Coding and Decoding Speech using a Biologically Inspired Coding System

1st Madhurananda Pahar *
Digital Signal Processing Lab
Department of Electrical and Electronic Engineering
University of Stellenbosch
Stellenbosch, 7600, South Africa
ORCID: 0000-0002-5926-0144

2nd Leslie S. Smith
Computing Science and Mathematics
University of Stirling
Stirling, UK
l.s.smith@cs.stir.ac.uk

***Abstract*—A spike (event) based sound coding technique has been presented in this study where the spikes are similar to the spikes exhibited by type 1 fibers of the auditory nerve. This lossy coding technique has already been shown useful for inter-aural time difference based sound source direction finding. Here, we show that decoding and resynthesising this code can produce intelligible speech even using a small number of spike trains. We have used few composite techniques including speaker verification to assess the effectiveness of the coding technique on a large number of TIMIT sentences. This biologically inspired coding technique can provide suitable input for a spiking neural network, as well as maintaining the accurate time structure of sound.**

***Index Terms*—spike coding, spike decoding, speech coding, biological inspiration, x-vector, i-vector, speaker diarisation, speaker recognition**

I. Introduction

This study describes an event-based coding, based on animal hearing systems. Most of these translate sound into a set of spike trains (events) transferred to the animal's brain along the type 1 fibers of the auditory nerve (AN). There are many type 1 fibers, and these are tonotopically arranged. Some fibers ("high spontaneous rate") respond to low intensity sounds, saturating (i.e. firing at their maximal rate) with louder sounds, and others ("low spontaneous rate") respond only to louder sounds. There are about 30,000 type 1 AN fibers in humans, and 50,000 in cats. At low frequencies, (up to about 3 kHz) the spikes from these fibers are approximately phase-locked to the incoming signal [1].

Although this type of coding technique is not unique in nature [2], it is particularly suited to processing using spiking neural networks, as it is an event-based spectro-temporal coding. Most neuromorphic silicon cochleas follow a similar method of encoding their output (reviewed in [3]): for each frequency band a number of spike trains are generated, emulating the sensitivity of different type 1 auditory nerve fibers. This study examines the effect of varying the number of frequency bands (bandpass channels) and the number of threshold levels on the quality of the spike coded and then decoded (resynthesised) speech, with a view to understanding the required number of bandpass channels and threshold levels: this becomes important for the design of neuromorphic silicon cochleae.

Similar work has been carried out by [4], where a hierarchical spike code is capable to capture complex structure in music, animal vocalizations and ambient natural sounds. A biologically inspired low-bit-rate spike encoder is also developed in [5] and then improved in [6] and this coding shows that high quality audio can be delivered at between 10 and 21 bits per spike. However, in the work presented here we are interested in minimising the number of spike while maintaining intelligibility.

In earlier work, we used an engineering approximation to the animal code (based on the Gammatone filterbank [7], followed by a simple neural network) to detect onsets in sounds, and to find the location of sound sources by considering inter-aural (inter-microphone) time differences at the onset of the sound [8], [9] and for identifying the type of musical instrument that produced single notes [10]. Considering everything that the brain does with sound must originally be based on these type 1 AN spikes, we became interested in re-creating the original sound from these spike trains.

Clearly, the brain does not decode sound from these spikes, but interprets them. But, by decoding the sound from its spike based code, we can discuss the effectiveness of the spike coding technique by comparing the quality of the decoded sounds with the original sounds and investigating into the intelligibility preserved in the decoded sounds from the original sounds.

Models such as Gammatone filter bank [7] with centre frequencies [11] are built on early auditory processing and they can be used as a basis for speech and sound coding [12].

This type of filter bank [13] has also been used in optimising quantizer for the spike amplitudes while coding audio. It has also helped in minimising computational costs.

The idea of using such biologically inspired codes or representations for resynthesis is considered in [14], [15] and [16] with the latter two using the Gammatone filterbank as front end. Neuromorphic cochleae also use approximately

*The first author performed the work during his PhD at the University of Stirling

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

logarithmically distributed centre frequencies [3]. This study uses a relatively small number of frequency bands, and a small number of threshold levels while coding the audio and after running few tests on the large TIMIT dataset by comparing the original and decoded audio, we confirm that it still manages to provide good quality of decoded audio. This suggests that our biologically inspired spike based coding technique, even when used with small numbers of bands and threshold levels, is capable to code and decode audio which maintains the information required for intelligibility, and is therefore a good candidate as input to an interpreting neural network.

II. Spike coding and decoding

Sounds are coded and decoded using software so that particular parameters (number of bands (N_f), and number of threshold levels (N_J)) can be easily adjusted, and the sound quality assessed.

A. The spike coding technique

The coding technique is described in detail in [8], [9], and shown in figure 1.

Briefly, the incoming sound $s(t)$ is passed through a Gammatone filter bank, to create N_f bandpassed signals, $s_i(t)$, for $i = 1 \dots N_f$, where N_f is the number of bandpassed channels. For each signal $s_i(t)$, N_J spike trains $P_{i,j}$ (with $i \in (1 \dots N_f)$, and $j \in (1 \dots N_J)$) are generated. Each spike marks a positive-going zero-crossing event in $s_i(t)$. The difference between the spike trains indexed by j (for fixed i) is in the level of the signal prior to the zero-crossing. For $j = 1$, a spike is generated when the signal exceeds a minimum (voltage) threshold (here set to 0.0002) in the previous quarter cycle (using the centre frequency of the i'th band to calculate the period). For $j = N_J$, the threshold is 0.0362, which makes the N_J'th band $10 \log_{10}((0.0362/0.0002)^2) = 45.2$ dB less sensitive. When $N_J > 2$, intermediate thresholds are calculated geometrically. There are thus $N_f * N_J$ spike trains, although when a spike is generated in spike train $P_{i,j}$, it is also generated in spike train $P_{i,j'}$ for $j' < j$. In practice, we generate only N_f trains of pulses, with each pulse coded by the maximal value of j for which a spike is generated. We thus store the event trains Q_i for $i = 1 \dots N_f$, with each element of each train being (t_i^k, j_i^k), that is, the time and the maximal j-value of each zero-crossing event, with k indexing the event number.

B. The spike decoding technique

Here, our aim is to generate a signal that could have resulted in the set of pulse trains Q_i. Clearly, this signal is not unique. Each Q_i is processed individually, and then the signals from each channel are summed. Figure 2 explains the decoding technique.

First, the spike trains are delay compensated to take account of the delays introduced by the Gammatone filterbank [17]. Next we consider pairs of consecutive spikes in the i'th channel, at times t_i^{k-1} and t_i^k, where $(t_i^k - t_i^{k-1}) < \frac{2}{f_i}$, where f_i is the centre frequency of the i'th bandpassed channel. For each of these, we generate a single cycle of a sine wave with period $(t_i^k - t_i^{k-1})$, that is, frequency $\frac{1}{(t_i^k - t_i^{k-1})}$. The amplitude of the sine wave depends on j_i^{k-1} and j_i^k. This amplitude value is associated with each j value, linearly based on the threshold levels used in the spike event generation. When $j_i^{k-1} \neq j_i^k$, a linear ramp is used to modulate the sine wave cycle.

Where there is a sequence of consecutive spikes in bandpass channel i, each less than $\frac{2}{f_i}$ apart, we add an extra ramped sine wave at the start and the end, with one end of the ramp set to 0, so that the concatenated set of sine waves starts and ends smoothly at 0. When there is a single isolated spike event, we treat it the same way, so that we generate a pair of sine waves, modulated to start and end at 0, with frequency f_i. These techniques can result in non-linear distortion of the signal: however, we note that the aim of this work is to show that intelligibility is maintained even for small numbers of spike trains, rather than to find the optimal decoding technique.

To create the final decoded signal, we concatenate all the sine waves (putting 0's in between them) for each frequency band with centre frequency f_i, then add up all the signals across the N_f bands. Lastly, we normalise the signal so that the maximal amplitude is in the appropriate range. A more detailed description may be found in chapter 3 of [18].

Clearly this is quite complex: however, we note that since the coding is not necessarily intended for decoding, but for interpretation, and in that case, it does not matter how complex decoding is. That said, we believe the coding to be quite efficient for situations in which there are many lines carrying data, and where the energy required to transfer each event is small (as is the case in animal nervous systems).

III. Results from TIMIT dataset

The system was initially tested on a variety of sounds including 'string', 'percussion', 'male and female speech' and the results were assessed both by human listeners (subjective testing) and by using the composite techniques (objective testing) described in [19]. The subjective testing results mentioned at the table 5.41 in [18] shows that our lossy coding technique is able to generate audio with differences which cannot be perceived by the human ear using 16 bandpass channels and 12 threshold levels. This shows that our lossy coding technique is able to generate audio with differences which cannot be perceived by human ear for 16 bandpass channels and 12 threshold levels. The subjective test only contained 20 questions which took about 5 to 8 minutes for each of 21 participants to complete.

Here, we also describe results on short duration speech utterances from the TIMIT dataset [20]. This consists of 1360 female utterances, and 3260 male utterances. The composite computer based assessment technique is again from [19]. In the work reported here, we used their composite objective measure, Covl, calculated as

$$\text{Covl} = 1.594 + 0.805 * \text{PESQ} - 0.512 * \text{LLR} - 0.007 * \text{WSS} \quad (1)$$

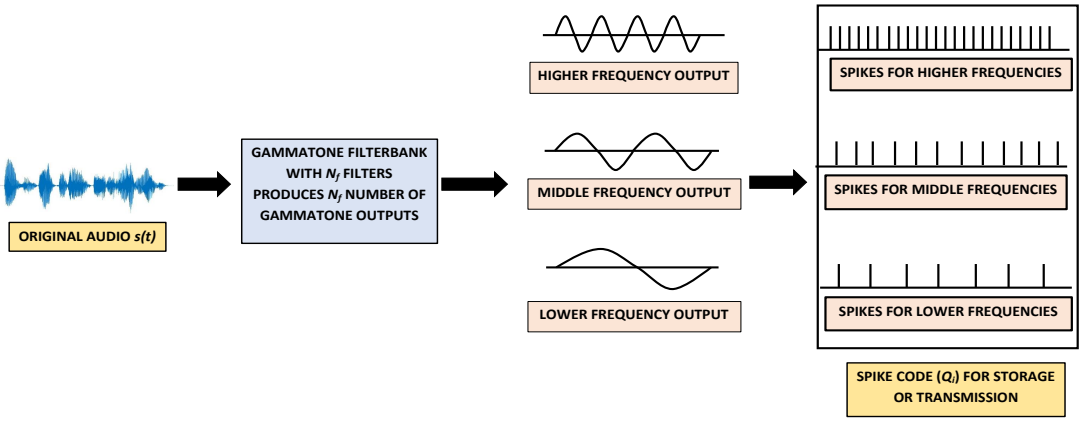

Fig. 1. Incoming sound $s(t)$ is passed through the gammatone filterbank with N_f number of centre frequencies. Each positive zero crossing is a spike at a threshold level, explained in the text. These are the spike trains Q_i which are ready to be stored or transmitted.

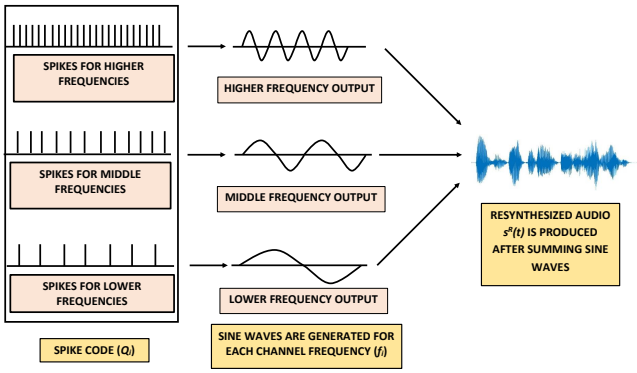

Fig. 2. The spike trains Q_i are processed using their threshold levels and bandpass channels centre frequencies, and sine waves are generated for each spike. These are summed to generate the resynthesized audio $S^R(t)$.

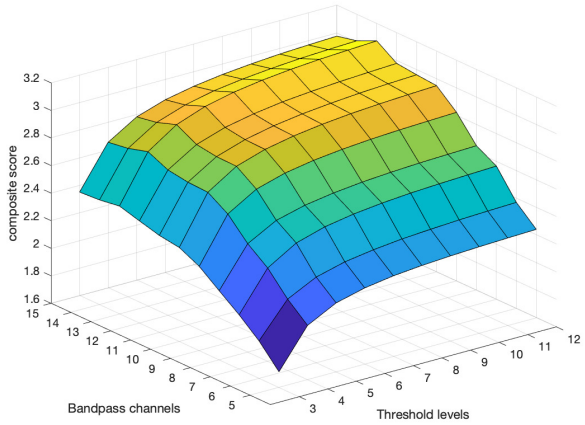

Fig. 3. Mean value of composite score (Covl) for 1360 resynthesised TIMIT female utterances, varying number of channels from 5 to 15, and number of threshold levels from 3 to 12.

where PESQ is the narrowband PESQ measure (ITU-T Recommendation P.862), LLR is log likelihood ratio [21], and WSS is the weighted slope spectral distance [22].

We varied the number of Gammatone filter bands, N_f, from 5 to 15 with steps of 2, with the lowest f_i set to 100 Hz, and the highest to 3950 Hz. The equivalent rectangular bandwidth was set to 1. In addition we varied the number of spike thresholds N_J from 3 to 12, with steps of 2. This gave the results shown in figures 3 and 4.

An instrumental intelligibility metric called $SIIB^{Gauss}$ has been developed in [23]. This is an evaluation method of comparing the intelligibility of two sounds. We have used $SIIB^{Gauss}$ over $SIIB$ as it takes less time to compute although both of them have state-of-the-art performance in evaluating the quality of a speech sample [24]. $SIIB^{Gauss}$ values have been calculated for the original audio with the resynthesised audio from the TIMIT dataset with the same Gammatone Filterbanks configuration. The TIMIT dataset is sampled at 8K samples/second, and thus cannot contain energy above 4Khz. Each is about 1 second long. The MATLAB codes developed and published in [24] requires the audio files to be over 20 seconds long. We therefore concatenated recordings from the same speaker to make the length of the audio files greater than 20 seconds.

A recording name in the TIMIT dataset can be as 'FAEM0SA1.AU'. Here, 'FAEM' is the code for the speaker ('F' for female and 'M' for male) and the 'SA1' is the code for the sentence. They are seperated by the digit '0'. We have used MATLAB script to concatenate all audio files for a speaker and then resynthesised them by using number of bandpass channels varying from 5 to 15 (in steps of 2), and number of threshold levels varying from 2 to 12 (in steps of 2). Then the mean $SIIB^{Gauss}$ values are generated comparing the original and resynthesised audio and shown in figure 6 and figure 5

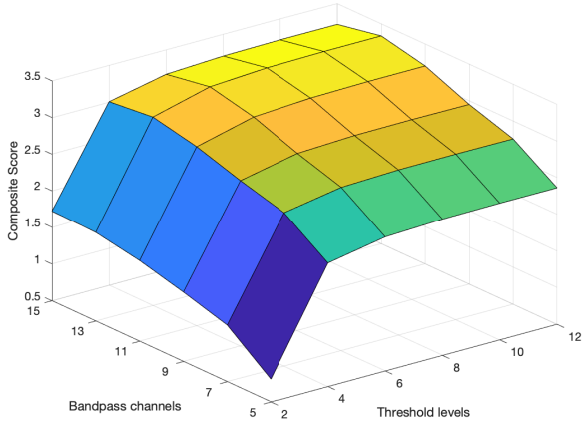

Fig. 4. Mean value of composite score (Covl) for 3260 resynthesised TIMIT male utterances, varying number of channels from 5 to 15 (in steps of 2), and number of threshold levels from 2 to 12 (in steps of 2).

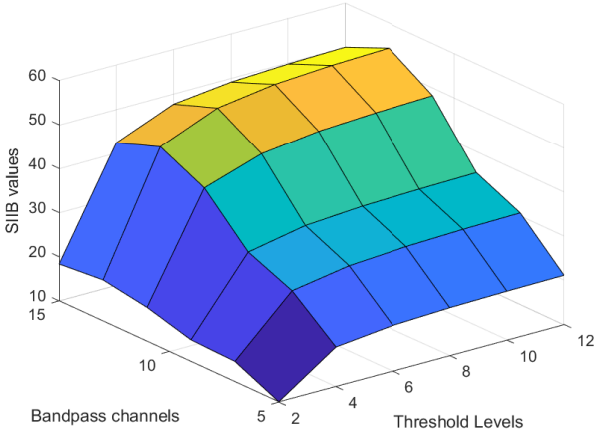

Fig. 6. Mean value of $SIIB^{Gauss}$ for original and resynthesised audio from 269 TIMIT male speakers, varying number of channels from 5 to 15 (in steps of 2), and number of threshold levels from 2 to 12 (in steps of 2).

for 269 males and 124 females.

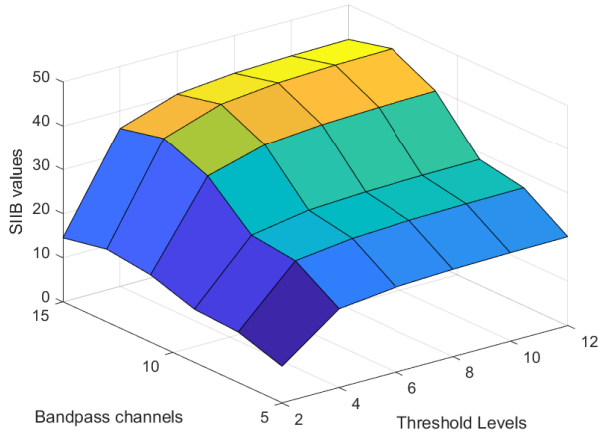

Fig. 5. Mean value of $SIIB^{Gauss}$ for original and resynthesised audio from 124 TIMIT female speakers, varying number of channels from 5 to 15 (in steps of 2), and number of threshold levels from 2 to 12 (in steps of 2).

Listening to the resynthesised utterances, male speech is always intelligible at 6 bandpass filters and 7 threshold levels: female speech seems to sound more distorted, though the intelligibility threshold remains about the same. The required number depends on both the speaker and what is being spoken (some of the original TIMIT utterances are difficult to understand for these non-US native English speakers). It is critical that the listener does not know what is being spoken, so that listeners need to hear the speech starting at the smallest number of bandpass channels and threshold levels. Initial experiments suggest that male voices can be intelligible with as few as 4 bandpass filters and 5 threshold levels, but this depends both on the speaker and the listener. Only an initial subjective assessment of few sounds has been carried out as explained in chapter 5 of [18]; a full-scale subjective intelligibility assessment is yet to be carried out.

Automatic speaker recognition and diarisation is possible by means of x-vectors [25] and we have developed a system shown in figure 7, to compare the effectiveness of a coding technique by recognising if the same speaker is diarised in both original and decoded audio.

As explained in figure 7, the number of channels has been varied from 5 to 15 (in steps of 2) and the number of threshold levels from 2 to 12 (in steps of 2) while decoding audio. Thus we have used 6 different numbers of bandpass channels and 6 sets of threshold levels which produces 36 decoded audio for every combination of bandpass channels and threshold levels. We concatenate the original audio file with the decoded audio files and this gives the total number of audio clip as 37. We have tested 20 male speakers and 20 female speakers and if we denote Φ as the symbol for concatenation, then the audio clip which is fed into the x-vector extractor after extracting mel-frequency cepstral coefficients (MFCC) features in figure 7 is $S(t)$, as shown in equation 2.

$$S(t) = \Phi_{n=1}^{20}(s^n(t) + \Phi_{d=1}^{36} s_d^n(t)) \tag{2}$$

The final audio $S(t)$ is approximately 6 hours long for both male and female. We have used a x-vector which is pre-trained on the multi-channel Wall Street Journal Audio-Visual data corpus [26] and on VoxCeleb data collection [27]. We have used Kaldi speech recognition toolkit [27] for extracting the x-vectors and then cluster the same speaker diarisation scores by means of a probabilistic linear discriminant analysis (PLDA). This produces a RTTM file [28] which is read in Python ([29] and [30]) and a 'speaker identifier strength' (η) is calculated by counting the number of times (γ) the same speaker is diarised

for an audio clip and the dividing it by the total number of audio clips (37) for a speaker, as shown in equation 3.

$$\eta = \frac{\gamma}{37} \tag{3}$$

where, $1 \leq \gamma \leq 37$

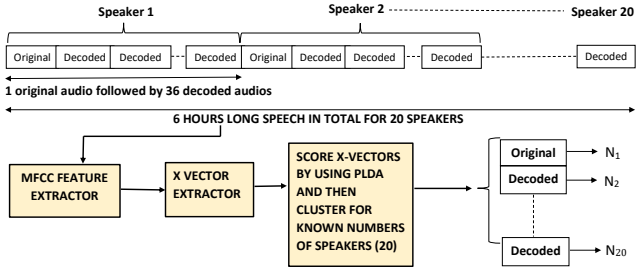

Fig. 7. **Speaker diarisation for the original and decoded audio:** the number of channels has been varied from 5 to 15 (in steps of 2) and the number of threshold levels from 2 to 12 (in steps of 2) for decoded audio and this gives us 36 decoded audio for one original audio clip. These 37 audio clips are concatenated together, as shown in equation 2, for one speaker and there are 20 (male/female) speakers in total and this produces a speech audio of length 6 hours combining 740 short audio clips. X-vectors are extracted from the MFCC features and then they are clustered by means of PLDA. Finally we have a speaker diarisation score (N_1, $N_2 \ldots N_{20}$) for every single audio clip in a RTTM file.

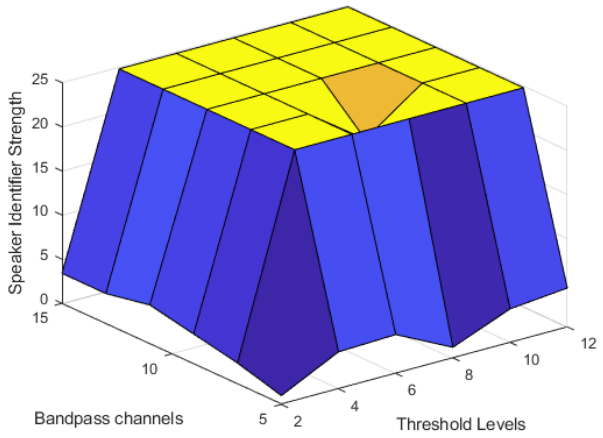

Fig. 8. **Speaker Identifier Strength:** For male speakers, higher number of threshold levels and bandpass channels decode the original audio in such a way that it can still be diarised as the same speaker.

Both figure 8 and 9 shows that using higher number of threshold levels and bandpass channels in our spike coding can produce decoded audio which can still be identified as coming from the same speaker. Some original and resynthesised sounds may be found at https://bit.ly/2l2XBZV.

IV. Discussion

The coding used here is a computationally tractable (and easily engineered) approximation to that used in the auditory nerve and similar codings are used in so-called silicon

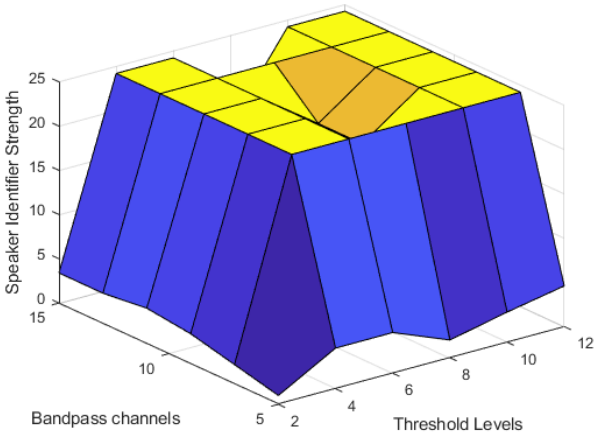

Fig. 9. **Speaker Identifier Strength:** For female speakers, higher number of threshold levels and bandpass channels decode the original audio in such a way that it can still be diarised as the same speaker, although it has more variation than male speakers.

cochleae [3]. The major differences between the coding used here and the type 1 AN fibers are that (i) we are using a far smaller number of bandpass filters than the approximately 3,500 transducing inner hair cells in the human cochlea (each drives about 10 type 1 AN fibers), (ii) we do not limit the number of events per second on each fiber (in real AN type 1 fibers, the maximal spiking rate is about 300 spikes per second), and (iii) we have a larger number of threshold levels (the literature suggests that there are low spontaneous rate and high spontaneous rate type 1 fibers) [1]. To some extent, these differences compensate for each other, at least at higher frequencies, since multiple fibers can have a high total spiking rate, but there is also stochasticity in real AN fiber firing, so that phase locking is lost after about 3 kHz.

We are interested in the *possibility* of decoding these events to produce intelligible speech: for this purpose, the complexity of the decoding process is not important. Clearly if one were interested in *using* this coding as a method of recording or transmitting sound, the size and complexity (duration) of the coding and decoding would be important. The amount of time taken for both coding and decoding depends primarily on the number of bandpass channels and the number of threshold levels. Using 7 bandpass channels, and 7 threshold levels on a Mac Mini (3.2 GHz Intel core i7, 32 Gbyte 2667 MHz memory, running Matlab R2019a under OSX10.14.5), coding and decoding 100 TIMIT utterances (total duration 294.47 seconds) takes 42.64 seconds; the coding itself takes 31.81 seconds. This suggests that decoding takes about 3.7% of the signal duration.

This work differs from [14] firstly in using the Gammatone filterbank, but more importantly in using the actual spike times from each filter output, rather than choosing the dominant frequency present every 20 ms. We note that (i) because filters are wideband, dominant frequencies will also be detected by

filters whose center frequency is distant, and (ii) the coding of the amplitude to use is much more impoverished when the number of thresholds is low. In [15], the Gammatone filterbank is used with a fixed number, 20, of filter bands. A power law based compression technique is used, followed by an ensemble of neurons for each bandpass channel. The firing rate is used to code the amplitude. In essence this is similar to what we do, by using geometrically related thresholds for the spike trains indexed by j. It does, however, require more spike trains. In [16], the same techniques as in [15] are used, but with more (50) bandpass channels. Using so many channels allows for omitting some in the reconstruction to enable better signal to noise ratios when there are interfering sounds. However, using more channels increases the amount of data required for code transmission.

A. Number of bandpass channels and threshold levels to use

The minimal amount of subjective testing used suggests that 6 or 7 bandpass channels and threshold levels are required to ensure intelligibility. The scores shown in figures 3, 4, 5 and 6 suggest that for a given number of bandpass channels and threshold levels, the intelligibility of male speech is higher than that of female speech. This is supported by the small amount of subjective testing done. This may well be due to the larger amount of the energy of female speech being at higher frequencies than is the case for male speech. Figure 8 and 9 shows that by using higher number of bandpass channels and threshold levels, our biologically inspired spike coding technique is able to preserve enough intelligence which enables a speaker recognition system, shown in figure 7, to diarise that both original and decoded audio is coming from the same speaker.

As can be seen from figures 3 and 4, and as is also clear from subjective testing, quality and intelligibility increases as the number of bandpass channels increases to about 11, but then plateaus. Something similar happens above about 8 threshold levels. Subjectively, there is a gradual increase in quality as the number of bandpass channels increases, but little audible difference above 9 threshold levels. The speakers can also be recognised successfully while using higher number of bandpass channels and threshold levels. This is the case for both male and female speech utterances.

B. Code size

One important question is the amount of data required to transmit the code. If one is storing the code, each event needs to store i, j and the event time t_i^k (or sufficient information to recreate these). For i, one needs $\lceil \log_2(N_f) \rceil$ bits, and for j, one needs $\lceil \log_2(N_J) \rceil$ bits. It is less clear how many bits are required to store the time, as this depends on the accuracy with which one needs to recreate the time. Timings within a few 10's of μs are required if one wants to be able to use time difference of arrival techniques for sound source direction finding, but for intelligibility alone, less precision is needed. One possibility is to use the sample number, or, to reduce the maximum value required, the number of samples since the last event (probably in this bandpass channel). At a sample rate of 16 Ksamples/second, 14 bits would allow times up to one second to be coded with $\pm 30\mu$s accuracy. Reducing the precision to $\pm 120\mu$s would allow times of up to 1 second to be coded in 12 bits. Thus, for $N_f = 16$ and $N_j = 8$, and accuracy of event times of $\pm 120\mu$s, 19 bits are required per spiking event.

If one was willing to store the events sorted by threshold level within bandpass channel number, the number of bits per event could be reduced where there are many events per threshold level within bandpass channel. This is usually the case. In this case, the channel number and threshold level would not need to be stored for every spiking event. Further, where many events occur closely spaced in the same bandpass channel, it would be possible to reduce the number of bits per event by coding the time between events rather than the event time. In addition, LZW coding [31] could be applied. However, these forms of coding would require the whole file to be read and decoded prior to generating any sound from the file, and cause a delay in playing the sound back.

In animals, the auditory nerve coding uses a separate nerve fiber for each spiral ganglion axon output (i.e. auditory nerve type 1 fiber). Signal times are not coded as such: the axonic spike marks its own time. Using N_f lines (one per bandpass channel), one would need to send a set of $\lceil \log_2(N_J) \rceil$ bits per event to code the threshold level, as the time would be the time of the signal, and the bandpass channel implicit in the choice of the line down which the pulse was sent. Going further, one could use a single higher speed line, and send a packet of length $\lceil \log_2(N_J) \rceil + \lceil \log_2(N_J) \rceil$ bits to code the channel number and threshold level. The time of the packet itself would code the event time. If we had $N_f = 16$, and a highest f_i of 4 KHz, the maximum number of events per second would be 21570, assuming that all bandpass channels spike at f_i events per second. Taking $N_J = 16$, this would require 8 bits per event, giving a maximal data rate of 172620 b/s. The data transfer rate required would be higher than this, to allow for protocols, say 300000 b/s. In addition, many events might be almost coincident, but we can only transfer one event at a time so that we need to ensure that the time between almost coincident events is small enough: a few 10's of μs. But serial lines of data rate greater than 10 Mb/s are commonplace so that this would be straightforward. We note however, that the event rate is normally much lower than this highest possible value: the highest value would only be reached if there was a positive-going zero-crossing in every bandpass channel at all times, which is unlikely to be the case.

C. Connection to neural networks

There are many ways in which sound (speech) can be coded for input to neural networks. One can, for example, calculate cepstral coefficients (e.g. MFCCs), as are used by the speech community. Why then are we suggesting something so different?

We are proposing an event-based scheme: in such a scheme, events arrive immediately after their occurrence. With an

MFCC (or other Fourier transform based technique), new values arrive regularly, perhaps every 25 or 40 milliseconds. Here, events arrive as they occur (or rather, with a delay that depends primarily on the speed of calculating bandpass transforms). Further, the coding maintains the fine time-structure of the sound signal, something that is lost with regular computation of a set of parameter values. While it is not clear that this is important for interpretation of clean speech, it is important for identifying foreground from background sound. Wideband sounds are the norm: they lead to the co-occurrence in time of events (spikes) across different parts of the spectrum. This co-occurrence can be used to group signals emanating from the same sound source [32]. This is simply not possible with regularly computed spectral coefficients.

We note that because this coding maintains precise timing, implementing it using signals from more than one microphone would enable time difference of arrival (TDOA) measurements to be made, assisting the identification of the direction of the sound source, though would need accurate time recording. Earlier work [9] used a similar coding to compute the TDOA at sound onsets (which are assumed to come from the direct path from the sound source). Onsets were detected using a neurally plausible technique based on depressing synapses and leaky integrate-and-fire neurons. If a similar coding technique, explained in this study, was used in a hearing aid or cochlear implant, this might enable the user to identify sound source direction.

V. Conclusions and future work

We have demonstrated a biologically inspired coding system, shown that it can be transmitted using a reasonable number of bits, and that it can be decoded quite quickly to produce intelligible speech. The intelligence has been measured by comparing the original and decoded signal by means of PESQ scores and $SIIB^{Gauss}$ values. Our findings also show that the same speaker can be verified in both original and decoded audio by using a pre-trained x-vector extractor.

Some parameter ranges have not yet been thoroughly investigated, for example the equivalent rectangular bandwidth, as well as the dynamic range. In addition, other measures of intelligibility could be calculated by running a thorough subjective test with more than 100 participants from different cultures and accent and more than 100 comparisons between original and resynthesised sounds. Another state-of-the-art algorithm used for speech and speaker verification is 'i-vector' [33] which can be extracted by using deep neural network [34] for short utterances like the recordings used in TIMIT dataset [35]. It will be interesting to see if the resynthesised audio can still be verified as coming from the same speaker by using a pre-trained 'i-vector extractor'.

We suggest that this type of coding is particularly applicable for providing input to spiking neural networks: the maintenance of precise timing, as well as of spectral and intensity information is important for foreground/background streaming for sounds, and not just for speech as well as for the interpretation of sound. Because coded sounds can be reconstructed while maintaining quality, we suggest that this biologically inspired coding technique maintains the information important for interpretation and speaker recognition.

Acknowledgements

Funding for Pahar during his Ph.D. was partially provided by the McGlashan Trust, the Sidney Perry Foundation, the CAM Syms Charitable Trust, and the University of Stirling.

References

[1] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 4th ed. Brill, May 2013.

[2] J. Maciokas, P. Goodman, and F. Harris, "Large-scale spike-timingdependent-plasticity model of bimodal (audio/visual) processing," *Technical Paper. Brain Computation Lab, University of Nevada, Reno, NV*, 2002.

[3] S.-C. Liu, T. Delbruck, G. Indiveri, and R. J. Douglas, "Silicon Cochleas," in *Event-based neuromorphic systems*, S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. J. Douglas, Eds. Chichester, UK: John Wiley and Sons, Dec. 2014, pp. 1–20.

[4] Y. Karklin, C. Ekanadham, and E. P. Simoncelli, "Hierarchical spike coding of sound," in *Advances in neural information processing systems*, 2012, pp. 3032–3040.

[5] R. Pichevar, H. Najaf-Zadeh, and L. Thibault, "A biologically-inspired low-bit-rate universal audio coder," *Audio Eng. Society Conv., Austria*, 2007.

[6] R. Pichevar, H. Najaf-Zadeh, L. Thibault, and H. Lahdili, "Differential graph-based coding of spikes in a biologically-inspired universal audio coder," *Audio Eng. Society Conv., Netherlands*, 2008.

[7] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Tech. Rep. Annex B of SVOS Final Report, 1987.

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

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

[10] M. J. Newton and L. S. Smith, "A neurally inspired musical instrument classification system based upon the sound onset," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, p. 4785, 2012.

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

[12] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*. IEEE, Aug. 2010, pp. 3809–3812.

[13] R. Pichevar, H. Najaf-Zadeh, H. Lahdili, and L. Thibault, "Entropy-constrained spike modulus quantization in a bio-inspired universal audio coder," in *2008 16th European Signal Processing Conference*. IEEE, 2008, pp. 1–5.

[14] O. Ghitza, "Auditory Nerve Representation Criteria for Speech Analysis/Synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 6, pp. 736–740, Jan. 1987.

[15] G. Kubin and W. Bastiaan Kleijn, "On speech coding in a perceptual domain," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. IEEE, Oct. 1998, pp. 205–208 vol.1.

[16] C. Feldbauer, G. Kubin, and W. B. Kleijn, "Anthropomorphic Coding of Speech and Audio: A Model Inversion Approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 571 618–18, Jun. 2005.

[17] M. Cooke, *Modelling auditory processing and organisation*. Cambridge University Press, 1993.

[18] M. Pahar, "A Novel Sound Reconstruction Technique based on a Spike Code (event) Representation," Ph.D. dissertation, University of Stirling, University of Stirling, 2016.

[19] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, Dec. 2007.

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

[21] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice-Hall, 1988.

[22] D. H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, May 1982.

[23] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *arXiv.org*, no. 11, pp. 2153–2166, Aug. 2017.

[24] S. V. Kuyk. (2018) Matlab code. [Online]. Available: https://stevenvankuyk.com/matlab_code/

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[26] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "Sms-wsj: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[27] Y. Yang, "Automatic speaker verification and diarization on voxceleb data collection," Ph.D. dissertation, Georgia Institute of Technology, 2020.

[28] N. Spring, "Rich transcription meeting recognition evaluation plan," 2005.

[29] M. Lutz, *Programming python*. "O'Reilly Media, Inc.", 2001.

[30] T. E. Oliphant, "Python for scientific computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 10–20, 2007.

[31] C. H. Lin, Y. Xie, and W. Wolf, "Lzw-based code compression for vliw embedded systems," in *Proceedings Design, Automation and Test in Europe Conference and Exhibition*, vol. 3. IEEE, 2004, pp. 76–81.

[32] D. Wang and G. J. Brown, *Computational auditory scene analysis*. John Wiley and sons, Oct. 2006.

[33] W. Li, T. Fu, and J. Zhu, "An improved i-vector extraction algorithm for speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 18, 2015.

[34] W. Wang, W. Song, C. Chen, Z. Zhang, and Y. Xin, "I-vector features and deep neural network modeling for language recognition," *Procedia computer science*, vol. 147, pp. 36–43, 2019.

[35] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances." 08 2011.