Automatic classification of 'front' and 'back' pronunciation variants of /r/ in the Götaland dialects of Swedish

Johan Frid

SOL, Lund University

Abstract

In the Götaland region of Sweden, there are two major pronunciation variants of the /r/ phoneme: the 'front' and the 'back' variants. We present a classification experiment where we try to classify which of the two variants an unknown speech sound is by using machine learning methods and acoustic features.

Specifically, the methods we use are Classification and Regression Trees, Logistic Model Trees, Multilayer Perceptrons and K-Nearest Neighbour and the features are Formants and Harmonicity (which is a measure of the balance between periodic and aperiodic energy) in bark-filtered speech.

The results show that the single best feature is 'trimmed mean' F2, that Formants+Harmonicity performs better than Formants alone and that the best overall correct classification score is about 89%, which is much better than a baseline method based on choosing the majority class, which gives 52%.

Introduction

The research presented here originated in the work done by Engstrand, Frid and Lindblom (in press) on the perceptual bridge between coronal and dorsal /r/ in Swedish. In that study, we tested the ability to distinguish front and back /r/ perceptually. Even though discrimination between synthetically produced front and back /r/ variants seemed somewhat difficult in that study, identification was comparably easy. This lead us to the question whether this identification also could be performed automatically.

A possible use of this research is in dialect identification. E.g., a dialect aware speech recognizer may have different phonetic models for different speech sounds. Prior knowledge about the dialect, provided by a system like the present one, may then be used to guide the system in choosing the most suitable models.

Material

We used all the words containing the /r/ phoneme in the wordlist section of the Götaland part of the Swedia 2000 database (Engstrand et al. 1997; Eriksson 2004). For the purposes of this study, we assumed that only the Götaland material would be relevant for the front/back distinction. The words were *dagar* 'days', *dör* 'dies', *dörr* 'door', *fara* 'danger', *rasa* 'collapse' and *särk* 'chemise'. The /r/ parts of each word had been segmented and labelled previously as part of the Swedia project. The author was not directly involved in this analysis process.

The material consisted of words from all the 37 locations in the Götaland part and included material spoken by both men and women. In total, the material consisted of 1995 words, of which 52% were 'front' pronunciations, and 48% were 'back' pronunciations. A simple baseline for this data set based on choosing the class with the majority of cases in it, would thus give 52% correctly classified cases.

Attributes

We used two sets of attributes: one based on conventional resonance frequencies in the vocal tract: *formants*, and one based on the balance between periodic and aperiodic acoustic energy in bark-filtered parts of the speech signal:

harmonicity. The idea behind this is that aperiodicity may occur in different parts of the spectrum for the different /r/ variants due to differences in articulation. Formant values were extracted using the Burg method (Press et al. 1992) in the Praat program (Boersma & Weenink 2007), and Harmonicity also by the Praat program.

Formants

Since automatic estimation of formant values sometimes may result in erroneous values (especially when formants are close to each other and one of them possibly is missed), we adopted the method of trimmed means to increase robustness. In effect, for each /r/ segment, we measured one set of formant values (full array of F1-F5) for every 10 ms and then calculated a trimmed mean (using the 20%-80% interval) per formant.

Another attempt at increasing robustness is to use 'formant tracking', an algorithm that tries to extract 'formant paths' by trying to keep each formant as close as possible to a specified value. This method, however, is reported to work best for vowels (Boersma & Weenink 2007) and we did not find that this method improved the results in the end.

Harmonicity

Harmonicity measures the balance between periodic and aperiodic energy in the acoustic signal. The following description is from the Praat manual (Boersma & Weenink 2007):

> A Harmonicity object represents the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio (HNR). Harmonicity is expressed in dB: if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is 10*log10(99/1) = 20 dB. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise.

Harmonicity was not measured over the whole frequency range but rather in the output of a bank of bark-scaled filters. In this way, aperiodicity in different parts of the spectrum is detected. The filters are shown in Figure 1. All in all, there were 21 bark filters, and therefore 21 different harmonicity measurement values.

Feature selection

Feature selection is a technique of selecting a subset of relevant features for building robust learning models.

In classification experiments like this one, features may be more or less useful, and to some extent even harmful, in predicting the correct class. In the case of the formants, it is not unusual

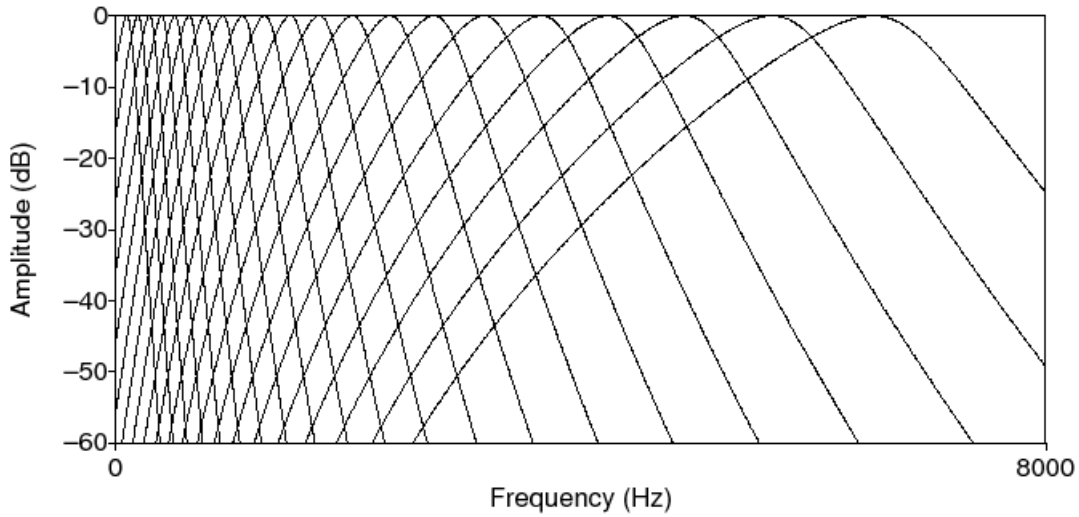

Figure 1. The bark-scaled auditive filters that were applied to the speech signal before calculation of harmonicity.

that the lower formants are better at distinguishing between speech sounds. This is partly because the lower formants vary more, partly because they often can be more reliably estimated than the higher ones.

In the case of the harmonicities, it is also expected that some of them will be better at distinguishing between the two speech sound classes than others, since aperiodicity simply may be missing or very weak in some spectral regions for both speech sounds. Furthermore, since the bark filters overlap, there may be correlations between the different harmonicity measurements since they, to some extent, cover overlapping parts of the spectrum.

We used a method in the WEKA toolbox (Witten & Frank 2005) that evaluates subsets of features by testing the individual predictive ability of each feature as well as the degree of redundancy between them. This method suggested that seven features should be used in the end: the trimmed F1, F2 and F3 values, and the harmonicity levels of bands 9, 12, 14 and 19 (these correspond to the bands with central frequencies at 1085, 1746, 2357 and 4883 Hz).

Classification

We used the WEKA toolbox for all training and testing, and the following methods (described in Witten & Frank 2005) were used:

1. **RepTree** (Reduced Error Pruning). A variant of CART (decision trees) where the maximum depth of the tree can be set.
2. **LMT** (Logistic Model Tree). This is also similar to CART, but with a logistic regression function at each node.
3. **MLP** (Multilayer Perceptron), or neural network. We used 4 hidden nodes and 500 training iterations.
4. **IBk** (k-nearest neighbour). This method classifies objects based on the closest training examples in the feature space. In this study, we used 33 training examples (this was determined by local optimization).

Training

All training and testing was done using 10-fold cross-validation, where the data is partitioned into 10 subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. Furthermore, this procedure was repeated 10 times, so in total each algorithm was trained and tested on 100 different data partitions. In order to test if Harmonicity improves the results, we did one full (100 partitions) training+test run using formants only and one using the combination of Formants + Harmonicity.

Testing

RepTree

As stated above, RepTree lets you set the depth of the tree. This is interesting since this lets you request a number of rules and then the method selects the best features for this number of rules. Thus you can see which of the features that are the most important. Table 1 shows how the algorithm performs for one, two and three rules (on the Formants + Harmonicity feature set).

Table 1. The performance of RepTree for different depth levels. The features and the values at which they split the data is also shown. R is back /r/, and r is front /r/.

1 level: **65.8**% tr_mean_f2 < 1215.5 : R tr_mean_f2 >= 1215.5 : r
2 levels: **74**% tr_mean_f2 < 1215.5 | tr_mean_f3 < 2384 : r | tr_mean_f3 >= 2384 : R tr_mean_f2 >= 1215.5 | tr_mean_f1 < 741 : r | tr_mean_f1 >= 741 : R
3 levels: **77**% tr_mean_f2 < 1215.5 | tr_mean_f3 < 2384 | | tr_mean_f2 < 939.5 : R | | tr_mean_f2 >= 939.5 : r | tr_mean_f3 >= 2384 : R tr_mean_f2 >= 1215.5 | tr_mean_f1 < 741 : r | tr_mean_f1 >= 741 | | mharm_bark19 < 4.04 : r | | mharm_bark19 >= 4.04 : R

From this, we see that F2 is the single most important feature. A single rule saying that the sound is a back /r/ if F2 is lower than 1215 Hz, and otherwise a front /r/ is correct for almost 66% of the instances. By adding rules for other formants and harmonicity measurements the figure is improved.

All methods

The results for all the methods and feature sets are presented in Table 2.

Table 2. Percent correct classifications. Mean of 100 different test runs on different partitions. Formants only were not tested with MLP.

	RepTree	**IBk**	**LMT**	**MLP**
Formants	83.8	86.8	85.5	-
Formants+ Harmonicity	85.6	89.3	88.7	89.1

In the table, we see that the IBk method has the highest mean percent correct for the F+H feature set. However, this is not a statistically significant difference from the results for LMT and MLP. Furthermore, we see that the scores for F+H always is higher than F only. This difference is statistically significant for all methods (MLP takes very long time, so we did not test it on Formants only, but we would expect the same results). The t-tests were performed in WEKA and they only tell you if a difference is significant or not and does not give you p-values. We also checked Harmonicity alone, but they perform much worse than the Formants.

Discussion

Even though the RepTree method produces the worst results it also gives you the most readable and interpretable results. We would like to point out one thing here and that is that the result that an F2 threshold at 1215 Hz is the single most useful rule corresponds to the findings in the identification test performed by Engstrand, Frid and Lindblom (in press). Here, listeners judged synthetic /r/ stimuli where F2 and F3 were varied, and the judgements shift from a majority of 'front' judgements to a majority of 'back' judgements when crossing the F2 threshold.

Conclusions

In this paper we have shown that front and back /r/ speech sounds in the Götaland region of Sweden is classified correctly in 89.3% of the instances. We have shown that the single best rule for guessing the place of articulation is to check if F2 is above or below 1215 Hz. Furthermore, we have shown that Harmonicity, a measure of the balance between periodic and aperiodic energy in different parts of the spectrum, is helpful in classifying these sounds.

References

Boersma P & Weenink D (2007). *Praat: doing phonetics by computer* (Version 4.5.18) [Computer program]. Retrieved March 30, 2007, from http://www.praat.org/

Engstrand O, Frid J & Lindblom B (in press). A perceptual bridge between coronal and dorsal /r/. In Solé M-J, Beddor P S & Ohala M (eds.), *Experimental Approaches to Phonology. In honor of John J. Ohala*, Oxford: Oxford University Press.

Eriksson A (2004). *SweDia 2000: A Swedish Dialect Database.* [WWW document]. URL gandalf.aksis.uib.no/~gjert/ESFWorkshop/Eriksson.pdf

Press W H, Teukolsky S A, Vetterling W T & Flannery B P (1992). *Numerical Recipes in C: the art of scientific computing*, Second Edition, Cambridge University Press.

Engstrand O, Bannert R, Bruce G, Elert C-C, & Eriksson A (1997). Phonetics and phonology of Swedish dialects around the year 2000: a research plan. *PHONUM* 4, 97–100.

Witten Ian H and Frank E (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.