RNN Classification of English Vowels: Nasalized or Not

Ling Liu and **Mans Hulden** and **Rebecca Scarborough**
Department of Linguistics
University of Colorado
first.last@colorado.edu

1 Introduction

Vowel nasality is phonemic in languages like French, Hindi and Navajo while allophonic in languages like English, Mandarin and Arabic. It is an important linguistic phenomenon perceived and used by listeners even in languages where it is coarticulatory and does not contrast meaning (Lahiri and Marslen-Wilson, 1991; Beddor et al., 2013; Scarborough, 2013).

Generally speaking, vowel nasality results from lowering the velum to allow airflow to escape through the nasal tract on top of oral articulation. In the case of coarticulation, if a vowel is preceded by a nasal consonant, the velum keeps lowered after the nasal pronunciation and causes the vowel nasalization as a carryover effect; when a vowel is followed by a nasal consonant, the velum lowers during the pronunciation of the vowel in anticipation of the nasal articulation, resulting in the anticipatory vowel nasalization. The nasality of a vowel varies in the time course due to factors such as context the sound appears in. It has been claimed that carryover nasality (i.e. vowel nasalization in the syllabic structure of NVC[1]) is less strong than anticipatory nasality (i.e. in a CVN syllabic structure) (e.g. Moll (1962), Ohala (1971)), and nasality involving both carryover and anticipatory effects (i.e. in an NVN context) is the strongest (e.g. Styler (2015)). Cohn (1990) found that in English the nasal airflow for carryover nasality has a decreasing tendency through the articulation of the vowel, for anticipatory nasality the nasal airflow has an increasing tendency, and for nasality involving both carryover and anticipatory effects the nasal airflow is more flat than in the other two contexts.

In comparison to the description of vowel nasality from the articulatory perspective, the acoustic and perceptual description and understanding of the phenomenon is more elusive. Styler (2015) summarized 29 features of nasality discussed in the literature and constructed two feature-based classifiers, a random forest classifier and a support vector machine (SVM) classifier, to categorize vowels into either nasal or oral, as a way to test which features or feature combinations are useful for nasality perception and measurement. The performances of both classifiers on the classification of English vowels in NVN and CVC contexts, were the best when all the 29 features were used, indicating the complexity of the phenomenon. In addition, most measurements about the nasality features were taken at the 1/3 and 2/3 duration points of each vowel, and thus most of the information the classifiers use is discrete, even though speech is continuous.

Recurrent neural networks (RNNs) are a type of deep learning architecture specifically designed to take advantage of sequential information (Lipton et al., 2015). A neural network (NN) classifier is not necessarily a good model to evaluate the contribution of any particular feature in isolation or feature combinations as a feature-based classifier can do. However, it has the advantage of making use of information from simpler features, freeing us from high-level feature engineering: we don't have to identify, isolate and measure features such as formant frequencies, amplitude and bandwidth at certain points. What's more, because it is good at taking advantage of sequential information, it can potentially be a stronger model of the phenomenon with a more holistic view of the speech signal.

For the current work, we constructed a vanilla RNN (Elman, 1990) classifier as an NN baseline model and a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) classifier as

[1]N stands for nasal consonant, V stands for vowel, and C stands for non-nasal consonant

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 318-321.
New York City, New York, January 3-6, 2019

a more advanced model. The classifiers are binary, which take as input the Mel Frequency Cepstral Coefficient representations (MFCCs) (Davis and Mermelstein, 1980) of English vowels isolated from words with CVC, NVN, CVN, or NVC syllabic structures and classify the vowels into the classes nasalized/non-nasalized. Three groups of binary classification were conducted: CVC vs NVN, CVC vs CVN, CVC vs NVC. The contributions and results of this paper are as follows:

- The LSTM classifier is very strong in classifying vowels as nasalized or not. It achieves an accuracy of 94.50% on CVC vs NVN which is almost 10% higher than the SVM classifier, the state-of-the-art feature-based classifier.
- The good classification result of the neural models suggest that MFCCs, a perceptually motivated representation of speech signal commonly used for automatic speech recognition, retain very informative components as regards nasalization, though information may be lost in various ways in the process of deriving the MFCCs (Huang et al., 2001; Zheng et al., 2001).
- The performance of the neural models supports phonetic claims about the degree of nasality and indicates that nasality study should not only take discrete measurements at discrete points but also take a more holistic view by combining features throughout the duration of the relevant speech signal.

2 Experiments

The data are a subset of the data collected by Styler (2015), which are sound files with corresponding textgrid annotations marking word boundaries and vowel boundaries on different tiers. Each sound file is an English monosyllabic word experimentally elicited in isolation. The consonants immediately before and after the vowel are /b/, /d/, /m/, or /n/. For each of the four contexts, the training set consists of 665 tokens, the test set includes 100 tokens. A development set is held out only for the CVC vs CVN context contrast, where 70 tokens are provided for each context. A Praat (Boersma and Weenink, 2018) script is used to automatically extract the vowel in each sound file as is annotated in the corresponding textgrid. The vowels in isolation are then converted into MFCCs using the

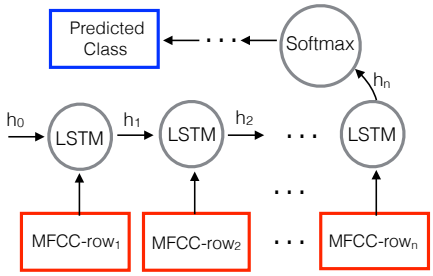


Figure 1: Architecture of the LSTM classifier. *MFCC-row$_i$* is the i*th* row in the MFCC representation of the vowel. h_i is the output generated by the i*th* step in the hidden layer, which will be the input to the next step together with the next row of MFCCs. The final output of the LSTM, i.e. h_n will be the condensed representation of the vowel and will be the input to the softmax function for classification.

`python_speech_features` library[2] with the signals framed into 25ms frames with a 10ms window step. The final representation of a vowel is a matrix with different numbers of rows depending on the length of the vowel where each row contains 13 MFCCs.

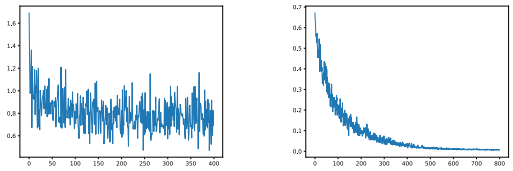
The RNN networks iterate over the MFCC matrix representation of the vowel one row after another, which outputs a condensed representation in the end. This condensed representation is input to a softmax function which produces an output as to whether the vowel is nasalized or not. Figure 1 illustrates the architecture of the RNN classifiers, with LSTM as an example. The architecture of the baseline RNN model differs from Figure 1 only in the hidden layer where it uses vanilla RNN rather than LSTM.

For both the baseline vanilla RNN classifier and the LSTM classifier, there is only one layer in the network, the optimization method is stochastic gradient descent (SGD), and the loss function is the negative log likelihood loss (NLLLoss). All weights are initialized to zeros. The hyperparameters for each of the classifiers is tuned on the development set. The setting of hyperparameters for the baseline vanilla RNN classifier is as follows: the hidden layer is of size 300, the learning rate is 0.0005, and 40,000 iterations were run with each iteration randomly picking a sound file from the training data. For the LSTM classifier, the hyper-

[2] https://github.com/jameslyons/python_speech_features

Context contrast	SVM	Vanilla RNN	LSTM
CVC-vs-NVN	84.76	71.00	**94.50**
CVC-vs-CVN	-	73.00	**90.00**
CVC-vs-NVC	-	51.00	**81.00**

Table 1: Accuracies(%) of classifiers for different context contrasts



(a) RNN: CVC vs NVN (b) LSTM: CVC vs NVN

Figure 2: Example plots of average losses over every 100 iterations during training process of the vanilla RNN classifier and the LSTM classifier for different context contrasts for CVC vs NVN. The other two context contrasts have the same pattern.

parameters are as follows: the hidden layer size is 300, the learning rate is 0.0005, and the number of iterations is 80,000. The systems are implemented using PyTorch (Paszke et al., 2017).

Experiments were conducted with three types of oral-nasal context contrasts. In other words, three groups of binary classification were conducted: one is to classify whether the vowel is between two oral consonants or is between two nasal consonants (i.e. CVC vs NVN), another is to classify whether the vowel is between two oral consonants or is immediately preceded by an oral consonant and immediately followed by a nasal consonant (i.e. CVC vs CVN), and the third is to classifier whether the vowel is between two oral consonants or is immediately preceded by a nasal consonant and immediately followed by an oral consonant (i.e. CVC vs NVC). Vowels between oral consonants are not nasalized, and vowels immediately preceded or followed by a nasal consonant are considered nasalized. The same parameters are used for classification in all context contrasts, though they were tuned only for CVC vs CVN. For each context contrast, we trained an ensemble of ten models and used majority voting to decide the final classification, i.e. the class which was predicted by the greatest number of models is the final classification decision of the ensemble.

3 Results and Discussion

Table 1 provides the accuracies of the RNN classifiers for different context contrasts, as well as the best performance of Styler (2015)'s models. Styler (2015) classified English nasality only for vowels in tokens with CVC and NVN contexts. The best performance was achieved by the SVM classifier using all 29 features in his experiments.

The baseline RNN model, i.e. the vanilla RNN classifier performs worse than the feature-based models by a large margin, while the improved RNN model, i.e. the LSTM classifier achieves an accuracy of about 10% higher than the best feature-based model for the same context contrast, i.e. CVC vs NVN. This indicates that MFCCs contain relevant information for nasality despite its lossy representation. This finding supports the use of MFCCs for automatic speech recognition, as they appear to retain enough of the signal to allow for subtle distinctions.

The vanilla RNN architecture has the vanishing gradients problem (Pascanu et al., 2013). As a result, the information at the beginning of the vowels in the output of the RNN after going over the MFCC matrix will have very little effect on the classification. The LSTM architecture can handle long-distance information better (Hochreiter and Schmidhuber, 1997) and thus produces better combinations representing information across the vowel. It is observed that the LSTM classifier outperforms the vanilla RNN classifier for every context contrast. This indicates that the features across the vowel, rather than at a certain place, be it the beginning or the end, are important for nasality measurement and perception.

For the LSTM classifiers which can take good advantage of information across the vowel, the classification accuracies become increasingly higher for context contrasts from CVC vs NVC, CVC vs CVN to CVC vs NVN. This agrees with the phonetic claim that NVN vowels in English are most nasalized, CVN vowels are less nasalized, and NVC vowels are least nasalized (Moll, 1962; Ohala, 1971; Styler, 2015). However, this pattern is not supported by the classification result of the vanilla RNN architecture, since the baseline NN model achieves the highest accuracy for CVC vs CVN. This may be because the hyperparameters were tuned for the CVC vs CVN context contrast. We can see in Figure 2 that the LSTM classifiers converge very well while the vanilla RNN clas-

sifiers do not. This indicates that the pattern of nasality degree supported by the LSTM classifier results is reliable, and can't be overthrown by the vanilla classification results.

4 Conclusion

We developed a baseline NN classifier of the vanilla RNN architecture, and an improved NN model of the LSTM architecture, to classify English vowels as to whether they are nasalized or not. The RNN models take as input the MFCC representations of the vowel, and do binary classifications. The models were trained and tested on three types of vowel context contrasts: CVC vs NVN, CVC vs CVN, and CVC vs NVC. The LSTM classifier largely outperforms the SVM classifier which uses 29 carefully designed and measured features for the same task. This indicates that MFCCs contain the relevant part of the speech signal to discriminate nasality, and that nasality is represented in the vowel not only as discrete features at certain points but that there is also a holistic aspect to the detection, and that retaining the signal throughout the relevant segments is important for nasality perception and measurement. In addition, the performance of the LSTM classifiers is strongest for CVC vs NVN, followed by CVC vs CVN, and the worst for CVC vs NVC, a result that agrees with phonetic claims about strength of coarticulatory nasality.

Neural network models as currently used are not necessarily the best choice for examining which specific features are useful for detection of nasality since it is not easy to interpret what the neural models attend to and so the result does not directly contribute to our understanding of the details of the phenomenon. Future work may compare various ways of representing speech sounds as input to NN models to see which leads to better or worse classification. It is possible to incorporate the features proposed by Styler (2015) into the model. It is also worth exploring data with both acoustic and nasal airflow information, though we do not currently have access to such a corpus.

References

Patrice Speeter Beddor, Kevin B. McGowan, Julie E. Boland, Andries W Coetzee, and Anthony Brasher. 2013. The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, 133(4):2350–2366.

Paul Boersma and David Weenink. 2018. Praat: doing phonetics by computer [computer program], version 6.0.43. http://www.praat.org/.

Abigail C. Cohn. 1990. *Phonetic and Phonological Rules of Nasalization. UCLA Ph. D.* Ph.D. thesis, dissertation.[UCLA Working Papers in Phonetics 76].

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and signal Processing*, 28(4):357–366.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. 2001. *Spoken language processing: A guide to theory, algorithm, and system development*, volume 1. Prentice Hall.

Aditi Lahiri and William Marslen-Wilson. 1991. The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, 38(3):245–294.

Zachary C. Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Kenneth L. Moll. 1962. Velopharyngeal closure on vowels. *Journal of Speech, Language, and Hearing Research*, 5(1):30–37.

John J. Ohala. 1971. Monitoring soft palate movements in speech. *The Journal of the Acoustical Society of America*, 50(1A):140–140.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch.

Rebecca Scarborough. 2013. Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics*, 41(6):491–508.

Will Styler. 2015. *On the Acoustical and Perceptual Features of Vowel Nasality*. Ph.D. thesis, University of Colorado Boulder.

Fang Zheng, Guoliang Zhang, and Zhanjiang Song. 2001. Comparison of different implementations of MFCC. *Journal of Computer science and Technology*, 16(6):582–589.