Contribution of wavelet-based features to identification of overlapping Azerbaijani vowels using machine learning

Payam Ghaffarvand-Mokari

University of Eastern Finland, School of Humanities
payam.ghaffarvand.mokari@uef.fi

ABSTRACT

While vowels are generally found to be distinct based on their first two or three formants, Azerbaijani mid-front rounded /œ/, high-back unround /ɯ/, and high-front rounded /y/ are found to be largely overlapping. The present study examines the contribution of wavelet-based decompositional features to semi-automatic identification of these vowels using random forest and neural networks. From a data sample of 607 vowel tokens produced by eight male speakers, first three formants and six Shannon entropy wavelet packets were extracted and served as the training features for the classifiers. Results showed that the addition of the wavelet-based features to formant information improved the classification accuracies by 8–11% across the classifiers. The highest classification accuracy was 83.1% achieved by random forest.

Keywords: Azerbaijani, vowels, wavelet packets, formants, machine learning.

1. INTRODUCTION

The identification of robust acoustic parameters for the classification of speech sounds is crucial for developments of automatic speech recognition systems, as well as for basic research in different areas of phonetics and phonology. Vowels are especially important for automatic speech processing as they contain useful information for speech and speaker identification/verification, clinical assessment of speech/voice disorders, and for forensic applications. Therefore, a better understanding of important features in classification of speech sounds across the languages is important.

First noted by Householder in 1972 [1], Azerbaijani mid-front rounded /œ/ and high-back unround /ɯ/ are found to be largely overlapping on a F1–F2 space [1, 2]. To a lesser extent, the high-front rounded /y/ is also found to be overlapping with the high-back unround /ɯ/ [2]. This study aims to investigate whether further inclusion of wavelet-based decompositional features for training of the classification models improves the classification of these phonologically distant but acoustically similar vowels.

Wavelet decomposition is a mathematical technique that involves breaking down a signal into different frequency components using wavelets, which are small waves that are localized in both time and frequency domains.

1.1. Background

Researchers have studied different sets of features in classification of vowels of different languages. For instance, Krocil et al. [4] have used F1, F2, and zero-crossing rate (ZCR) in classification of Czech vowels. Their results from a heuristic classifier, revealed the classification accuracy ranged from 54.5% to 97.5% for different vowels.

For Hindi vowels, Biswas et al. [5] used different combination of F1, F2, F3, Gammatone Frequency Cepstral Coefficients (GFCC), and Mel-Frequency Cepstral Coefficients (MFCC) and found that using formants together with GFCC outperformed other combinations of features in different noisy conditions.

Classification of American English vowels in noisy and noise-free conditions has been evaluated using MFCCs and formants [6]. It was found that the classification accuracies based on the formant features and MFCCs were almost the same.

Korkmaz et al. [7] studied the classification of eight Turkish vowels using a feature set optimized by Genetic Algorithm. They used formants, energy, ZCR, MFCCs, and Wavelet Decomposition Shannon Entropy features as their input. Their results from a data set based on vowel tokens uttered by 10 male speakers showed that the feature vector optimized by Genetic Algorithm method reached 100% classification accuracy.

Among the limited number of studies on classification of Azerbaijani vowels are the studies by Ghaffarvand-Mokari and Werner [2] and Imamverdiyev and Sukhostat [3]. Ghaffarvand-Mokari and Werner [2] found that the classification accuracies using simple Linear Discriminant Analysis (LDA) with first two formants as predictors were very low for Azerbaijani /œ/-/ɯ/ vowels. They found that further inclusion of the third formant to the LDA models, improves the classifications accuracy but still the largest confusion is between /œ/-/ɯ/

vowels. The /y/ vowel was also found to be overlapping with /ɯ/. Imamverdiyev and Sukhostat [3] evaluated classification of Azerbaijani vowels based on MFCCs using Support Vector Machine (SVM). Their findings show that /œ/, /ɯ/, /y/, and /e/ vowels were poorly classified (with accuracy rates of 45.5% – 60%) compared to other vowels (82.6% – 95.7%).

1.2. Present study

The present paper aims at evaluating the contribution of Wavelet Decomposition Shannon Entropy features in identification of the /œ/, /ɯ/, and /y/ vowels using random forest and neural networks as classifiers.

The findings would help identification of a subset of acoustic features that carry most of the information for efficient classification of these vowels.

2. METHODS

2.1. Speech data

A part of previously recorded corpora for studying Azerbaijani fricative [8] was used as the speech data for the present study. The corpora were recorded from speakers of southern Azerbaijani dialect. Recorded data from 8 male speakers were used who had relatively balanced number of reliable tokens across /ɯ/, /œ/, and /y/ vowels. Recordings used for this study had the vowels in V position of consonant-vowel-consonant (C_1VC_2) contexts. The C_1 varied over nine different fricative consonants and C_2 was kept identical. The target words were embedded in a fixed carrier phrase “burdɑki _ kælmæsidi” (here is the _ word). Each target vowel was repeated 3 times, yielding a total of 648 tokens per speaker (3 vowels × 9 different C_1 consonants × 3 repetitions).

After removing unreliable and wrong pronunciations, a total of 607 tokens were submitted for the analyses. Recordings were done using a condenser microphone (Blue Bluebird, USA) microphone with a sound card (RME Fireface 800, Germany) connected to a PC computer through a controller card (VIA VT6307, Taiwan) in a sound-attenuated booth. Recordings were saved as wav files with sampling frequency of 44.1 kHz and 16-bit quantization. Tokens were segmented and labeled manually based on the waveforms and spectrograms. To minimize coarticulation effects from neighboring consonants, the middle 50% of the vowel duration were extracted and used for the analyses.

2.2. Feature extractions

The vowel parts of the signals were labelled manually in Praat [9] and formants were calculated using a Praat script with Burg algorithm with range of 0-5,000 Hz for five formants. Fig. 1 shows the spectrograms of the middle parts of the vowels uttered by a male speaker. F1, F2 and F3 values were then measured for each vowel by taking an average of the 50% of the vowel centred around the midpoint.

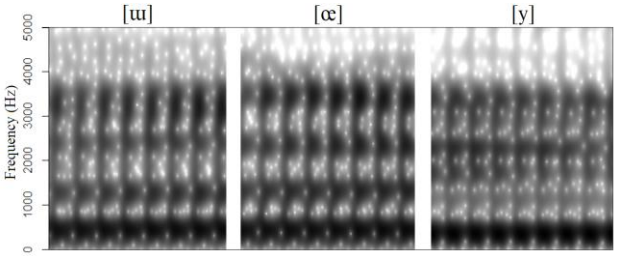

Figure 1: Spectrograms of the middle 50 ms of /ɯ/, /œ/, and /y/ vowels in a h-vowel-r context uttered by a male speaker.

Following the method used in [7], wavelet transform on the selected parts of the vowels were applied to obtain decompositional features in MATLAB. The ‘wavelet transform’ divides signal into sub-signals with low and high frequencies. A Daubechies-filtered 5-depth wavelet packet decomposition was used. Six features selected by considering d1-d5 and a5 sub-signals (Fig. 2).

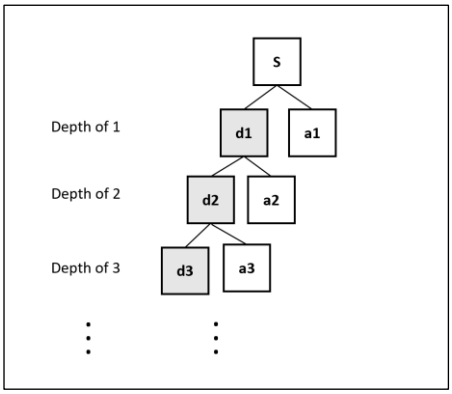

Figure 2: The structure of wavelet packet decomposition. Shaded boxes indicate selected sub-signals.

After having sub-signals, the Shannon entropies of them were calculated. Finally, Shannon entropy values of sub-signals resulted in the original signal’s decompositional features.

2.3. Classifiers

Random forest [10] and neural networks [11] were used to predict the vowels based on the formants and wavelet-based features.

Random forest with 300 decision trees were used to classify the vowels and to analyse the importance of each variable in correctly identifying the vowels.

A feed-forward neural network was used which included input, hidden, and output layers. In this network, the information moves in only one forward direction from the input layer. The neurons in each layer are connected to the neurons of the subsequent layer. Ten hidden layers used for the classifications. The input layer has one neuron for each variable and the output layer has one neuron for each predicted vowel.

The classifiers were trained on a randomly chosen 70% of the data and evaluated on the remaining 30% of the data. The training and testing data had a balanced number of the predicted vowels. Ten replications of each classification were done. Each replication had a randomly partitioned data into training and test subsets.

3. RESULTS

The /œ/ and /ɯ/ vowels, and to some extent /y/, were found to be overlapping on a F1–F2 space (Fig. 3).

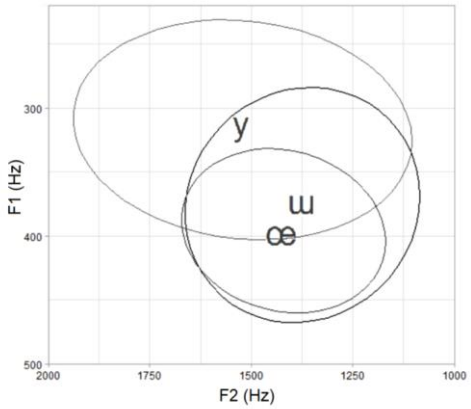

Figure 3: Distribution of the /œ/, /ɯ/, and /y/ vowels in a F1 × F2 (Hz) space with ellipses representing one standard deviations from the mean.

Results from 10 replications of the classifications using random forest and neural networks showed accuracy rates did not vary much between the classifiers for the models with F1 and F2 as predictors, but for the models with all three formants and those with formants and wavelet-based features, random forest resulted in higher accuracies (Fig. 4). The mean accuracy with F1 and F2 as the input variables was 63.6% (95% CI [60.8–66.4]) and 64.7% (95% CI [62.3–67.1]) for neural networks and random forest, respectively. After inclusion of F3, mean accuracy rate reached 69% (95% CI [67.1–70.9]) for neural networks and to 72% (95% CI [70–74.1]) for random forest. Finally, the inclusion of wavelet-based features further improved the mean accuracy for neural networks to 77.1% (95% CI [74.6-79.5]) and for random forest to 83.1% (95% CI [81.1–85.1]) with a maximum of 87.8% accuracy in one of the replications.

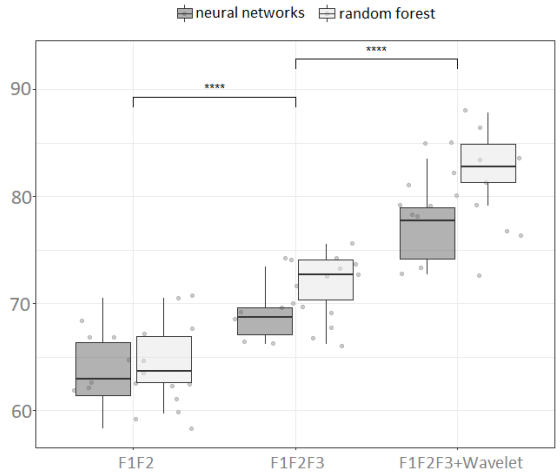

Figure 4: The performance (% of accuracy) of the random forest and neural networks as classifiers in identification of /œ/, /ɯ/, and /y/ based on ten replications with different sets of predictors.

The accuracies across the sets of features were compared by a series of Mann-Whitney Wilcoxon tests. As indicated in Fig. 4 by the three asterisks (***), the accuracy of F1F2 set was significantly different from F1F2F3 set for both classifiers (all $p < 0.001$). The accuracy of F1F2F3 set was also significantly different compared to F1F2F3 + wavelet-based features set for both classifiers (all $p < 0.001$).

3.1. Separate analyses for each vowel

Further, the accuracy of identifications was also analysed separately for vowels to explore how different sets of features contribute to correct identification of each vowel. Mean identification accuracies for /œ/, /ɯ/, and /y/ with random forest as the classifier are presented in Fig 5. The /y/ vowel was identified with mean accuracy of 81.5% (95% CI [78.3–85]) based on F1 and F2. The inclusion of F3 didn't result in noticeable improvement for the identification of /y/ (M = 82%, 95% CI [77.8–86.8]), yet further inclusion of the wavelet-based features resulted in better identification accuracy (M = 89.6%, 95% CI [86.6–92.5]).

For /ɯ/ vowel, the mean identification accuracy with F1 and F2 was 62.6% (95% CI [60.1–65]) and after inclusion of F3, it reached 71.7% (95% CI [69.4–74]). Further inclusion of the wavelet-based features resulted in mean accuracy of 81.8% (95% CI [79.4–84.1]).

For /œ/ vowel, the correct identification rate was close to a chance level[1] with F1 and F2 as predictors (M = 46.1%, 95% CI [41–51.3]). Inclusion of F3 to the models resulted in noticeable improvements in the

identification rates (M = 60.6%, 95% CI [55.5–65.7]). The inclusion of wavelet-based features also substantially improved the identification of /œ/ (M = 78.7, 95% CI [73.9–83.6]).

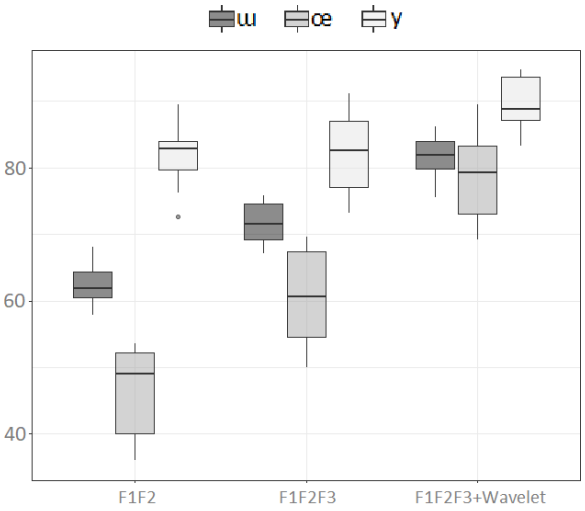

Figure 5: Identification accuracy (%) of /œ/, /ɯ/, and /y/ vowels using random forest as classifier based on ten replications.

3.2. Important features for classifications

To find out which features were the most important ones in identification of the vowels, an estimate of the importance of each predictor for random forest models were calculated. Fig. 6 shows the ranked importance of the features in terms of the mean decrease in the accuracy of the model when excluding a predictor. By a large extent F1 was the most important predictor, followed by F3, F2, d3, and a5.

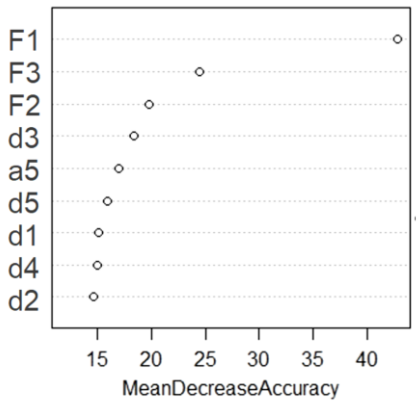

Figure 6: The mean decrease in the accuracy (%) of the random forest model when excluding a predictor (a high decrease means that the variable has salient predictive power)

4. DISCUSSION

This paper was set out to investigate whether formants and wavelet based acoustic cues contain enough information to correctly classify Azerbaijani overlapping /œ/, /ɯ/, and /y/ vowels. Further, two different computational classifiers – random forest and feed-forward neural networks – were evaluated in classifying the vowels using different sets of acoustic cues.

It was found that the accuracy of classifying the vowels based on F1 and F2 was quite low especially for /œ/ vowel. This is in line with findings of [2] that have found large overlap in classification of these vowels based on F1 and F2. Additionally, in line with the findings of [2], further inclusion of F3 to the predictors improved the overall classification accuracy, but it did not have noticeable effect on correct identification of /y/ vowel (Fig 5). This indicates that F3 is an important predictor for /œ/-/ɯ/ classification but not for /y/.

As for the main aim of this paper, wavelet-based features found to improve the classification accuracies by 8.1%–11.1% for random forest. This is in line with the findings Korkmaz et al. [7] which shows five wavelet-based features were selected by Genetic Algorithm as important features in the classification of Turkish vowels. Overall, wavelet-based features seem to contain useful information for classification of the Azerbaijani /œ/-/ɯ/-/y/ vowels. However, they do not contain enough information for optimal classifications. Since wavelet transforms captures both short-term and long-term variations in the signal's characteristics, it is possible that they capture (part of) the information important for listeners in discrimination these overlapping vowels.

Finally, the classification accuracies using random forest were higher than for neural networks. Since neural networks normally need larger data for better trainings of the models, the observed results are possibly due to relatively small size of the data set in this study. Random forests are found to be appropriate for analysis of small data sets and for taking the possible correlation of variables into account [12].

This study has some limitations, probably the most important being the relatively small sample size and that it only analyses data from male speakers. Future studies should investigate the classification of overlapping vowels using larger sample sizes and including vowels produced by female speakers, as well as employing other feature sets and classification algorithms. Future studies should also consider inclusion of all other Azerbaijani vowels and different word contexts in the classifications. Finally, subsequent studies could explore whether there are differences in the acoustic classification outcomes for the vowels of northern and southern Azerbaijani dialects.

5. REFERENCES

[1] Householder, F. 1972. Vowel overlap in Azerbaijani. In: Valdman, A. (eds), *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*. De Gruyter Mouton, 229–230.

[2] Ghaffarvand-Mokari, P., Werner, S. 2015. An acoustic description of spectral and temporal characteristics of Azerbaijani vowels. *Poznan Studies in Contemporary Linguistics*, 52(3), 503–518.

[3] Imamverdiyev, Y., Sukhostat, L. 2011. SVM based recognition of Azerbaijani vowels. In *5th International Conference on Application of Information and Communication Technologies (AICT)* IEEE, 1–4.

[4] Krocil, J., Machacek, Z., Koziorek, J., Martinek, R., Nedoma, J., Fajkus, M. 2018. Improved method of heuristic classification of vowels from an acoustic signal. *Turkish Journal of Electrical Engineering and Computer Sciences*, 26(6), 2900–2914.

[5] Biswas, A. S. T. I. K., Sahu, P. K., Bhowmick, A. N. I. R. B. A. N., & Chandra, M. (2014). Hindi vowel classification using GFCC and formant analysis in sensor mismatch condition. *WSEAS Trans Syst*, *13*, 130–143.

[6] De Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., & Bourlard, H. 2004. Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America*, 116(3), 1781–1792.

[7] Korkmaz, Y., Boyacı, A., Tuncer, T. 2019. Turkish vowel classification based on acoustical and decompositional features optimized by Genetic Algorithm. *Applied Acoustics*, 154, 28–35.

[8] Ghaffarvand-Mokari, P., Mahdinezhad Sardhaei, N. 2020. Predictive power of cepstral coefficients and spectral moments in the classification of Azerbaijani fricatives. *The Journal of the Acoustical Society of America*, 147(3), EL228–EL234.

[9] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341–345.

[10] Breiman, L. 2001. Random forests. *Machine learning*, 45(1), 5–32.

[11] Parks, R. W., Levine, D. S., Long, D. L. 1998. *Computational Neuroscience Fundamentals of Neural Network Modeling: Neuropsychology and Cognitive Neuroscience*. MIT, Cambridge, MA.

[12] Tagliamonte, S. A., Baayen, R. H. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language variation and change*, 24(2), 135–178.

[1] The baseline was measured as the accuracy of a model that makes completely random guesses. The random baseline was the square of the proportion of each vowel in the data.