A comparison of formant and CNN models for vowel frame recognition

Ondrej Šuch^{1,2}, Santiago Barreda³, and Anton Mojsej²

 Mathematical Institue, Slovak Academy of Sciences 955 01 Banská Bystrica, Slovakia ondrejs@savbb.sk,
Žilinská Univerzita v Žiline, Univerzitná 8215/1, 010 26 Žilina
University of California Davis, 956 16 CA, USA

Abstract: Models for the classification of vowels are of continuing interest in both phonetics and for the development of automatic speech recognition (ASR) systems. The phonetics researchers favor linear classifiers based on formants, whereas ASR systems have adopted deep neural networks in recent years. In our work we compare the performance of several kinds of convolutional neural networks (CNN) with the linear classifier on the task of classifying short vowel frames from TIMIT corpus.

Our primary hypothesis was that the CNN models would prove significantly more precise than linear models, including during consonant-vowel and vowel-consonant transitions, while obviating the inherent difficulties of formant tracking. We confirmed the hypothesis, although the improvement was modest.

Our secondary goal was to investigate the possibility to mitigate loudness sensitivity of CNN models, and determine whether the mitigation would have deleterious effect on the classification performance of CNN models. Our experiments indicate that the loudness invariant CNN architecture performs equally well to traditional spectrum based convolutional models.

Keywords: MNIST, convolutional network, pairwise coupling, one-on-one classification, binary classification, dropout

1 Introduction

The difference in a single English vowel may convey nearly a dozen different meanings (e.g. hVd examples in [1]). Models that people use to differentiate vowels have been under study for many decades, both in phonetics as well as in automated speech recognition (ASR) systems. Recognition models can be broadly divided between two paradigms based on their use of the dynamic information in vowel sounds.

First, one may idealize a vowel as a stationary sound. This is a reasonable approximation since speakers can stretch almost any vowel to many times its ordinary duration without much effort. This view was adopted for instance in classical studies of English vowels in [2, 1], or [3].

Second, one may view a vowel as a dynamic entity evolving in time to both accommodate its acoustic surroundings as well as to express inherent vowel dynamic. This view fits much better to acoustics observed in conversational speech, where higher articulation rates lead speakers to shorten stationary vowel centers to the point of nearly eliminating them. This view was already hinted on in [2], reiterated in [3], and thoroughly examined in [4].

A model belonging to the first paradigm is sometimes called a frame recognition model, since it can be trained based on extracted vowel windows (frames), and it can categorize a vowel given a previously unseen sound window (frame). We will call a model belonging to the second paradigm a segmental model.

There are two principal methodologies that are used to create frame recognition models. For phonetics, the typical approach is to create a linear model based on the first three formants, which are known to be key determinants of a vowel quality. The second methodology is a powerful class of classification models that has garnered much attention in automated speech recognition systems, namely convolutional neural networks (CNN) [5, 6].

The goal of this work is to compare, both qualitatively and quantitatively, classification results obtained by the classical phonetic approach based on formants and the newer method based on convolutional neural networks with the eye on suitability of adoption of CNN for phonetic analyses. A key potential problem is the known dependence of deep neural network models on large amount of data [7]. While large amounts of data are commonly used in ASR research, phonetics datasets are usually much smaller.

2 Dataset

The performance of classifiers is easiest to investigate in cases when they are faced with a difficult recognition problem. A well-known pair of vowels that repeatedly shows up in listening experiments as difficult to distinguish is the aa/ao pair [1, 3, 8]. Therefore we looked for vowels that could be confused with this pair of vowels.

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

word	aa	ae	ah	ao	ax	eh	er	OW	uh
all	24	0	0	850	0	0	0	2	0
dark	858	0	10	6	0	2	0	0	0
wash	408	2	8	454	0	0	2	0	2
water	104	0	34	724	4	0	0	0	10

Table 1: TIMIT annotations of the first syllables in selected words in SA1 sentences.

word	ah	ao	ax	ix	ow	oy	uh
don't	10	2	2	2	856	0	4
oily	0	90	0	0	66	674	0

Table 2: TIMIT annotations of the first syllables in selected words in SA2 sentences.

We opted to use the TIMIT corpus [9], which has been the focus of many automated speech recognition studies. Compared to single syllable laboratory recordings, the corpus has the advantage of being closer to natural speech, and it comes with annotations dividing speech into individual phonemes.

We used vowel segments of male speakers from SA1 and SA2 sentences of the TIMIT corpus (we exluded female speakers due to the well-known problem of F1 determination for voices with high fundamental frequency, as well as due to likely separate models required for different sexes [10, 11]). The segments were the first vocalic segments in the words "dark", "wash", "water", "all", from SA1 sentences and "don't", "oily" from SA2 sentences. We excluded segments from those instances of words "oily", which were annotated with three vocalic segments. Tables 1 and 2 indicate phonetic annotation of the segments chosen by the TIMIT authors. The tables indicate that the words are prototypes respectively for segments /ao/, /aa/, /ow/ and /oy/.

In this work we used classes implied by the words, rather than the classes provided by TIMIT annotators. The primary reason is that those classes are less ambiguous, and possibly more reliable [12]. This approach also matches current trend in ASR to directly output lexical labels [13, 14, 15]. Our decision makes little difference for words "all", "dark", "don't", "oily", because these are typically annotated with a single TIMIT phoneme label. But it may affect classification results of the phonemes from words "wash", "water", which were labelled mostly by /aa/ and /ao/ annotations in TIMIT corpus.

The analysis window was set to be 24ms (384 samples). For each vowel altogether 21 frames were extracted equally spaced within the segment duration as indicated by TIMIT segmentation boundaries. Thus there were altogether 109410 samples divided among TRAIN and TEST datasets approximately in ratio 2:1. Extraction of formants and LPC spectra was performed using phonTools R package [16]. We normalized features for CNN, LSTM and k-NN models, which we will describe in detail in the next section.



Figure 1: Time evolution of the medians of the first three formants (solid colors). Data of one speaker (MREB0) is indicated in broken lines.

3 Classification models

Model creation was done using R software [17] and Keras machine learning library [18].

3.1 Frame classification models

We considered three types of frame classification models. The first was the logistic generalized linear model (GLM) [19, 20], which is of kind commonly used in phonetic studies of vowel qualities. The second one was convolutional neural network [21], which is a powerful classifier used in image processing, but also commonly applied in deep learning ASR systems. We expected that a linear model may have a problem capturing consonant-vowel and vowel-consonant transitions at the starts and ends of the segments. Therefore we included also k-nearest neighbor (k-NN) non-parametric classifier [19, 20], which should be able to learn nonlinear boundaries.

Neural network models Currently, search for neural network architectures is an active research field. We used ad hoc designs as shown in Figure 2.

For convolutional networks we used architectures shown in the upper part of Figure 2. All convolutional layers were 1-D with kernel of length 3; we used 5 neurons on the first layer and 10 on the second. ReLU nonlinearity was used both in convolutional and dense layers. The dense layer had 32 neurons. Finally, we distinguish several variations of the model based on provided spectral input:

- CNN-A. The input was 192 point Fourier logperiodogram.
- CNN-B. The input was LPC spectrum of order 19, sampled at 192 equally spaced frequencies.



Figure 2: Schematic layout of the three deep neural network architectures considered in the paper.

- CNN-C. The input was as in CNN-B, but intensity normalized per equation (2) below.
- CNN-D. The input was as in CNN-C but limited to components below 3,5kHz.

The purpose of the variations was to try providing data to CNN which is closer to data that formant models use (e.g. the third formant F3 almost never exceeds value 3,5kHz).

Finally, let us explain the rationale behind intensityinvariant models CNN-C and CNN-D. If the sound source is further away from the listener, obviously the sound is quieter. In the context of frame classification, there should be no change in prediction if the sound envelope is uniformly shifted by an equal constant. Admittedly, the prediction may change, if more context were available, because then a decrease of intensity may constitute a phonological contrast (e.g. syllabic stress). Let us note that any model based on formants satisfies this invariance property, by virtue of ignoring formant amplitudes. To achieve an equivalent property with CNN, one can transform the input tensor. Thus instead of supplying input vector **s** representing log-magnitudes of spectral components

$$\mathbf{s} = (s_1, s_2, \dots, s_M) \tag{1}$$

we provide as input the vector $N(\mathbf{s})$ defined as follows:

$$N(\mathbf{s}) := (s_2 - s_1, s_3 - s_1, \dots, s_M - s_1)$$
(2)

The key feature of this transform is that the sounds differing purely in intensity, say by d decibels, map to the same input vector since

$$N(s_1 + d, s_2 + d, \dots, s_M + d) = N(s_1, s_2, \dots, s_M)$$
(3)

Moreover, the topographic distribution of input vector components, which is important for CNN, is maintained, and only a very low frequency component is missing, which should not be relevant for vowel identification.

3.2 Segment classification models

We considered two kinds of segmental models. First, we supplied frame information from two frames: the onset at 10% and the offset at 90% position. This is inspired by research of Nearey and Morrison [22], who indicated that onset + offset data fits well with perceptual experiments. We considered two implementations of the model, a CNN with stacked onset+offset spectral information as shown in Figure 2 upper right, and logistic regression (GLM) model.

Second, we trained LSTM (long short-term memory) network [23] with Fourier and LPC spectra (denoted respectively LSTM-A and LSTM-B) with architecture as shown in the bottom of Figure 2. The LSTM layer used 4 units.

4 Results

We have opted for pairwise classification paradigm, where we classify frames (or segments) from a pair of words at a time. This decision is motivated primarily by our desire to elucidate strengths and weaknesses of various models, which may be confounded in the multiclass setting. We note that the paradigm has analogues and applications in phonetics (concept of minimal word pairs [24], forced A/B choice experimental design [25], or comparing accents by contrasting pairs of vowels in various words [26]).

4.1 Summary of accuracy with respect to location

Interpretability for classification models is highly desirable for recognition models in phonetics. For frame classification models it is possible to analyze their performance based on the position within vowel. Summary analysis is presented in Figure 3, from which we draw the following conclusions:

Paradigm	Model type	Model	Average
1 uruurgin	Model type	wieder	accuracy
Frame classifi- cation models	convolutional	CNN-A	83.2 %
	neural network models	CNN-B	83.2%
		CNN-C	82.1 %
		CNN-D	83.1%
	formant-based	k-NN	74 %
	models	GLM	73.9 %
Segmental nodels	onset+offset	CNN-A	96.8 %
	models	GLM	95.4 %
	recursive neural	LSTM-A	96.9 %
	network models	LSTM-B	96.2 %

Table 3: Average accuracy of frame and segmental models. Note that for frame classification, the input to CNN-A is a 1D tensor consisting of a spectrum, whereas in segmental classification, we joined onset and offset spectra to form a 2D tensor as shown in upper right in Figure 2.

- CNN models are uniformly superior to the other two classes of models (k-NN and GLM).
- There is negligible performance difference among models CNN-A through CNN-D.
- Formant models (GLM as well as k-NN) in several cases closely follow convolutional neural networks (e.g. all-dark, or dark-wash, or dark-oily). This indicates that very often formant positions are sufficient for classification.

4.2 Accuracy improvement with segmental models

We trained also segmental models in order to gauge the effect of restricting classification to frames rather than segments. The results are summarized in Table 3.

Clearly, segmental models can make good use of extra information, since all six segments are well classified. The most likely explanation is the ability of segmental models to discriminate consonant-vowel/vowel-consonant transitions.

4.3 Multi-class classification

For ASR applications, especially for the systems based on hidden Markov models [27], it is desirable to obtain multi-class likelihoods. Passage from pairwise models to multi-class ones can be achieved by any of multiple pairwise-coupling techniques [28, 29, 30]. We performed this coupling using the second method suggested in [29], which is widely used in machine learning [31]. The multiclass models were composed of 6 pairwise models trained on frames from the first syllables of all possible pairs of words "all", "dark", "don't", and "oily". We used those four words, because they are the prototypes of the four phonemes /ao/, /aa/, /ow/, /oy/, that account for most TIMIT annotations of our dataset (see Tables 1 and 2). The likelihood plots are shown in Figure 4.

It is hard to argue which model predicted likelihoods that are more correct, since there is no ground truth. Subjectively, one may feel that CNN models are too certain about their predictions, especially since the predictions are considerably less smooth than that of the formants based GLM model. The small size of the dataset may be a possible cause.

5 Conclusion

Our experiments indicate that it is feasible to construct CNN models for frame classification for medium size phonetic corpora. The resulting models showed superior performance to formant-based models even without extensive parameter optimization, which is in accordance with findings of a recent study on Korean vowels [32]. It is also possible to make CNN models intensity-invariant akin to formant-based models without noticeable loss of performance. The only downside we observed (Section 4.3) was that the likelihoods predicted by CNN models lacked temporal smoothness compared to the formant model.

Acknowledgment

Work on this paper was partially supported by grant VEGA 2/0144/18. We are thankful to Peter Tarábek for his suggestion to use the stacked input architecture for onset+offset CNN models, which proved much more accurate than the previously investigated branched architecture.

References

- Peterson, G.E., Barney, H.L.: Control methods used in a study of the vowels, J. Acoust. Soc. Am. 24 (1952) 175–184
- [2] Steinberg, J. C., Potter, R. K.: Toward the specification of speech, J. Acoust. Soc. Am. 22 (1950)
- [3] Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels, J. Acoust. Soc. Am. 97 5 (1995) 3099–3111
- [4] Morrison, G. S., Assmann, P. F.: Vowel inherent spectral change, Springer (2012)
- [5] Abdel-Hamid, O., Mohamed, A-R., Jiang H., Deng L., Penn G., Yu, D.: Convolutional neural networks for speech recognition, IEEE/ACM Transactions On Audio, Speech, and Language Processing, 22 10 (2014) 1533–1545
- [6] Tóth, L.: Convolutional deep rectifier neural nets for phone recognition, INTERSPEECH, (2013)
- [7] Hestess, J. et al: Deep learning scaling is predictable, empirically, https://arxiv.org/abs/1712.00409
- [8] Zahorian, S.A., Jagharghi, A.J.: Spectral-shape features versus formants as acoustic correlates for vowels, J. Acoust. Soc. Am. 94 4 (1993)



Figure 3: Average accuracy of various classification models based on the position of frame within a vowel, conditioned on pairs of words. Dotted line indicates 50% accuracy rate.



Figure 4: Frame likelihoods produced by multiclass models for speaker MREB0.

- [9] Garofolo, J. S. et al: TIMIT Acoustic-phonetic continuous speech corpus LDC93S1. Web download. Philadelphia: Linguistic Data Consortium (1993)
- [10] Nearey T.M., Assmann, P.F.: Probabilistic 'slidingtemplate' models for indirect vowel normalization, In Experimental Approaches to Phonology, Eds. M.J. Solé, P.S. Beddor, and M. Ohala, Oxford University Press (2007) 246– 269
- [11] Barreda S., Nearey, T.M.: The direct and indirect roles of fundamental frequency in vowel perception, J. Acoust. Soc. Am. 131 (1) (2012) 466–477

- [12] Šuch, O., Beňuš, Š, Tinajová, A., A new method to combine probability estimates from pairwise binary classifiers, ITAT 2015 (2015)
- [13] Graves, A., Jaitly N.: Towards end-to-end speech recognition with recurrent neural networks, International conference on machine learning (2014)
- [14] Chan W., Jaitly N., Le Q.V., Vinyals O.: Listen, attend and spel: a neural nework for lage vocabulary conversational speech recognition, ICASSP 2016 (2016)
- [15] Park D.S., Chan W., Zhang Y., Chiu C-C., Zoph B., Cubuk E.D., Le Q-V.: SpecAugment: A simple data augmentation method for automatic speech recognition, arXiv:1904.08779 (2019)
- [16] Barreda, S.: phonTools R package, version 0.2-2.1 (2015)
- [17] R software for statistical computing, https://www.rproject.org
- [18] Chollet, F. et al, Keras, https://keras.io (2015)
- [19] James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, Springer (2013)
- [20] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning – Data Mining, Inference, and Prediction. New York: Springer (2009)
- [21] LeCun Y., Bengio, Y.: Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, The Handbook of Brain Theory and Neural Networks. MIT Press (1995)
- [22] Morrison G. S., Nearey, T. M.: Testing theories of spectral change, J. Acoust. Soc. Am., **122** (2007), EL15–EL22
- [23] Hochreiter, S. Schmidhuber, J.: Long short-term memory, Neural computation, 9(8) (1997), 1735–1780
- [24] Ladefoged, P.: Vowels and consonants, Blackwell (2001)
- [25] Jeffress, L. A.: Masking, in J. V. Tobias ed., Foundations of modern auditory theory, Academic Press (1970)
- [26] Huckvale, M.: ACCDIST: a metric for comparing speaker's accents, In Proc. Int. Conf. on Spoken Lang. Proc., Juju Island, Korea (2004)
- [27] Rabiner, L. R.: A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 77, no. 2 (1989) 257–286
- [28] Hastie, T., Tibshirani, R.: Classification by pairwise coupling, The Annals of Statistics, 26(2) (1998) 451--471.
- [29] Wu T-F., Lin C-J., Weng R.C.: Probability estimates for multi-class classification by pairwise coupling, Jounral of Machine Learning Reserach (2004) 975–1005
- [30] Šuch, O., Barreda, S.: Bayes covariant multi-class classification, Pattern Recognition Letters, 84 (2016) 99–106
- [31] Chang, C-C., Lin C-J., LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent System and Technology*, vol 2., issue 3 (2011)
- [32] Kim, H., Lee, W. S., Yoo, J., Park, M., Ahn, K.H.: Origin of the higher difficulty in the recognition of vowels compared to handwritten digits in deep neural networks, J. Korean Physical Soc. 74, No.1 (2019) 12–18