A computational model of sound recognition used to analyze the capacity and adaptability in learning vowel classes

Jeffrey Spencer[1,2,*] Neil McLachan[3], and David B. Grayden[1,2]

[1] *Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Australia*

[2] *Centre for Neural Engineering, University of Melbourne, Melbourne, Australia*

[3] *Centre for Music, Mind, and Wellbeing, Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Australia*

Sound recognition is likely to initiate early in auditory processing and use stored representations (spectrotemporal templates) to compare against spectral information from auditory brainstem responses over time. A computational model of sound recognition is developed using neurobiologically plausible operations. The adaptability and number of templates required for the computational model to correctly recognize 10 Klatt-synthesized vowels is determined to be around 1250 templates when trained with random fundamental frequencies from the male pitch range and randomized variation of the first three formants of each vowel. To investigate the ability to adapt to noise and other unheard vowel utterances, test sets with 1000 randomly generated Klatt vowels in babble at signal-to-noise ratios (SNRs) of 20 dB, 10 dB, 5 dB, 0 dB, and −5 dB are generated. The vowel recognition rates at each SNR are 99.7%, 99.6%, 97.0%, 77.6%, and 54.0%, respectively. Also, a test set of four vowel recordings from four speakers is tested with no noise, giving 100% recognition rate. These data suggest that storage of auditory representations for speech at the spectrotemporal resolution of the auditory nerve over a typical range of spoken pitch does not require excessive memory resources or computing to implement on parallel computer systems.

INTRODUCTION

Most research on sound recognition in computational systems has been on automatic speech recognition systems. Automatic Speech Recognition (ASR) has primarily used Hidden Markov Models (HMMs) to model the statistics of the acoustic features in human speech (Rabiner, 1989). The performance of ASR systems using HMMs is significantly worse in noisy compared to clean conditions especially in non-stationary noise such as babble noise. The recognition accuracy for vowel identification of current automatic speech recognition systems at −5 dB SNR is comparable to human vowel identification scores at −15 dB SNR (Kalinli *et al.*, 2010; Mi *et al.*, 2013). This performance gap increases even further if these ASR systems are only trained with clean data instead of trained with both clean and noisy data (Kalinli *et al.*, 2010;

*Corresponding author: jeffspencerd@gmail.com

Proceedings of ISAAR 2013: Auditory Plasticity – Listening with the Brain. 4th symposium on Auditory and Audiological Research. August 2013, Nyborg, Denmark. Edited by T. Dau, S. Santurette, J. C. Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. ISBN: 978-87-990013-4-7. The Danavox Jubilee Foundation, 2014.

Pearce and Hirsch, 2000).

Template-based or exemplar-based approaches to sound recognition modeled on the neurobiology and psycholinguistics of the auditory system can provide more accurate modeling of auditory signals (Deng and Strik, 2007). A limitation often stated in the past for template-based systems is the computational power and memory resources required is too excessive (De Wachter *et al.*, 2007). This limitation has become less of a problem in recent years with availability of large increases in computing power and memory storage. Furthermore, template-based systems can be implemented in parallel in near real-time systems, such as field-programmable gate arrays (FPGAs) or graphics processing units (GPUs).

A recent neurobiological model of the auditory system, the Object-Attribute Model (OAM), postulates that long-term memory modulates neural spectrotemporal receptive fields through recognition mechanisms early in cortical processing (McLachlan and Wilson, 2010). Temporal information is not available at the onset of the sound, necessitating the use of sequential slices of spectral information to compare to long-term memory templates through time to recognize sounds. Hebbian learning enables the creation and adaptation of the long-term memory templates for commonly occurring sound timbres (McLachlan and Wilson, 2010). The stages of the computational sound recognition model described here are based on the mechanisms in the OAM and use neurobiologically plausible mechanisms for spectrotemporal processing of the auditory signals fed to the model.

Research on categorical vowel perception has shown that vowels are not perceived categorically but are rather perceived along a continuum (Schouten *et al.*, 2003). Human listeners do not make unanimous decisions about vowels when perceptual vowel boundaries overlap in the F1/F2 vowel space (Peterson and Barney, 1952; Hillenbrand and Gayvert, 1993; Neel, 2008). Although vowels have overlap in the perceptual boundaries, vowels presented in isolation do have a region in the F1/F2 vowel space where they are most consistently identified (Fairbanks and Grubb, 1961). This F1/F2 vowel-formant space is the region containing the points where at least 75% of the listeners correctly identified the produced vowel. This suggests that a centroid in the F1/F2 vowel space best represents a vowel.

The model is first trained until recognition accuracy proceeds to near 100% for ten Klatt-synthesized (Klatt, 1980) vowels from the most representative vowel region of the F1/F2/F3 vowel space in the seminal work of Peterson and Barney (1952). The training determines if the number of templates and memory storage required for correct recognition is feasible on current computer systems. Furthermore, the recognition accuracy of the model with different resolutions of fundamental frequency in the templates is compared to determine if fine pitch information is required for recognition. Then, the template database is used to explore the benefit and adaptability of the recognition system with speech babble noise added to Klatt-synthesized vowels at multiple SNRs. In addition, the template database is used for recognition of a small

set of recorded vowels /ae, ɜ, i, u/ from two male speakers to see if the model can adapt from synthesized to real speech.

METHOD

The model is implemented in the programming language Python using the numpy and scipy packages (Oliphant, 2007). The beginning processing stages are similar to many other models involving the auditory periphery (Slaney, 1993). The stages include a Gammatone filter bank similar to Slaney (1993), half-wave rectification as an approximation of hair-cell transduction, and the formation of specific loudness by short-duration temporal integration at each filter channel (Viemeister and Wakefield, 1991). Next, lateral inhibition is performed to sharpen the spectral resolution by off-frequency inhibition (McLachlan, 2009). Following lateral inhibition is a nonlinear dynamic saturation stage that provides loudness invariance and noise robustness. The saturation stage calculates saturation thresholds across every filter channel independently. At each filter, a Gaussian function weights the neighboring filters, and the mean of the center filter and the weighted neighboring filters is taken as the saturation threshold for that filter channel. Therefore, the saturation threshold can increase in only specific regions of the spectrum to rise above the noise level. This saturation mechanism can not only rise above white noise but also non-stationary noise such as babble noise. More specific details of the processing stages of the computational model can be found in McLachlan (2011).

The model is based on normal-hearing listener classification of the vowels /i, ɪ, ɛ, æ, ʌ, a, ɔ, ʊ, u, ɜ/. The fundamental frequency (F0) and first three formants (F1/F2/F3) of these vowels are taken from the unanimously classified male spoken vowels in Peterson and Barney (1952). The unanimously classified vowels are the vowels that are heard as the intended vowel by all 70 listeners. The minimum and maximum F0 values (93-203 Hz) are used to define the range for choosing the fundamental frequency values. The F1(x_1), F2(x_2), and F3(x_3) values for each vowel are fitted with a three-dimensional multivariate Gaussian distribution,

$$f_{\mathbf{x}}(x_1,x_2,x_3) = \frac{1}{\sqrt{(2\pi)^3|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right), \qquad \text{(Eq. 1)}$$

where Σ is the covariance matrix, $|\Sigma|$ is the determinant of Σ, and μ is the vector of means. The tolerance region of the distribution is the region in which at least p percentage of the points are enclosed. The tolerance region is defined as

$$(x-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) \leq c = \chi_3^2(p), \qquad \text{(Eq. 2)}$$

where c is the tolerance factor and $\chi_3^2(p)$ is the percent point function for probability p of the chi-squared distribution with three degrees of freedom (Krishnamoorthy and Mathew, 2009).

The surfaces in Fig. 1 are ellipsoids defined by a constant probability containing 30% ($p = 0.3$) of the points ($c = 1.42$). Enclosing 30% of the data points is roughly one

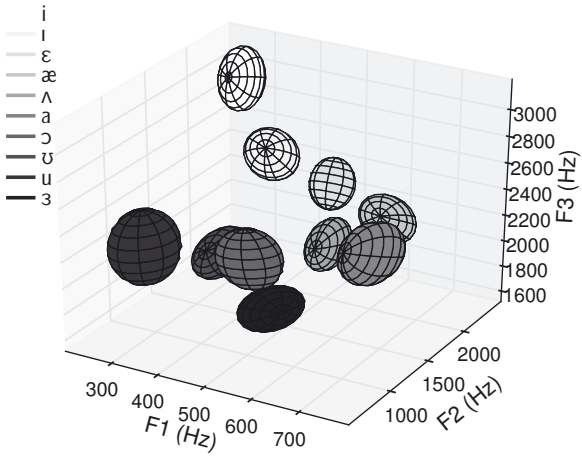

Fig. 1: Multivariate Gaussian distribution for each male spoken vowel that are unanimously classified (Peterson and Barney, 1952). The ellipsoids enclose at least 30% of the points in each distribution. This corresponds to a tolerance factor of $c = 1.42365$ and roughly one standard deviation in each direction F1/F2/F3.

standard deviation in each formant direction from the centroid of each vowel. These vowel ellipsoids are the F1/F2/F3 vowel spaces that best represent the spoken vowels from Peterson and Barney (1952) and recognition accuracy is expected to reach 100%.

Klatt-synthesized (Klatt, 1980) vowels are initially computed from the means of the formants of each vowel ellipsoid at the mean pitch from the male vowel recordings. The bandwidth for each formant frequency is determined by a fifth-order polynomial fit to measured closed-glottis bandwidths (Hawks and Miller, 1995). Separate fifth-order polynomials are fit to the data below and above 500 Hz. The other parameters fed into the Klatt-synthesizer, besides the default Klatt-synthesizer (Klatt and Klatt, 1990) values are F4 = 3400 Hz, F5 = 4000 Hz, sampling frequency = 12 kHz, and duration = 300 ms. This creates 10 Klatt-synthesized vowels at the mean values in F0, F1, F2, and F3 space.

Spectral templates are computed from the 10 Klatt-synthesized vowels as the initial training set. Then, to determine the number of vowel templates required for the correct recognition of the 10 vowels, the program is iterated by choosing a random vowel, random fundamental frequency (93-203 Hz), and random formant parameters from the multivariate Gaussian distribution of the chosen vowel at each iteration. Multiple training sets are computed at different fundamental-frequency resolutions. The fundamental frequency in the training sets is chosen randomly at semitone intervals of 3, 1, 0.25, and 0.1 calculated from the lowest male F0 (93 Hz) or at random from the set of rational numbers. These parameters are used to Klatt-synthesize a new

vowel, which is fed through the model. The output from the model is then compared to all vowels currently in the stored template database. The most activated template in the stored template database is checked to see if it comes from the same vowel as the computed spectral template. If the vowel does not match (a misclassification), the computed spectral template is added to the training database, and then the next iteration begins.

The procedure for selecting a random fundamental frequency from the rational numbers during the training phase is used to generate a testing set. The testing set is Klatt-synthesized vowels with no noise added (clean) and Klatt-synthesized vowels with babble noise added at SNRs of 20 dB, 10 dB, 5 dB, 0 dB, and -5 dB. For vowels with added babble, a Klatt-synthesized vowel (sig_{kl}) and a random 300-ms section of the babble noise (sig_{noi}) recording from the Aurora2 dataset (Pearce and Hirsch, 2000) are chosen at each iteration. The two signals are added together after determining the proper coefficient to multiply by the noise to get the desired SNR. The final input signal is $sig_{fin} = sig_{kl} + sig_{noi}\sqrt{\frac{p_{sig}}{p_{noi}}}10^{\frac{-SNR_{dB}}{20}}$ where the power of the 300-ms Klatt-synthesized vowel is p_{sig} and the power of the 300-ms segment of babble noise is p_{noi}. This iteration is done 100 times for each of the ten vowels for a total of 1000 inputs at each SNR. 1000 vowel inputs or 100 for each vowel are also generated in the clean condition. The Klatt-synthesized vowels in the testing set are then fed through the model and compared to the template databases built during training for each training set. Furthermore, recorded vowels produced from two native male English speakers for four of the ten vowels are also compared against the completely random template database. The recorded vowels are /ae, ɜ, i, u/.

RESULTS

The model requires 400,000 iterations to build the training database. The total number of templates compared to iteration number in each training set is shown in Fig. 2a. The number of templates stored for four hundred thousand iterations at semitone intervals of 3, 1, 0.25, 0.1, and random is 189, 505, 1009, 1294, and 1324, respectively. The recognition accuracy by the end of the four hundred thousand iterations for the training sets at 3, 1, 1/4, 1/10, and random semitone intervals is 96.8%, 98.8%, 99.5%, 99.9%, and 99.9%, respectively. The addition of finer frequency resolution adds fine pitch information that is not needed for high rates of recognition for the vowels. The recognition rate drops only 3% from selecting the fundamental frequency using 3 semitone intervals (5 frequencies total) to completely random selection of the fundamental. Furthermore, the number of templates stored drops substantially from 1324 with the completely random training set to only 189 for the training set at 3 semitone intervals. This fine pitch resolution, although not being necessary for recognition of American English vowels, is necessary for tonal languages and emotional prosody and could be stored in the templates if required. With fine pitch resolution stored in the templates, the model is also capable of detecting pitch at the accuracy of highly trained musicians (around 0.1 of a semitone) (Moore, 2003).

The percentage of the total templates (1324) stored for each vowel is shown in Fig. 2b. The most added templates are for the vowels /ʊ/ and /ɛ/, which both overlap other vowels in the first and second formant space (Peterson and Barney, 1952). These two vowels are among the worst performers in being classified in Peterson and Barney (1952) as well. The vowel with the least added templates is /i/, which also causes the least perceptual confusions with other vowels for human listeners in both quiet and noise (Peterson and Barney, 1952; Mi *et al.*, 2013). Also, /i/ is the least overlapping vowel in the formant space as seen in Fig. 1. This vowel was expected to require less templates and cause less confusions than any other vowel. Fig. 2c shows the percentage that each vowel contributes to the total number of misclassifications. This shows that the vowels that overlap the most not only require the most templates for correct recognition but also cause the most false classifications.

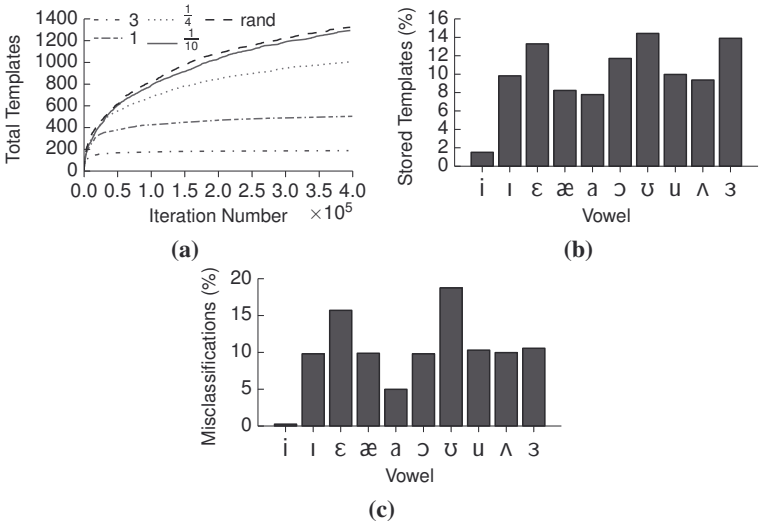

Fig. 2: Vowel training performance. (a) Total number of templates stored compared to the training iteration number. This is training with fundamental frequency semitone interval resolution of 3, 1, $\frac{1}{4}$, and $\frac{1}{10}$ semitones. Also training with random fundamental frequencies. Each new template corresponds to a misclassification error in the training set. (b) The percentage of the total stored templates for each vowel. The training is using random fundamental frequencies shown in Fig. 2a. (c) The percentage of the total misclassifications for each vowel. A misclassification means that a vowel is classified instead of the intended vowel that should be classified for that input. The training is using random fundamental frequencies shown in Fig. 2a.

The Klatt-synthesized template database from the randomly selected fundamental

frequencies is then tested with Klatt-synthesized vowels with babble speech added at SNRs of 20 dB, 10 dB, 5 dB, 0 dB, and −5 dB. The results using 100 test inputs for each vowel (1000 total inputs) at SNRs of 20 dB, 10 dB, 5 dB, 0 dB, and −5 dB are 99.7%, 99.6%, 97.0%, 77.6%, and 54.0%, respectively. The Klatt-synthesized template database is also tested against two male native English speakers' recordings of four vowels (/ae, ɜ, i, u/) for a total of 8 vowel recordings. The results with no noise show that all 8 of the recorded vowels are classified correctly when using the Klatt-synthesized template database. Although this is a very small sample size of recorded speech, this is a promising result.

CONCLUSION

The model trains to near 100% recognition performance on the 10 Klatt-synthesized vowels with a total of 1324 templates stored. The template database can be reduced by 86% with only a 3% loss in recognition accuracy by using sparse fundamental-frequency resolution. This reduced storage requires only 12 Mb of memory, which is well within the memory storage requirements to compute on parallel architectures such as FPGAs and GPUs. Furthermore, a 10-ms input passed through the model with 300 filter channels compared to a spectral template requires roughly 2 μs. The total time required for the recognition decision would then be dependent on the particular hardware chosen but is roughly the comparison for one spectral template (2 μs) times the number of templates divided by the number of processors. Therefore, the recognition decision does not become a very time-limiting step in the computation on a parallel architecture with sufficient cores, and the model can perform with near real-time performance. The model also performs exceptionally well when tested with Klatt-synthesized vowels with babble noise added at SNRs of 20 dB, 10 dB, 5 dB, 0 dB, and −5 dB. The vowel recognition rates at each SNR are 99.7%, 99.6%, 97.0%, 77.6%, and 54.0%, respectively. Furthermore the Klatt-synthesized vowel template database correctly recognizes recorded speech from two male speakers for the four vowels (/ae, ɜ, i, u/). The further exploration of the computational mechanisms in the model may elucidate how the brain adapts to learn language.

REFERENCES

De Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., and Van Compernolle, D. (**2007**). "Template-based continuous speech recognition", IEEE T. Audio Speech, **15**, 1377-1390.

Deng, L. and Strik, H. (**2007**). "Structure-based and template-based automatic speech recognition – Comparing parametric and non-parametric approaches", in *Interspeech 2007*, pp. 2608-2611.

Fairbanks, G., and Grubb, P. (**1961**). "A psychophysical investigation of vowel formants", J. Speech Hear. Res., **4**, 203-219.

Hawks, J.W., and Miller, J.D. (**1995**). "A formant bandwidth estimation procedure for vowel synthesis", J. Acoust. Soc. Am., **97**, 1343-1344.

Hillenbrand, J., and Gayvert, R.T. (**1993**). "Identification of steady-state vowels synthesized from the Peterson and Barney measurements", J. Acoust. Soc. Am., **94**, 668-674.

Kalinli, O., Seltzer, M.L., Droppo, J., and Acero, A. (**2010**). "Noise adaptive training for robust automatic speech recognition", IEEE T. Audio Speech, **18**, 1889-1901.

Klatt, D.H. (**1980**). "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am., **67**, 971-995.

Klatt, D.H., and Klatt, L.C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am., **87**, 820-857.

Krishnamoorthy, K., and Mathew, T. (**2009**). "The multivariate normal distribution", in *Statistical Tolerance Regions: Theory, Applications, and Computation* (John Wiley & Sons, Inc.), pp. 225-247.

McLachlan, N. (**2009**). "A vomputational model of human pitch strength and height judgments.", Hear. Res., **249**, 23-35.

McLachlan, N., and Wilson, S. (**2010**). "The central role of recognition in auditory perception: a neurobiological model", Psychol. Rev., **117**, 175-196.

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

Mi, L., Tao, S., Wang, W., Dong, Q., Jin, S.-H., and Liu, C. (**2013**). "English vowel identification in long-term speech-shaped noise and multi-talker babble for English and Chinese listeners", J. Acoust. Soc. Am., **133**, EL391-EL397.

Moore, B.C.J. (**2003**). *An Introduction to the Psychology of Hearing*, 3rd Ed. (Academic Press).

Neel, A.T. (**2008**). "Vowel space characteristics and vowel identification accuracy", J. Speech Lang. Hear. Res., **51**, 574-585.

Oliphant, T.E. (**2007**). "Python for scientific computing", Comput. Sci. Eng., **9**, 10-20.

Pearce, D., and Hirsch, H.-G. (**2000**). "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in *ISCA ITRW ASR-2000* (Paris, France), pp. 181-188.

Peterson, G.E., and Barney, H.L. (**1952**). "Control methods used in a study of the vowels", J. Acoust. Soc. Am., **24**, 175-184.

Rabiner, L.R. (**1989**). "A tutorial on hidden Markov models and selected applications in speech recognition", P. IEEE, **77**, 257-286.

Schouten, B., Gerrits, E., and Van Hessen, A. (**2003**). "The end of categorical perception as we know it", Speech Commun., **41**, 71-80.

Slaney, M. (**1993**). "An efficient implementation of the Patterson-Holdsworth auditory filter bank", Apple Computer Technical Report #35.

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