in the stimulus.

For the female speaker, over 50% of the word onsets were not detected by the PSM algorithm in 25% of the stimuli with background levels of speech babble equal to or below 0 dB SNR (the two conditions in which speech intelligibility deteriorated). Word recognition rates were all zero for these stimuli. Furthermore, in some stimuli the last word was not detected by the algorithm due to reduced emphasis on this word by the speaker. These problems only occurred in these conditions and substantially contributed to the poorer word intelligibility results. Decreasing the threshold for onset detection increased the number of word onsets detected in these conditions, but caused many spurious onsets to be created in other experimental conditions that degraded the speech quality. In this experiment, all input variables of the algorithm were held constant across all conditions to enable comparison of the algorithms performance; however, in real-world contexts, the onset detection variable could be varied by the user to suit conditions and the algorithm could be expected to perform as well for female as male speakers in speech babble.

4. General Discussion

Overall, the results confirm that the PSM algorithm could segregate periodic from non-periodic components of a signal within the time required to boost the salience of spoken phonemes. This confirms that it is possible to enhance the tonal components of speech and other animal vocalizations by periodicity processing in the brain stem.

The observation that word intelligibility for normal hearing listeners was not reduced by the PSM algorithm under most of the experimental conditions shows that auditory components of speech that are essential for recognition were not greatly distorted by the algorithm. Significant differences in the word intelligibility between processed and unprocessed stimuli may have emerged with a larger number of participants, although the present analysis of the data suggests that these differences would not be great if the problem of onset detection for female speakers was resolved.

Failure of the PSM algorithm to improve speech intelligibility for these listeners may be due to the ability of healthy listeners to undertake the same signal processing that was undertaken by the algorithm. Since the algorithm responded more strongly to lower frequency voiced components than formants, improvements in speech intelligibility for normal listeners may require a stronger weighting of the outputs of filter channels in the regions of the frequencies of the three main speech formants than was employed in this study (Yoo et al., 2007). Finally, the algorithm cannot enhance non-voiced components of speech and, by amplifying voiced components, it distorts envelopes of voiced components over the first 20-30 ms. These factors may have limited speech intelligibility gains, particularly when listeners first encounter the processed speech. However it should be noted that other noise reduction strategies currently employed in hearing aids have not been able to increase speech intelligibility (Brons et al., 2013).

The PSM algorithm may provide greater benefits to impaired auditory systems that have poorer frequency resolution or limited access to periodicity information (such as cochlear implants). Furthermore, reduction of noise by this algorithm may reduce the discomfort associated with hearing prostheses in noisy environments. Implementation for real-time processing in hearing prostheses will require a new method of selecting the frequency of re-sampling in each channel (Section 2.2.5) and for selecting signal portions to include in the long term averages of background noise used to set onset thresholds. The latter might be achieved by the user resetting the background when it changes and no speech is present, or by computational identification of these conditions from periodicity and spectral average data generated by the algorithm. Finally, in real-time processing the filter gains could be modulated by periodicity information without the need to mix processed and unprocessed signals. The algorithm will then need to be compared with others that use enhancement of pitch perception to improve speech perception performance for CI recipients (Vandali and van Hoesel, 2011; Milczynski et al., 2012).

Animals are more likely to recognize important vocal signals that are embedded in environmental noise if they have evolved enhanced sensitivity to tonal sounds due to brainstem periodicity processing. However, it is difficult to find an evolutionary advantage for animals to possess fine pitch perception due to periodicity processing. Even among humans, the frequency information required to recognize phonemes, and the pitch variations associated with speech prosody, are greater than the resolution of the auditory nerve ([Patel et al., 2006](#)). These frequency variations can be rapidly detected by changes in spike rates across the AN without reliance on periodicity processing. The pitch matching accuracy of non-musicians for pure tones is often about the resolution of AN spectral processing, but improves rapidly with training ([McLachlan et al., 2013](#); [Hutchins and Peretz, 2011](#)), so it is likely that fine pitch perception based on periodicity cues is a learnt skill. Furthermore, increased congruence of brainstem EEG recordings with waveform periodicity has been observed for the timbre of the instrument that musicians have learnt ([Strait et al., 2012](#)). In the context of the PSM, the observation that fine pitch processing based on periodicity is enhanced by training points to pitch being a form of loudness perception at highly specific frequencies that are associated with specific timbres and cultural traditions through music training ([McLachlan, 2009](#); [McLachlan, 2011](#); [McLachlan and Wilson, 2010](#); [McLachlan et al., 2012](#)). This suggests that periodicity processing in the brain stem evolved to enhance the salience of vocalizations, rather than for processing pitch at resolutions finer than the resolution of the AN.

Acknowledgements

This work was supported by Australian Research Council Discovery Project grants DP1094830 and DP120103039.

Appendix A. The algorithm and user variables used in this study.

The implementation is described for a sample rate of 11,025 Hz. If other sampling rates are used, the number of samples used in the temporal integration windows would need to be adjusted accordingly.

1. Filter and rectification

A 4th order Gammatone filter bank is defined in the time domain (sample number, m) by:

$$GT(m) = am^3 e^{-2\pi f_c m} \cos(2\pi f_c m)$$

where the filter center frequencies, f_c (Hz), are defined by the number of channels, N , uniformly distributed between the highest and lowest frequency channels (f_{max} and f_{min} , respectively) along the ERB scale according to:

$$f_c(n) = f_{max} \exp\left(\frac{n}{N} (\log(f_{min} + k) - \log(f_{max} + k))\right) - k$$

where $k = 228.83$ derived from parameters of best fit with human auditory tuning curves (Moore and Glasberg, 1983; Slaney, 1993), $N = 100$ channels, $f_{min} = 80$ Hz, and $f_{max} = 5,500$ Hz.

The bandwidth of channel n is defined by:

$$f(n) = 0.11 f_c(n) + 25.2$$

The rectified filter output $y_r(n, m)$ for input signal $x(m)$ is computed by:

$$y_r(n, m) = \text{abs}(GT(t) * x(m))$$

where $*$ denotes the convolution operator, and $\text{abs}()$ returns the absolute value.

2. Onset detection

The onset detector y_o is computed by:

$$y_o(n, m) = \sum_{n'=n}^{n+w_n} \sum_{m'=m+1}^{m+w_m} ((y_r(n', m') - y_r(n', m'-1))) \quad 1$$

where $n \leq N - w_n$ and $m \leq M - w_m$. M is the total number of samples in each recording, $w_n = 10$ channels, and $w_m = 11$ samples (~ 1 ms).

An onset threshold, T_o is dynamically set for each sentence in relation to the mean amplitude in each recording before speech commenced by:

$$T_o(n) = \frac{d^2}{w_{noise}} \sum_{n'=n}^{n+w_n} \sum_{m=1}^{w_{noise}} y_r(n', m) \quad 2$$

where $d = 1.1$ and $w_{noise} = 5,500$ samples (~ 500 ms).

If, for any n , $y_o(n, m) \geq T_o(n)$, an onset flag is set, $F(m) = 1$.

To prevent multiple onsets associated with the stimulus fine temporal structure (such as individual glottal pulses), onsets within a fixed refractory time, w_o , after a first onset are discounted:

$$y_o(n, m; m + w_o) = 0 \quad 3$$

where $w_o = 330$ samples (~ 30 ms).

To speed the re-sampling step, an array of onset sample numbers, O , indexed by i is generated by:

$$O(i) = m, \text{ if } F(m) = 1 \quad 4$$

Finally, if $O(i) - O(i-1) \geq b$, then $O(i) = O(i-1) + b$, where the buffer length is, $b = 2205$ samples (200 ms).

3. Compute coincidence detection

The rectified filter outputs, $y_r(n, m)$ are copied into a three dimensional array $y_{rs}(n, m, j)$ that is initialized to zero, where j is a range of re-sampling frequencies for a maximum of 10 periods, p after each onset, i .

for each channel n

for each onset i

for each re-sampling period $j = 37-19$ samples (equivalent to 300-580 Hz)

for each period p , provided $O(i) + j/p \leq O(i+1)$

$$y_{rs}(n, O(i) + j/p, j - j_{min} + 1) = y_r(n, O(i) + j/p) \quad 5$$

where $j_{min} = 28$ samples.

4. Periodicity computation

The amplitudes of y_{rs} are saturated when they exceed a threshold:

$$y_s(n, m, j) = H(y_{rs}(n, m, j) - T_s) \quad 6$$

where $H(\cdot)$ is the Heaviside step function and $T_s = 0.15 \max(y_{rs})$.

The saturated output is accumulated over the analysis frame after each onset:

$$y_a(n, m, j) = \sum_{i=O(i)}^{O(i+1)-1} y_s(n, m, j), \text{ where } y_a(n, O(i), j) = 0 \quad 7$$

5. Detect dominant BMF and compute filter gains

Take y_a with the maximum amplitude in each analysis frame as the filter gain:

$$g(n, m) = y_a(n, m, j_{ma}) \quad 8$$

where j_{ma} is the sampling period for which y_a is maximal.

Smooth the filter gain transition at the end of each analysis frame with a linear offset ramp:

$$g_r(n, m) = g(n, m) + g(n, (Q(l) - 1)) \frac{(w_r - (m - Q(l)))}{w_r}$$

9

for $0 \leq m - Q(l) \leq w_r$, and $w_r = 330$ samples (30 ms).

6. Filter gain implementation and signal resynthesis

The Gammatone filtered signal is multiplied by the computed filter gains and the output (S_{out}) is computed by summing the filter outputs:

$$S_{out}(m) = \sum_n g_r(n, m) y_g(n, m).$$

10

References

- Alain, C., 2007. Breaking the wave: Effects of attention and learning on concurrent sound perception, *Hear. Res.* 228, 225-236.
- Assmann, P. F., Summerfield, Q., 1990. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies, *J. Acoust. Soc. Am.* 88, 680-697.
- Beerends, J. G., Houtsma, A. J. M., 1989. Pitch identification of simultaneous diotic and dichotic two-tone complexes. *J. Acoust. Soc. Am.* 85, 813-819.
- Bench, J., Bamford, J., 1979. *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children.* Academic Press, Orlando, pp 481-486.
- Blackburn, C. C., Sachs, M. B., 1991. Regularity analysis in a compartmental model of chopper units in the anteroventral cochlear nucleus. *J. Neurophysiol.* 65, 606-629.
- Blackburn, C. C., Sachs, M. B., 1992. Effects of off-BF tones on responses of chopper units in ventral cochlear nucleus; I. Regularity and temporal adaptation patterns. *J. Neurophysiol.*, 68, 124-143.
- Brons, I., Houber, R., Dreschler, W. A., 2013. Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear Hear.* 34, 29-41.
- Cant, N. B., Benson, C. G., 2003. Parallel auditory pathways: projection patterns of the different neuronal populations in the dorsal and ventral cochlear nuclei. *Brain Res. Bull.*, 60, 457-474.

Cariani, P. A., 2001. Neural timing nets. *Neural Networks*, 14, 737-753.

Cheng, Y. M., O'Shaughnessy, D., 1991. Speech enhancement based conceptually on auditory evidence. *IEEE Trans. Signal Process.* 39, 1943-1954.

Covey, E., Casseday, J. H., 1991. The monaural nuclei of the lateral lemniscus in an echolocating bat: parallel pathways for analyzing temporal features of sound. *J. Neurosci.* 11, 3456-3470.

Dau, T., Kollmeier, B., Kohlrausch, A., 1997. Modeling auditory processing of amplitude modulation. I. Detection and Masking with narrow-band carriers. *J. Acoust. Soc. Am.* 102, 2892-2905.

de Cheveigné, A., Kawahara, H., 1999. Multiple period estimation and pitch perception model, *Speech Comm.* 27, 175-185.

Dicke, U., Ewert, S. D., Dau, T., Kollmeier, B., 2007. A neural circuit transforming temporal periodicity information into a rate-based representation in the mammalian auditory system. *J. Acoust. Soc. Am.* 121, 310-326.

Ehret, G., Schreiner, C. E., 2005. Spectral and intensity coding. In *The Inferior Colliculus*. Springer, New York, pp. 319-345.

Fastl, H., Stoll, G., 1979. Scaling of pitch strength. *Hear. Res.*, 1, 293-301.

Fastl, H., 1989. Pitch strength of pure tones. Proceedings of the 13th International Congress on Acoustics, pp. 11-14.

Ferragamo, M. J., Oertel, D., 2002. Octopus cells of the mammalian ventral cochlear nucleus sense the rate of depolarization. *J. Neurophysiol.* 87, 2262–2270.

Frisina, R., D., Smith, R., I., Chamberlain, S., C., 1990. Encoding of amplitude modulation in the gerbil cochlear nucleus. I. A hierarchy of enhancement. *Hear. Res.*, 44, 90-122.

Gai, Y., Carney, L. H., 2006. Temporal measures and neural strategies for detection of tones in noise based on responses in anteroventral cochlear nucleus. *J. Neurophysiol.*, 96, 2451–2464.

Guérin, A., Le Bouquin Jeannès, R., Bès, J., Faucon, G., Lorenzi, C., 2006. Evaluation of two computational models of amplitude modulation coding in the inferior colliculus. *Hear. Res.* 211, 54-62.

Hazan, V., Simpson, A. 1998. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Commun.* 24, 211-226.

Herzke, T., Hohmann, V. (2007). Improved numerical methods for Gammatone filterbank analysis and synthesis. *Acta Acust.* 93, 498-500.

Howell, P., Rosen, S., 1983. Perception of rise time and explanations of the affricate/fricative contrast. *Speech Commun.* 2, 164–166.

Hsieh, I-H., Saberi, K., 2007. Temporal integration in absolute pitch identification of absolute pitch, *Hear. Res.* 233, 108-116.

Hu, G., Wang, D., 2008. Segregation of unvoiced speech from nonspeech interference. *J. Acoust. Soc. Am.* 124, 1306-1319.

Hutchins, S. Peretz, I., 2011. A frog in your throat or in your Ear? Searching for the causes of poor singing. *J. Exp. Psych. Gen.* 141, 76-97.

Ishizuka, K., Nakatani, T., 2006. A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition. *Speech Commun.* 48, 1447-1457.

Kidd Jr., G., Mason, C. R., Brantley, M. A., Owen, G. A., 1989. Roving-level tone-in-noise detection. *J. Acoust. Soc. Am.* 38, 106-112.

Krumbholz, K., Patterson, R. D., Seither-Preisler, A., Lammertmann, C., Lütkenhöner, B. 2003. Neuromagnetic evidence for a pitch processing center in Heschl's gyrus. *Cerebral Cortex*, 13, 765–772.

Kryter, K. D., Pearsons, K. S., 1965. Judged noisiness of a band of random noise containing an audible pure tone *J. Acoust. Soc. Am.* 38, 106-112.

Langner, G., 2005. Neuronal mechanisms underlying the perception of pitch and harmony. *Ann. N.Y. Acad. Sci.* 1060, 50–52.

Langner, G., Albert, M., Briede, T., 2002. Temporal and spatial coding of periodicity information in the inferior colliculus of awake chinchilla (*Chinchilla laniger*). *Hear. Res.* 168, 110-130.

McLachlan, N. M., 2009. A computational model of human pitch strength and height judgments. *Hear. Res.* 249, 23-35.

McLachlan, N. M., Wilson, S. J., 2010. The Central Role of Recognition in Auditory Perception: A Neurobiological Model. *Psych. Rev.* 117, 175-196.

McLachlan, N. M., 2011. A neurocognitive model of recognition and pitch segregation. *J. Acoust. Soc. Am.* 130, 2845-2854.

McLachlan, N. M., Marco, D. J. T., Wilson, S. J., 2013. Consonance and Pitch. *J. Exp. Psych. Gen.*, doi: 10.1037/a0030830.

Meddis, R., Hewitt, M. J., 1992. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* 91, 233-245.

Meddis, R., O'Mard, L. P., 2006. Virtual pitch in a computational physiological model. *J. Acoust. Soc. Am.* 120, 3861-3869.

Milczynski, M., Chang, J. E., Wouters, J., van Wieringen, A. 2012. Perception of Mandarin Chinese with cochlear implants using enhanced temporal pitch cues. *Hear. Res.* 285, 1-12.

Moore, B. C. J., Glasberg, B. R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* 74, 750-753.

Nelson, P., C. Carney, L., H., 2004. A phenomenological model of peripheral and central neural responses to amplitude modulated tones. *J. Acoust. Soc. Am.* 116, 2173-2186.

Oertel, D., 1985. Use of brain slices in the study of the auditory system: Spatial and temporal summation of synaptic inputs in cells in the anteroventral cochlear nucleus of the mouse. *J. Acoust. Soc. Am.* 78, 329-333.

Oertel, D., 1997. Encoding of Timing in the Brain Stem Auditory Nuclei of Vertebrates. *Neuron* 19, 959-962.

Oertel, D., Bal, R., Gardner, S. M., Smith, P. H., Joris, P. X., 2000. Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. *Proc. Nat. Acad. Sci. USA.* 97, 11773–11779.

Patel, A. D., Iversen, J. R., Rosenberg, J. C., 2006. Comparing the rhythm and melody of speech and music: The case of British English and French. *J. Acoust. Soc. Am.* 119, 3034-3047.

Patterson, R. D., Allerhand, M. H., Gicùère, C., 1995. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *J. Acoust. Soc. Am.* 98, 1890-1894.

Paolini, A. G., Clarey, J. C., Needham, K., Clark, G. M., 2005. Balanced inhibition and excitation underlies spike firing regularity in ventral cochlear nucleus chopper neurons. *Eur. J. Neurosci.* 21, 1236-1248.

Pickles, J. O., 2008. An introduction to the physiology of hearing, 3rd edition. Emerald: Bingley, pp. 155-183.

Plomp, R., 1994. Noise, amplification, and compression: Considerations of three main issues in hearing aid design. *Ear Hear.* 15, 2–12.

Rakowski, A., 1996. Pitch strength, pitch value, and pitch distance. *Acustica* 82, S80.

Richardson, U., Thomson, J. M., Scott S. K., Goswami U., 2004. Auditory Processing Skills and Phonological Representation in Dyslexic Children. *Dyslexia* 10, 215-233.

Riquelme, R., Saldanã, E., Osen, K. K., Ottersen, O. P., Merchán M. A., 2001. Colocalization of GABA and glycine in the ventral nucleus of the lateral lemniscus in rat: an in situ hybridization and semiquantitative immunocytochemical study. *J. Comp. Neurol.* 432, 409-424.

Robinson, K., Patterson, R. D., 1995. The stimulus duration required to identify vowels, their octave and their pitch chroma. *J. Acoust. Soc. Am.* 98, 1858-1865.

Schofield, B. R., 2005. Olivary and Lemniscal Projections. In *The Inferior Colliculus*. Springer, New York, pp. 140.

Seither-Preisler, A., Patterson, R. D., Krumbholz, K., Seither, S., Lütkenhöner, B., 2006. From noise to pitch: Transient and sustained responses of the auditory evoked field. *Hear. Res.*, 218, 50-63.

Slaney, M., 1993. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. *Apple Computer Technical Report #35*.

Smith, R. L., Zwislocki, J. J., 1975. Short term adaptation and incremental responses in single auditory-nerve fibres. *Biol. Cybernet.* 17, 169-182.

Soeta, Y., Maruo, T., Ando, Y., 2004. Annoyance of bandpass-filtered noises in relation to the factor extracted from autocorrelation function (L). *J. Acoust. Soc. Am.* 116, 3275-3278.

Soeta, Y., Yanai, K., Nakagawa, S., Kotani, K., Horii, K., 2007. Loudness in relation to iterated rippled noise. *J. Sound Vib.* 304, 415-419.

Stiebler, I., 1986. Tone-threshold mapping in the inferior colliculus of the house-mouse. *Neurosci. Lett.* 65, 336-340.

Strait, D. L., Chan, K., Ashley, R., Kraus, N., 2012. Specialization among the specialized: Auditory brainstem function is tuned in to timbre. *Cortex*, 48, 360-362.

Strange, W., Jenkins, J. J., Johnson, T. L. 1983. Dynamic specification of coarticulated vowels. *J. Acoust. Soc. Am.* 74, 695-705.

Turicchia, L., Sarpeshkar, R. 2005. A bio-inspired companding strategy for spectral enhancement. *IEEE Trans. Speech Audio Process.* 13, 243-253.

Vandali, A. E., van Hoesel, R. J. M. 2011. Development of a temporal fundamental frequency coding strategy for cochlear implants. *J. Acoust. Soc. Am.* 129, 4023-4036.

Viemeister, N. F., Wakefield G. H., 1991. Temporal integration and multiple looks. *J. Acoust. Soc. Am.* 90, 858 – 865.

Wiegerebe, L., Meddis, R., 2004. The representation of periodic sounds in simulated sustained chopper units of the ventral cochlear nucleus. *J. Acoust. Soc. Am.*, 115, 1207- 1218.

Wiegerebe, L., Winter, I. M., 2001. Temporal Representation of Iterated Rippled Noise as a Function of Delay and Sound Level in the Ventral Cochlear Nucleus. *J Neurophysiol.* 85, 1206-1219.

Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C., 2007. Speech signal modification to increase intelligibility in noisy environments. *J. Acoust. Soc. Am.*, 122, 1138-1149.

Zwicker, E., Fastl, H., 1999. *Psycho-acoustics Facts and models*, 2nd Edition. Springer-Verlag, Berlin, pp. 111-115.

Fig. 1. Schematic representation of the neurobiological model. CF = characteristic frequency, BMF = best modulation frequency.

Fig. 2. Flow diagram of the computational algorithm. AN = auditory nerve, IC = inferior colliculus, and BMF = best modulation frequency.

Fig. 3. Spectrogram of a male speaker in the presence of pink noise at -2 dB SNR (top), compared to the algorithm output (middle), with the analysis frame onsets (bottom) that were detected or interpolated by the algorithm.

Fig. 4. Re-sampled amplitudes that exceed the amplitude threshold for the analysis frame corresponding to the word “sits”, starting at 0.75 seconds in the 581 Hz filter channel. The re-sampling period of 394 Hz produced the most consistently high amplitudes.

Fig 5. Filter gains for the entire sentence based on the accumulated count of re-sampled amplitudes shown in Fig. 4.

Fig. 6. Comparison of word recognition rates by male and female speakers at various SNR with pink noise and 4-speaker speech babble with and without processing by the algorithm described in Appendix A.

Table 1. Stimulus conditions.

Speaker gender	Noise type	SNR (dB)
Male	Pink noise	-4
		-2
		0
	babble	-2
		0
		2
Female	Pink noise	-4
		-2
		0
	babble	-2
		0
		2

Highlights

- There is a fundamental relationship between pitch strength and loudness.
- These psychoacoustic findings inspired the formulation of a conceptual neurobiological model.
- A computer algorithm to enhance speech in noise was based on this model.
- In most experimental conditions noise was reduced by 14 dB without reducing speech intelligibility for normal hearing participants.
- This model raises the possibility that periodicity processing in the brain stem evolved to boost the salience of animal vocalizations.

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

McLachlan, NM;Grayden, DB

Title:

Enhancement of speech perception in noise by periodicity processing: A neurobiological model and signal processing algorithm

Date:

2014-02-01

Citation:

McLachlan, N. M. & Grayden, D. B. (2014). Enhancement of speech perception in noise by periodicity processing: A neurobiological model and signal processing algorithm. SPEECH COMMUNICATION, 57, pp.114-125. <https://doi.org/10.1016/j.specom.2013.09.007>.

Persistent Link:

<http://hdl.handle.net/11343/44154>