Interspeech 2018
2-6 September 2018, Hyderabad

Automated Classification of Vowel-Gesture Parameters using External Broadband Excitation

Balamurali B T, Jer-Ming Chen

Audio Research Group, Singapore University of Technology & Design, Singapore
balamurali_bt@sutd.edu.sg, jerming_chen@sutd.edu.sg

Abstract

External broadband signal excitation applied at the speaker (or singer)'s mouth has previously been successfully used to estimate acoustic resonances of the vocal tract during speaking and singing. In this study, we used a modified, low cost, light-weight, pocket-sized and simplified version of this measurement technique, with reduced sampling time and improved low frequency detection, so that such vocal tract measurements may be easily deployed 'in the field' and facilitate a more 'ecological/natural' tracking of phonatory gestures. This system was investigated with 6 volunteer speakers phonating 17 English vowels, and the relative impedance spectrum γ ('gamma') was measured. Although the $\gamma(f)$ signal measured here for each phonatory gesture is somewhat noisier than the original technique, it is still believed to carry some important cues associated with vocal tract configuration that produce these vowels. Features were identified both in the amplitude and phase of $\gamma(f)$ and three ensemble classifiers namely random forest, gradient boosting and adaboost were trained using them. The prediction output from these classifiers were combined using soft voting to predict a class label (front-central-back; open-close). This yielded an accuracy exceeding 80% in classifying the six nominal regions of the vowel plane.

Index Terms: Vocal tract impedance, Machine learning, Relative impedance spectrum, Ensemble classifiers

1. Introduction

In this study, a modified version of the technique previously reported by [1-3] using an external broadband excitation was applied to speaker's lips during phonation of 17 English vowels. This method has previously been used to successfully estimate acoustic resonances of the vocal tract during singing and speaking [4-7].

While the technique was previously used to only estimate vocal tract resonance frequencies, we now utilize the rest of the associated acoustic impedance spectrum data collected during phonation to assist with the automated classification of vowels. Further, the measurement hardware used has been reduced and simplified so that such vocal tract measurements may be easily deployed 'in the field' and facilitate a more 'ecological/natural' tracking of phonatory gestures.

Here we report the first implementation of machine learning techniques to analyze relative impedance spectrum (γ) information associated with speech, towards vowel cluster classification.

The remainder of this paper is organized as follows. Section 2 describes the data collection procedure for this investigation. Experimental methodology is described in Section 3. This includes a brief overview of the features used for the classification process, description about the classifiers and various classifier configurations. Section 4 contains some results produced as part of this investigation. Finally, conclusions of this investigation can be found in Section 5.

2. Data Collection

2.1. Hardware

The hardware offered in our current method is a smaller handheld version of the earlier hardware, with enhanced acoustic coupling and new electronics that allow greater signal:noise response of the detection system at the low frequency limit (~10dB boost @200 Hz) and reduced sampling time to 0.75 seconds (from 3 seconds): these reliably improve our system's tracking ability for the first speech resonance for a range of male and female speakers, allows easy deployment 'in the field' and facilitates a more 'ecological/natural' tracking of phonatory gestures.

In each measurement, we record both the relative impedance (gamma) and relative phase information from 200 Hz to 4000 Hz, which includes the first, second and third speech resonances tracked for each vowel gesture, as well as the audio (.wav) file of the vowel phonation along with the broadband excitation signal introduced at the speaker's lips.

The broadband excitation signal consists of harmonics of frequency 5.383 Hz (44100 Hz/(2^13)), between 200 and 4000 Hz, summed, with optimized phases to improve the signal to noise [8]. The signal is delivered to a portable Nakamichi "mini cube" speaker (5 x 5 x 5 cm). This source of acoustic flow was placed at the speaker's lips, with the lips resting on the speaker grill. Also located on the grill is a small electret microphone (Optimus 33-3013), which records both the sound of the speaker's voice along with the excitation signal interacting with the subject's vocal tract and the radiation field.

An initial calibration is made with the subject's mouth closed: the measurement field is loaded purely by the impedance of the radiation field (Z_{Rad}) as seen at the subject's lips, baffled by the subject's face, and the signal is then adjusted such that the measured pressure signal at the lips is independent of frequency. With the calibration in place, subsequent measurements – during phonation – are then made with the subject vocalizing naturally (i.e., with the lips open). This resulting measurement (relative impedance spectrum, γ) is a ratio of the vocal tract impedance (Z_{Tract}) operating in parallel with Z_{Rad}, with respect to that measured earlier during calibration with the mouth closed (Z_{Rad}), in response to the same acoustic flow. (Because the output impedance of the

10.21437/Interspeech.2018-1756

acoustic flow source is rather larger than both Z_{Tract} and Z_{Rad}, the acoustic flow may be assumed to be constant.)

While maxima in *γ*(*f*) have been shown to identify vocal tract resonances with a possible resolution of +/- 20, in this study, we are using other parameter features in the *γ* spectrum as the input for training a machine learning system, to identify target vowels

2.2. Subjects

6 speakers were recruited, and these speakers had the following characteristics:

- 4 men, 2 women;
- 4 native English speakers (2 Australian English, 2 Singaporean English), 2 non-native English speakers;
- 2 East Asian, 4 Caucasian

2.3. Experimental Protocol

Subjects were asked to phonate words of the form: h-*V*-d, where *V* is the target vowel between the consonants "h" and "d". 13 English vowels + 4 rhotacized/retroflex vowels were explored, and the 17 target words used are as follows

- Heed, hid, head, haired, haiRed
- Who'd, herd, heRd, hud, hard, haRd, had
- Hood, hoe'd hoard, hoaRd, hod

(in this paper, the uppercase "R" indicates a rhotacized/retroflex version of these vowel sound is used). The relative distribution of these words and their corresponding target vowels – based loosely on the vowel plane [9] – is shown in Figure 1. (Note: the exact position of these vowels depends strictly on the English accent used by the speaker.)

While holding the target vowel for about 2-3 seconds, a gamma measurement was made. With 4-5 'takes' of each target vowel, 17 target vowels: ~80 vowel gestures for each subject.

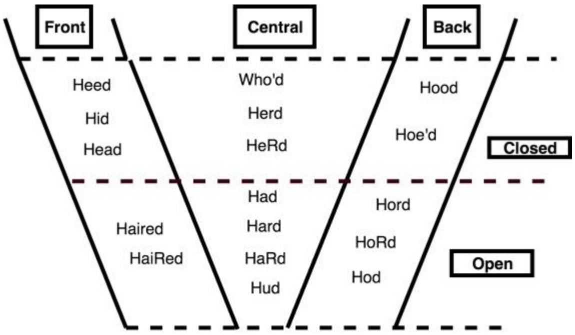


Figure 1: *Distribution of target words indicating[loosely] the relative relationships with target vowels used*

3. Methodology

3.1. Experimental setup

Figure 2 demonstrates the experimental setup. A classifier system using ensemble learning techniques was trained to classify the phonation labels. Ensemble learning perturbs-and-combines a number of machine learning techniques together. Three ensemble learning classifiers namely random forest, gradient boosting and an adaboost classifier were trained and tested as part of this investigation. The chosen classifiers for this investigation contain multitude of decision trees and have been trained using *γ*(*f*) (both amplitude and phase) extracted from phonatory gestures. To predict a class label, these classifiers were combined using soft voting (i.e., weighted averaging the predicted probabilities from the chosen classifiers). Details about *γ* and classifier systems can be found in the following subsections.

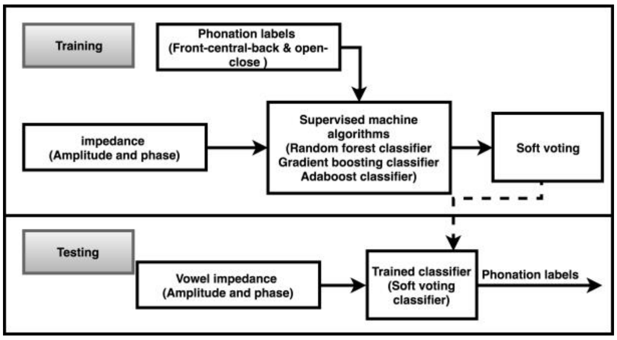

Training
Phonation labels (Front-central-back & open-close)
impedance (Amplitude and phase)
Supervised machine algorithms (Random forest classifier Gradient boosting classifier Adaboost classifier)
Soft voting
Testing
Vowel impedance (Amplitude and phase)
Trained classifier (Soft voting classifier)
Phonation labels

Figure 2: *Experimental Methodology.*

3.2. Training vs Testing Data

The available data (*γ*(*f*) measured during phonation) is split into two sets, a training set and a test set. The classification model is built/trained using training data set and its performance is evaluated using the test set. The test set has never been seen by the model therefore the resulting performance can be considered as good guide to what can be expected when the model is applied to unseen data. Very often, the proportion chosen is 70% for the training set and 30% for the test and we have followed the same in this investigation [10]. The rationale behind this 70-30 split is that more training data makes a better classification model whilst more test data results in accurate error estimate.

3.3. Relative Impedance Spectrum *γ*(*f*) Analysis

As mentioned earlier, amplitude and phase extracted from phonated signal were used as features to train the classifier system. The extracted features were post-processed using Savitzky-Golay FIR smoothing filter [11]. Further, the frequency of interest was limited between 200 and 4000 Hz. Figure 3 shows the resulting relative impedance spectrum *γ*(*f*) (*amplitude* and *phase*) for the target vowel sound [*a*] while phonating the word 'Had'.

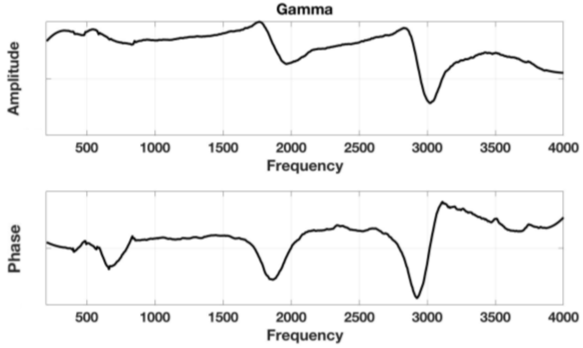


Figure 3: *Relative impedance spectrum γ(f)* (*amplitude and phase) for the target vowel* [*a*] *measured while phonating the word 'Had'.*

It is worth noting here that while two resonances are clearly indicated in the *amplitude* plot by the mid-point of the steep negative slope (~1850 Hz and ~2900 Hz) and accompanied with sharp minima in the phase plot, a rather weaker resonance at lower frequencies may additionally identified in the *phase* plot at ~650 Hz.

3.4. Supervised Classifier Systems

3.4.1. Random forest classifier

Random forest uses multitude of decision trees. Decision trees split the given samples into many homogenous sets based on the significance of input feature values. The top most node in a decision tree contains samples from the entire population and is highly non-homogenous. However, upon splitting to sub-nodes, homogeneity of samples in such sub-nodes increase [12].

In the case of random forest, decision from many trees are considered as opposed to a single decision tree. The final predicted label will be that one which most of the trees voted for. An alternate way of combining decision is by averaging probabilistic prediction from each individual decision tree. In this investigation mode of the class labels predicted by the individual tree was chosen to be the final predicted label [13, 14].

3.4.2. Gradient Boosting classifier

Gradient boosting also uses ensemble of decision trees. In gradient boosting, a simple regression predictor (for e.g. one-layer decision tree) is fitted for the data and then amount of the error per data point in the predictions is computed (i.e., error residual). A model (for e.g. another one-layer decision tree) is then created to predict this error residual. Finally, a new model is created by combining these two predictors (i.e., the original predictor and error residual prediction model). This new model will be more complex and is more accurate than the one-layer decision tree with which the data was initially fitted. This process continues over and over again for the number of classifiers one wants in the ensemble [15, 16].

3.4.3. Adaboost classifier

Adaboost learning starts with fitting a simple classifier (for e.g. a one-layer decision tree) on the data. Such a one-layer decision tree will result in a lot of error data points (i.e., misclassified data points). Weights are now assigned to original data points: higher weights to misclassified data points and lower weights to correctly classified data. The successive classifier will be trained in such a way that it tries to correctly classify those error data points or in other words tries to achieve a low weighted error. The process continues by increasing weights for incorrectly classified points and decreasing those of correctly classified for the next round and continues over the number of classifiers in the ensemble [15, 17].

Number of decision tree estimators used in this investigation for the random forest, gradient boosting and adaboost classifiers are 250, 250 and 100 respectively. Further, we assign equal weights while soft-voting to estimate the final prediction.

4. Results

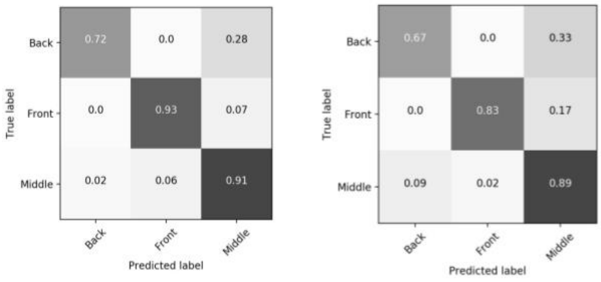
4.1. Front-central-back classification

Table 1 shows the results in terms of classification accuracy for front-central-back classifier system. Accuracy is estimated by comparing predicted labels with the actual labels. It is clear from the result that classifiers trained using gamma amplitude has outperformed those classifiers trained using phase feature. And this trend is consistent irrespective of the chosen classifier type. Now comparing the performance of various classifiers, random forest is marginally better than the other two for both amplitude and phase features.

Table 1: *Classification result of front-central-back classifier*

Classifier type	**Feature accuracy**	
	Amplitude	**Phase**
Random forest	90.4	82.4
Gradient boosting	84.0	78.4
Adaboost	89.6	79.2
Soft voting	86.4	80.8

The accuracy result after soft voting the classifiers is further analyzed by looking at the confusion matrix and this is shown in Figure 4. Confusion matrix can be used to assess the performance of a classifier. For a good classifier, the resulting confusion matrix will look dominantly diagonal (i.e., values closer to one). All the off-diagonal elements in a confusion matrix represent percentage of misclassified data.



(a) (b)

Figure 4: *Confusion matrix showing accuracy of softvoting classifier for front-central-back classification using (a) amplitude (b) phase*

It can be clearly seen from the confusion matrix that percentage of misclassification is higher when try to classify the input between front-middle and back-middle phonation. This is what one should expect. The amount of misclassification is marginally higher when the chosen feature is phase of gamma compared to the amplitude result.

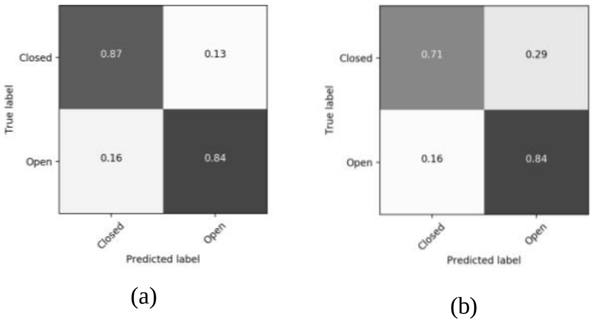
4.2. Closed-open classification

The result of closed-open vowel plane classifier is shown in Table 2. Classification accuracy was found to be marginally better when amplitude of gamma is used to classify the openness of a vowel and this is true irrespective of the classifier type. However, the random forest which was found to have an upper hand while classifying phonation to front-central-back tends not to be the best among the three in this context and its performance falls marginally short behind adaboost and gradient boost. However, with gamma phase as the feature, random forest captures back its top spot.

The accuracy after soft voting the classifiers was analyzed using confusion matrix (see Figure 5). With gamma amplitude as feature, the percentage of misclassification between closed-open and open-closed type is low and they are more or less the same. However, with gamma phase, the closed phonation misclassified as open phonation was found to be slightly higher.

Table 2: *Classification result of closed-open classifier.*

Classifier type	**Feature accuracy**	
	Amplitude	**Phase**
Random forest	81.1	77.9
Gradient boosting	83.2	71.5
Adaboost	83.2	73.7
Soft voting	85.3	77.9



(a)

(b)

Figure 5: *Confusion matrix showing accuracy of softvoting classifier for closed-open classification using (a) amplitude (b) phase*

5. Conclusions

We have shown the implementation of machine learning techniques to successfully classify vowels by analyzing relative impedance spectrum (γ) information collected using impedance measurement hardware which is simpler and more rudimentary than reported previously.

The successful vowel classification using impedance data associated with speech produced, collected under such rudimentary recording conditions on a small handheld device, opens up opportunities for further implementation via mobile computing devices (e.g. smartphones, tablets, etc.), leading to potential work for voice tracking and training.

6. Acknowledgements

Part of this work was funded by the Australian Research Council (ARC) Linkage Project Grant (LP110100570) undertaken at the University of New South Wales with Professors Joe Wolfe and John Smith. We also thank our volunteer subjects.

7. References

[1] Y. Swerdlin, J. Smith, and J. Wolfe, "The effect of whisper and creak vocal mechanisms on vocal tract resonances," *The Journal of the Acoustical Society of America,* vol. 127, no. 4, pp. 2590-2598, 2010.

[2] J. Epps, J. Smith, and J. Wolfe, "A novel instrument to measure acoustic resonances of the vocal tract during phonation," *Measurement Science and Technology,* vol. 8, no. 10, p. 1112, 1997.

[3] A. Dowd, J. Smith, and J. Wolfe, "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time," *Language and Speech,* vol. 41, no. 1, pp. 1-20, 1998.

[4] M. Garnier, N. Henrich, J. Smith, and J. Wolfe, "Vocal tract adjustments in the high soprano range," *The Journal of the Acoustical Society of America,* vol. 127, no. 6, pp. 3771-3780, 2010.

[5] E. Joliveau, J. Smith, and J. Wolfe, "Acoustics: tuning of vocal tract resonance by sopranos," *Nature,* vol. 427, no. 6970, p. 116, 2004.

[6] E. Joliveau, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: The soprano voice," *The Journal of the Acoustical Society of America,* vol. 116, no. 4, pp. 2434-2439, 2004.

[7] T. Donaldson, D. Wang, J. Smith, and J. Wolfe, "Vocal tract resonances: a preliminary study of sex differences for young Australians," *Acoustics Australia,* vol. 31, no. 3, pp. 95-98, 2003.

[8] J. R. Smith, "Phasing of harmonic components to optimize measured signal-to-noise ratios of transfer functions," *Measurement Science and Technology,* vol. 6, no. 9, p. 1343, 1995.

[9] "IPA chart available under a Creative Commons Attribution-Sharealike 3.0 Unported License http://www.internationalphoneticassociation.org/content/ipa-chart,. Copyright © 2015 International Phonetic Association.," ed.

[10] P. S. Crowther and R. J. Cox, "A method for optimal division of data sets for use in neural networks," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 2005, pp. 1-7: Springer.

[11] H. H. Madden, "Comments on the Savitzky-Golay convolution method for least-squares-fit smoothing and differentiation of digital data," *Analytical chemistry,* vol. 50, no. 9, pp. 1383-1386, 1978.

[12] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news,* vol. 2, no. 3, pp. 18-22, 2002.

[13] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[14] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238,* 2013.

[15] A. Ihler. CS178: Machine Learning and Data Mining, Information & Computer Science,UC Irvine. Downloaded from http://sli.ics.uci.edu/Classes/2011W-178 on February 2017 [Online].

[16] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean, "Boosting algorithms as gradient descent," in *Advances in neural information processing systems*, 2000, pp. 512-518.

[17] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for AdaBoost," *Machine learning,* vol. 42, no. 3, pp. 287-320, 2001.