A Unifying Framework for Neuro-Inspired, Data-Driven Detection of Low-Level Auditory Features

Lotte Weerts (lw1115@imperial.ac.uk)
Department of Electrical and Electronic Engineering, Imperial College London
South Kensington Campus, London SW7 2AZ, United Kingdom

Claudia Clopath (c.clopath@imperial.ac.uk)
Department of Bioengineering, Imperial College London
South Kensington Campus, London SW7 2AZ, United Kingdom

Dan F.M. Goodman (d.goodman@imperial.ac.uk)
Department of Electrical and Electronic Engineering, Imperial College London
South Kensington Campus, London SW7 2AZ, United Kingdom

Abstract

Our understanding of hearing and speech recognition rests on controlled experiments requiring simple stimuli. However, these stimuli often lack the characteristics of complex sounds such as speech. We propose an approach that combines neural modelling with machine learning to determine relevant low-level auditory features. Our approach bridges the gap between detailed neuronal models that capture specific auditory responses, and research on the statistics of real-world speech data and speech recognition. First, we introduce a feature detection model with a modest number of parameters that is compatible with auditory physiology. In order to objectively determine relevant feature detectors within the model parameter space, the model is tested in a speech classification task, using a simple classifier that approximates the information bottleneck. This framework allows us to determine the best model parameters and their neurophysiological and psychoacoustic implications. We show that our model can capture a variety of well-studied features (such as amplitude modulations and onsets) and allows us to unify concepts from different areas of hearing research. Our approach has various potential applications. Firstly, it could lead to new, testable experimental hypotheses for understanding hearing. Moreover, promising features could be directly applied as a new acoustic front-end for speech recognition systems.

Keywords: neural modelling; phoneme detection; information coding; speech recognition; machine learning

Introduction

Our current general understanding of hearing relies heavily on experimental approaches that require simple stimuli to allow for controlled experiments. Moreover, these simple stimuli are often based on specific auditory features that have been derived from fields such as signal processing (e.g. amplitude modulations (AM) and frequency modulations (FM)) and music (e.g. timbre and pitch). These features cannot necessarily be easily related to neurophysiology. In this work, we aim to expand the assumptions that underlie the basic features that are examined in auditory research by harnessing the strengths of both neural modelling and machine learning. Rather than focusing on well-known features, we propose a neuro-inspired auditory feature detection model that is compatible with auditory physiology and is capable of detecting a variety of auditory features. We then apply a data-driven machine learning approach to explore the model parameter space and objectively identify relevant features. Our approach allows us to bridge the gap between detailed neuronal models that capture specific auditory responses, and research on the statistics of real-world speech data and its relationship to speech recognition. Importantly, our feature detection model can capture a wide variety of well-studied features using specific parameter choices, and allows us to unify several concepts from different areas of hearing research.

Method

Neuronal Feature Detection Model

Our feature detection model is inspired by the early auditory pathway, which includes the auditory periphery and early brainstem. The model utilises precisely timed inhibition as a mechanism for feature detection. This mechanism is thought to be employed at various locations in the early auditory pathway to improve noise-robust speech processing, such as in onset and offset sensitive octopus cells in the Ventral Cochlear Nucleus (VCN), spectral notch detection through wideband inhibition in the Dorsal Cochlear Nucleus (DCN) and AM sensitive neurons in the Inferior Colliculus (IC). Our model is a simplification and generalisation of previous proposals (see e.g. Carney, Li, and McDonough (2015); Skorheim, Razak, and Bazhenov (2014); Smith and Fraser (2004) for neuronal models for AM, FM and onset sensitivity, respectively) and unifies the detection of a variety of features that tend to be treated separately in the literature, such as onsets and AMs.

The first stage of the model approximates the auditory nerve (AN) fiber response. Each AN fiber response x_{f_c} for a signal x is estimated as follows:

$$x_{f_c} = h(g_{f_c}(x))^{\frac{1}{3}} \tag{1}$$

Here, $h(\cdot) = max(\cdot, 0)$ denotes a half-wave rectification and $g_{f_c}(\cdot)$ denotes a Gammatone filter that extracts the frequency component of x centred around f_c. Given two sets of centre frequencies E and I, the model output $o[t]$ at a given time t is computed as follows:

$$o[t] = h(\sum_{E_i \in E} f_{\tau_E}(x_{E_i})[t + d_i] - r \sum_{I_j \in I} f_{\tau_I}(x_{I_j})[t + d_j]) \quad (2)$$

Here, $f_{\tau_I}(\cdot)$ and $f_{\tau_E}(\cdot)$ are low pass filters with time constants τ_E and τ_I of inhibitory and excitatory populations. The delay $d_i \in D$ denotes the delay of each AN response, r is the ratio between the strength of the inhibitory and excitatory stream. Our model thus relies on six sets of parameters: E, I, D, r, τ_e and τ_i. In order to keep the parameter space tractable, the model variants in the results presented here were restricted to up to two excitatory centre frequencies with wideband inhibition centred around E, but this could be expanded in future work.

Analysis Through Classification

The parameters of the model determine the type of features it responds to. To study the importance of each of these features in a speech processing context, the feature detectors are employed in a phoneme classification task. First, our model is applied to a given phoneme signal x to obtain our model output o. The phoneme data used here originates from the LibriSpeech dataset (Panayotov, Chen, Povey, & Khudanpur, 2015), which was aligned per phoneme using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017). The model output is used as input to a one-layer Long-Short-Term-Memory recurrent neural network classifier (LSTM), a machine learning algorithm that is well-suited for classifying sequenced data. The LSTM is used to classify the phoneme label y. We employ a small LSTM to ensure that high accuracy can only be achieved by features that have a relatively simple relationship to the label, an idea that is motivated by information theory. The Data Processing Inequality dictates that the model output o can only reduce the information about label y compared to the information that was available in x. We are interested in features that provide a useful compression of x and only retain the relevant information about y, while forgetting the irrelevant information present in x, a general principle known as the Information Bottleneck (IB) (Tishby, Pereira, & Bialek, 2000). Determining the IB directly is computationally intractable for high-dimensional variables (such as sound waves). However, our results indicate that there is a high correlation between the LSTM classification accuracy and the IB principle in a reduced setting, where the IB is made tractable by compressing the high-dimensional data (data not shown).

Results

Phoneme Classification Performance

We applied 1000 model variants with randomly sampled parameter settings to a classification task of twelve phonemes, consisting of four vowels (ʌ, ɪ, ɛ, æ), four fricative consonants (f, v, s, z) and four plosive consonants (p, b, t, d). A summation of the confusions of the 100 best-performing model variants in the three different groups shows that, as expected, most confusions are made within the three phoneme groups (fig 1a). A correlation analysis of the accuracies between the three groups indicates that model variants that are well-suited for fricative detection tend to be good at plosive and vowel recognition as well, but the same is not true when comparing model performance for plosive and vowel recognition (fig 1b). This reflects the lack of overlap in spectral content of plosives and vowels compared to fricatives, as fricatives contain the noisy waveform that is characteristic for plosives as well as a more periodic pattern that is usually associated with vowels.

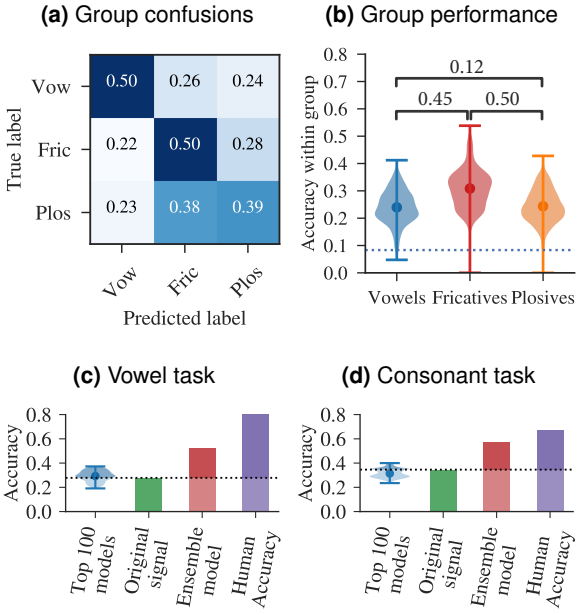

Figure 1: **a** Mean between-group confusion matrix for best model variants. **b** Distribution of within-group accuracies for all model variants, and between-group accuracy correlations. **c-d** Accuracy achieved with the best model variants from **a**, the original filtered signal, and ensemble models on a 10-vowel (**c**) and 13-consonant (**d**) classification task. Human performance on a similar task from Meyer et al. (2010) is included.

The 100 best model variants were applied to a more difficult 10-vowel recognition task (fig 1c) and 13 consonant recognition (fig 1d) task. For certain parameter settings, a single feature detector can outperform accuracy obtained with the original filtered signal. This suggests that the detected features allow for a more accessible representation of the relevant information in the original signal. The accuracy can be further improved by employing an ensemble of feature detectors as input for the LSTM instead of single model variants. Using an ensemble of the 23 feature detectors that individually performed best for each vowel and consonant increases accuracy by around 30%. The achieved accuracy is particularly high for consonants, where performance comes close to human performance in a similar task (fig 1d-c).

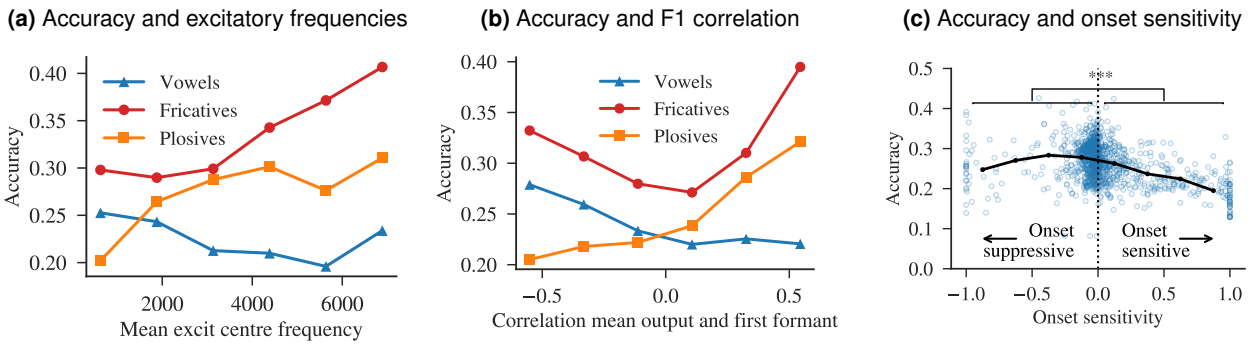

Figure 2: **a** Within-group accuracy in relation to the mean centre frequency for 12-phoneme task **b** Within-group accuracy and the correlation of the average model output and the first formant (or spectral peak) in the 12-phoneme dataset. Formants were estimated using parselmouth, a Python library for Praat (Boersma & Weenink, 2018; Jadoul et al., 2018). **c** Accuracy versus onset sensitivity, the mean accuracy of onset suppressive versus sensitive model variants is significantly different ($p \ll 0.0005$)

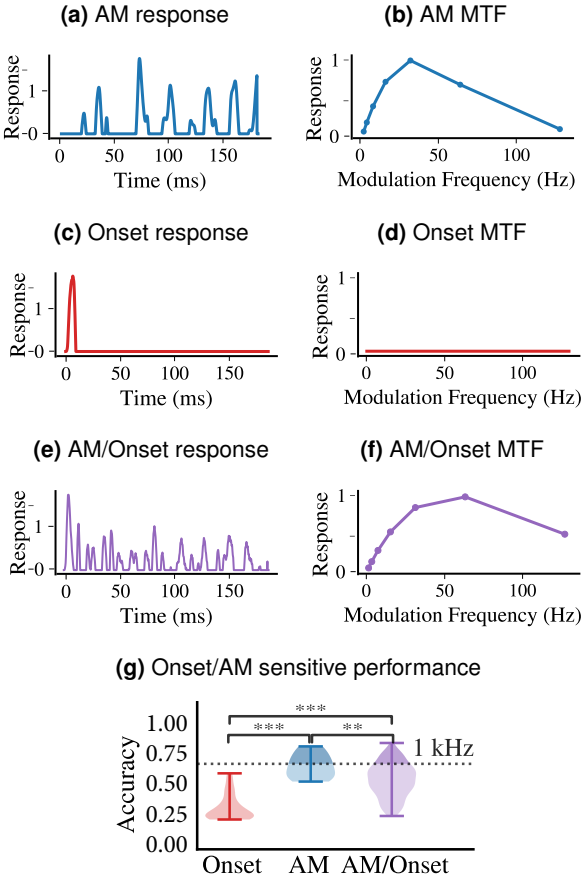

Figure 3: **a-f** Responses and Modulation Transfer Function (MTFs) of model variants (measured as the normalised mean response to various AMs) for onset (**a,b**) AM (**c,d**) and both AM/onset (**e,f**) model variants. **g** Distribution of accuracies of three model types (params $E = I = 1$ kHz and $\tau_i > \tau_e$ and $d_i > d_e$.)in a fricative classification task ($p < 0.005$ and $p < 0.0005$ for ** and ***).

Relevant Model Parameters and Detected Features

In this section, we discuss several relevant features detected by our model that performed well in the context of a phoneme classification task. We also relate these features back to psychoacoustic and neurophysiological findings.

AN Fibers Around Formants and Spectral Peaks Indicate Good Performance The accuracy within the different phoneme groups is heavily influenced by the centre frequencies of the model variants. We find that vowel recognition is facilitated by lower centre frequencies, whereas best-performing model variants for fricatives and particularly plosives tend to have a high centre frequency (fig 2a). This effect is closely related to the formants in vowels, which tend to lie between 200-1000 Hz, and spectral peaks in consonants, which tend to be much higher. This relation can be made explicit by correlating the formants of all phonemes with the mean output of the model for these phonemes (fig 2b). Models that are good at classifying vowels tend to have a negative correlation (i.e. models have a high output when the formants are low) whereas a positive correlation (i.e. model responds to high frequency formants) is typical for improved plosive recognition. Fricatives, which have high formants but can also have slower periodic components, benefit from either. Model outputs with a correlation close to zero (i.e. models that do not respond to spectral peaks) perform badly in all groups.

AM Sensitivity Improves Fricative Recognition Our model can be used to detect a range of features, in particular onsets, AMs or combinations of the two (fig 3a-f). We analysed the performance of 350 model variants that could be grouped into one of these three classes based on their onset response and Modulation Transfer Function. Here, each model variant is restricted to 1kHz centre frequency for excitation and inhibition and can only have positive inhibitory delays. When applied to a classification task of four fricative consonants, our results indicate that AM sensitivity is important for fricative recognition (fig 3g). Importantly, the accuracy of certain model variants is higher than the accuracy achieved by directly using the 1kHz channel as input, indicating that the

feature detector extracts useful features from the original signal. These results are in line with results from experimental research which show that consonant recognition is degraded when certain AMs are removed (e.g. van der Horst, Leeuw, and Dreschler (1999)).

Onset Suppression Facilitates Phoneme Classification We further extend our analysis of well-known features by investigating the typical onset sensitivity of a range of model variants (fig 2c). The onset sensitivity is computed as $(\mu_o - \mu_r)/(\mu_o + \mu_r)$, where μ_o and μ_r refer to the mean response to the first 50 periods plus the median delay, and last 200 periods of a preferred frequency sound. An onset sensitivity of -1 or 1 indicates a strongly onset suppressive or sensitive response, respectively. Our analysis shows that strong onset sensitivity generally leads to worse performance, whereas onset suppression improves performance. These results are in line with the hypothesis that the suppression of onset noise (or 'spectral splatter'), as observed in the mammalian auditory brainstem, can improve the clarity of a neural harmonic representation (Spencer et al., 2015).

Conclusion

We have introduced a neuro-inspired feature detector model as well as an analysis methodology that can be used to detect important features. Our approach provides a unifying framework that can be used to confirm and explain existing concepts, such as the importance of spectral peak detection and amplitude modulations. Moreover, it can be used to make suggestions for future work in less well-known theories of hearing, such as the importance of spectral splatter suppression (Spencer et al., 2015). Furthermore, the framework lends itself well for extensions to different settings, such as a variety of speech-related tasks (e.g. speaker identification) and environments (e.g. noise-robust features). Our proposed approach has various potential applications. Firstly, it could lead to new, testable experimental hypotheses for understanding hearing, both on the level of features as well as the underlying neurophysiological parameters. Moreover, promising features could be directly applied as a new acoustic front-end for speech recognition systems.

Acknowledgments

This work was partly supported by a Titan Xp donated by the NVIDIA Corporation, The Royal Society grant RG170298 and the Engineering and Physical Sciences Research Council (grant number EP/L016737/1).

References

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer [Computer program].* Version 6.0.43, retrieved 8 September 2018 http://www.praat.org.

Carney, L. H., Li, T., & McDonough, J. M. (2015). Speech Coding in the Brain: Representation of Vowel Formants by Midbrain Neurons Tuned to Sound Fluctuations. *eNeuro*, *2*(4). doi: 10.1523/ENEURO.0004-15.2015

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. doi: https://doi.org/10.1016/j.wocn.2018.07.001

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Interspeech, Stockholm, Sweden*.

Meyer, B. T., Jürgens, T., Wesker, T., Brand, T., & Kollmeier, B. (2010). Human phoneme recognition depending on speech-intrinsic variability. *The Journal of the Acoustical Society of America*, *128*(5), 3126–3141.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Icassp, ieee international conference on acoustics, speech and signal processing - proceedings* (Vol. 2015-Augus, pp. 5206–5210). doi: 10.1109/ICASSP.2015.7178964

Skorheim, S., Razak, K., & Bazhenov, M. (2014). Network models of frequency modulated sweep detection. *PLoS ONE*, *9*(12). doi: 10.1371/journal.pone.0115196

Smith, L. S., & Fraser, D. S. (2004). Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses. *IEEE Transactions on Neural Networks*, *15*(5), 1125–1134. doi: 10.1109/TNN.2004.832831

Spencer, M. J., Nayagam, D. A., Clarey, J. C., Paolini, A. G., Meffin, H., Burkitt, A. N., & Grayden, D. B. (2015). Broadband onset inhibition can suppress spectral splatter in the auditory brainstem. *PloS one*, *10*(5), e0126500.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

van der Horst, R., Leeuw, A. R., & Dreschler, W. A. (1999). Importance of temporal-envelope cues in consonant recognition. *The Journal of the Acoustical Society of America*, *105*(3), 1801–1809. doi: 10.1121/1.426718