COMPARING HUMAN AND MACHINE VOWEL CLASSIFICATION

Uwe D. Reichel, Katalin Mády

Institute of Phonetics and Speech Processing,
University of Munich, Schellingstr. 3, 80799 Munich, Germany
{reichelu, mady}@phonetik.uni-muenchen.de

ABSTRACT

In this study we compare human ability to identify vowels with a machine learning approach. A perception experiment for 14 Hungarian vowels in isolation and embedded in a carrier word was accomplished, and a C4.5 decision tree was trained on the same material. A comparison between the identification results of the subjects and the classifier showed that in three of four conditions (isolated vowel quantity and identity, embedded vowel identity) the performance of the classifier was superior and in one condition (embedded vowel quantity) equal to the subjects' performance. This outcome can be explained by perceptual limits of the subjects and by stimulus properties. The classifier's performance was significantly weakened by replacing the continuous spectral information by binary 3-Bark thresholds as proposed in phonetic literature [8]. Parts of the resulting decision trees can be interpreted phonetically, which could qualify this classifier as a tool for phonetic research.

Keywords: vowels, Hungarian, perception, classification, decision trees

1. INTRODUCTION

The Hungarian vowel system comprises 14 or 7 phonemes depending on whether vowel quantity is considered as a distinctive feature or not: /i y e ø a o u/(+ :). Like in other languages among the relevant acoustic cues for their identification are formant frequency [4], duration [4] and intrinsic pitch [3]. How these and other features influence vowel identification is examined in [5]. In this study we would like to investigate how a machine learning approach, namely the C4.5 decision tree [7], performs given this feature pool in comparison with human subjects and how it utilises these features.

2. PERCEPTION EXPERIMENT

The reference of human subject vowel identification originates from an identification experiment, in which 280 vowel stimuli consisting of the 14 vowel phonemes in equal amount presented 10 times each in voiceless and voiced context were presented to 37 subjects (10 male, 27 female, age from 14 to 44 years). The stimuli were derived from an electromagnetic articulography (EMA) examination of a male Hungarian speaker (age: 21) at the ZAS, University of Berlin. The presentation of the stimuli was carried out in two passes: 1) in an embedded condition and 2) in an isolated condition. In condition 1 the stimuli V were presented within a carrier word /aCVCa/, where C_C is a consonantal bracket of either both voiced or voiceless velar consonants. In condition 2 the stimuli consisted of the central 40 ms of the vowels, weighted by a Tukey window (taper sections set to 5 ms each) to remove cracks at the segment boundaries. Each stimulus could be listened to at most twice. The subjects were instructed in advance that the same number of short and long vowels is presented and that condition 2 is based on the vowels known from condition 1.

3. AUTOMATIC CLASSIFICATION

3.1. Classifier

As a classifier the C4.5 decision tree [7] was chosen. In this tree each object (vowel stimulus) is represented as a path from the root to the leaf connected to the object's class. Each non-terminal node is associated with an attribute according to which the set of objects is further divided. During the recursive branching of the tree in the training procedure the choice of the attribute A most suited to split the set of objects is guided by a measure closely related to information gain which gives the mean reduction of the amount of bits needed to encode the object class given that the value of A is known. The branching stops if all objects belong to the same class or cannot be further distinguished by their features. Subsequent recursive pruning against over-adaption is guided by the comparison of the expected error rate of a subtree and the error rate expected after its condension to a leaf.

There are two main reasons for the choice of this classifier: first, due to its pruning mechanism operating on the training material, no development set is required to avoid over-adaption. This is highly desirable to face the data sparseness problem in our study. Second, the outcome of the training process is

quite transparent, i.e. decision trees can in principal be interpreted to a certain extent.

3.2. Data, Dependent Variables, and Features

Data The data for training and testing consists of the 280 vowel stimuli of the perception experiment.

Dependent Variables The stimulus classes to be predicted are vowel quantity and identity, both for isolated and embedded stimuli. As a reference we took the classes intended by the speaker and not the listener judgements since the judgements were difficult to interpret due to only moderate inter-subject agreement (see Table 3). Furthermore taking just the reliable proportions of judgements as training and test material would have seriously worsened the data sparseness problem.

Features For the embedded task spectral and temporal features were extracted (cf. Table 1). Temporal features comprised vowel duration and duration of the surrounding consonants C_1 and C_2 normalised by the overall duration of the CVC sequence. This normalisation was carried out to prevent the trees from over-adaption to consonant length. Spectral features comprised fundamental frequency F0, distances between the first three formants expressed in Bark and the presence or absence of voicing VOI(C) in the consonant bracket. The choice of formant distances instead of absolute formant values is guided by the findings of Syrdal and Gopal [8] indicating that relative frequency measures are more closely connected to vowel identification than the absolute ones. Additionally, identification according to the binary feature matrix of formant distances less or greater than three Bark as proposed in [8] was trained. For this reason the respective spectral features were binarized. We omitted the F4–F3 distance since according to the binary feature matrix it did not play a role for classification of the vowels in question here.

For the isolated task only those features were used which were available to the human subjects. These were F0 and formant distances.

Temporal features were derived from manual segmentation of the data. To get spectral feature values *Praat 4.5.01* software was utilised. F0 was extracted by auto-correlation, for formants the Burg-LPC was used. As feature values medians were derived from the measurements within a 20 ms window, which was moved in 5 ms steps within the center of the vowel (40 ms).

4. RESULTS

Due to data sparseness 30-fold cross validation was applied in order to get testable mean accuracies. At each step the training set comprised 80% of the data and the test set the remainder.

Table 1: Features used for vowel classification.

temporal	DUR(V), DUR(C_1)/DUR(C_1VC_2), DUR(C_2)/DUR(C_1VC_2)
spectral	F0, F1–F0, F2–F1, F3–F2 (*Bark-scale*), VOI(C)

The performances of the automatic classifier and human subjects were compared with respect to vowel quantity and identity which led to four scenarios: (isolated|embedded) x (quantity|identity).

Table 2: Mean classification results for C4.5 trees (based on two different feature pools) and subjects in %. ** indicates that the difference between mean C4.5 and subject performance is highly significant (two-sided t-test, $\alpha = 0.001$).

		C4.5	**C4.5 (3Bark)**	**Subjects**
Isolated	Quantity	77.42**	61.70	60.31
	Identity	73.02**	35.02	53.31
Embedded	Quantity	95.90**	95.14	92.33
	Identity	83.16	54.01	85.21

Table 2 shows, that in three of four cases the automatic classifier outperforms the human subjects (two-sided t-test, $\alpha = 0.001$; it was appropriate to apply a parametric test since the compared samples are normally distributed according to the Lillie test). The difference is over all apparent for the isolated vowel presentation. Furthermore, it can be seen that classification by binary distance features relating on 3-Bark thresholds is apparently worse than by using more detailed spectral information.

Feature normalization to the intervall $[0, 1]$ did not lead to any significant improvement, so the results are not shown here. This finding may be due to the homogeneous training material given just one speaker and one stable carrier sentence.

All observed performances were better than chance guesses (χ^2-test, $\alpha = 0.001$).

Table 3: Inter-subject agreement (**ISA**) and proportion of cases in which the most frequent judgment was correct (**MFJC**) (in %).

		ISA	**MFJC**
Isolated	Quantity	82.14	58.99
	Identity	72.16	58.63
Embedded	Quantity	92.75	97.48
	Identity	85.63	96.76

As can be seen by the results in Table 2 as well as by the inter-subject agreements in Table 3, subjects performed worse for isolated vowels than for embedded ones.

Figure 1: Decison tree for quantity of embedded vowels. Information in brackets: number of stimuli counted at the respective leaf and percentage of correctly classified stimuli.

DUR ≤ 0.10:
- **VOI** $= 1$: ***short*** (48; 100%)
- **VOI** $= 0$:
  - **DUR** ≤ 0.08: ***short*** (49; 100%)
  - **DUR** > 0.08:
    - **F1–F0** ≤ 1.88: ***long*** (9; 92%)
    - **F1–F0** > 1.88: ***short*** (3; 100%)

DUR > 0.10:
- **DUR** > 0.12: ***long*** (74; 100%)
- **DUR** ≤ 0.12:
  - **VOI** $= 0$: ***long*** (26; 100%)
  - **VOI** $= 1$:
    - **LC2** ≤ 0.27: ***short*** (9; 100%)
    - **LC2** > 0.27: ***long*** (4; 75%)

Decision Tree Example Figure 1 shows a decision tree for embedded vowel quantity, the root on the left, the leaves on the right. To give an explanatory example, the first path through the tree has to be read the following way: if the vowel's duration is below 100 milliseconds and the surrounding consonants are voiced, then it is a short vowel. For 48 stimuli in the training data this combination of feature values was sufficient for correct classification (100% correctly classified).

5. DISCUSSION

5.1. Task Dependent Performances

Embedded vs. Isolated Vowels As can be seen in Tables 2 and 3, for both the subjects and the automatic classifier identification of isolated vowels is more difficult than identification of the embedded ones. This finding is reflected in lower performances and concerning the subjects also in lower inter-subject agreement as well as in a lower degree of correspondence between the most frequent judgements and the reference in the isolated vowel conditions.

Since temporal features were only available for the embedded vowels and not for the isolated ones it can be concluded that they are needed by man and machine to enhance vowel identification capability.

Vowel Quantity vs. Identity Subjects and machine performed better in determining quantity than vowel identity both in the isolated and the embedded vowel condition. This finding can be explained by the different complexities of these two tasks: determining quantity is simply the choice between two alternatives and therefore easier than choosing one of 14 alternatives when determining vowel identity. In accordance with this explanation κ-statistics considering the accuracy-by-chance baselines reveal higher 'identity' performances compared to 'quantity' performances in the isolated vowel condition for both classifier and subjects.

5.2. Machine vs. Human Classifier

The claim that the automatic classifier performed better than the subjects in three of four cases, is in our opinion justified by the significant accuracy differences and the rather moderate inter-subject agreements, the latter speaking against the possibility that human identification performance would not be worse but just different.

The most apparent performance differences could be observed in the isolated vowel conditions. Possible reasons are discussed in the following:

Perceptual Limits One of the constraints of the human perception system is its minimum need for about 6 periods for stable pitch determination [2]. Due to this inertia a stimulus duration of 40 ms as given in the isolated vowel conditions requires a maximum F0 period of about 6.5 ms, which corresponds to a minimum of 150 Hz for correct F0 determination. The pitch feature is therefore not reliable for human listeners in the isolated vowel conditions.

Erroneously Assumed Features In contrast to automatic classifiers subjects have no access to a feature pool definition for vowel identification. This causes problems when signal properties that are just connected to the mode of presentation are misunderstood as vowel features. The frequent erroneous classification of originally long vowels as short ones in the isolated condition (see [5] for further details) could be ascribed to the possibility that subjects considered the presentation time of 40 ms as a vowel duration feature, despite the preceding instruction. Automatic classifiers are not vulnerable to such misinterpretations.

Naturalness of the Stimuli Since the stimuli were derived from an EMA session they are not completely natural and therefore may not have exactly matched the learned vowel patterns of the subjects. The classifiers in contrast have been trained only on the EMA stimuli, and therefore their performance was not disturbed by any potential mismatch in reference to completely natural vowels.

5.3. 3-Bark Criterion

As shown in Table 2 the 3-Bark Criterion of [8] is not suitable for automatic vowel classification. The classifier needs more detailed spectral information.

5.4. Phonetic Interpretation of Decision Trees

Since the trees predict the intended vowel classes, their interpretation refers to intended production and not to perception. Most of the observations refer to the tree shown in Figure 1, but they also apply for parts of the remaining three trees.

Vowel Duration The tree in Figure 1 shows a strong tendency to classify vowels with shorter duration (DUR $\leq$ 100 ms) as short vowels, and else as long vowels (92% each).

Diffuse vs. Compact Vowels F2–F1 is the top-level feature of the decision trees for vowel identity. Lower values correspond to compact back vowels, higher values to diffuse front vowels. In correspondence the trees show the clear tendency to divide the stimuli according to this feature into these two classes (however, the chosen split point is too high):

- embedded condition:
  $F2 - F1 > 7.52$ Bark: 100% front vowels,
  else: 75.4% back vowels
- isolated condition:
  $F2 - F1 > 7.52$ Bark: 100% front vowels,
  else: 76.6% back vowels

Intrinsic Vowel Features For a variety of languages it has been observed that high vowels have a shorter intrinsic duration than low vowels (see [6] for English and [3] for Hungarian). This finding is reflected in the combination of the DUR and the F1–F0 feature in the tree in Figure 1. Since vowel height is inversely related to the distance between F1 and F0, the following interpretation is possible:

> Given no further durational distinction ($0.1s \geq$ DUR $> 0.08s$):
> - $F1 - F0 \leq 1.88$ Bark, i.e. higher vowels are classified as *long*
> - $F1 - F0 > 1.88$ Bark, i.e. lower vowels are classified as *short*

The given DUR range does not reach intrinsic duration values required for lower vowels.

Consonantal Context As has been observed for languages like German [1] and Hungarian [4] there is a tendency towards longer vowel duration before voiced consonants. In accordance with these findings the feature VOI has an impact on dividing long and short vowels in the tree in Figure 1.

- Vowels with short duration (DUR $\leq 0.1s$) are always classified as *short* in the voiced case, since duration is further reduced by attributing parts of it to consonant voicing. Vowels with the same duration classified as *long* appear only in the unvoiced case.
- Vowels with moderate duration ($0.1s <$ DUR $\leq 0.12s$) are always classified as *long* in the unvoiced case, since duration could not be attributed to consonant voicing.

Tense vs. Lax Vowels In the embedded condition /e:/ and /ɪ/ are separated at two tree nodes not by spectral properties but by duration, which corresponds well to the findings in [5], which show a high degree of spectral overlap between these two classes.

> Given no further spectral distinction:
> - shorter duration: /ɪ/ (18 cases, 100% correct)
> - longer duration: /e:/ (17 cases, 100% correct)

This behaviour is conform with the finding that lax vowels like /ɪ/ are shorter than tense vowels like /e:/ (at least in stressed position as given here) [4].

6. CONCLUSIONS

As was shown, machine learning methods are able to outperform human subjects in vowel identification tasks even when being trained on sparse data. In future studies it is to be tested how machine classification performs facing data with more variation (e.g. due to varying place of stop articulation) and if an enlargement of the feature pool e.g. by adding formant transition slope could lead to an improvement of performance.

Transparent C4.5 trees in particular can partly be interpreted phonetically. It would be interesting to examine whether also new phonetic knowledge could be derived this way.

Given more data and more uniform subject judgements decision trees could also be trained on subject responses instead of the intended vowel class to examine the perceptual aspects of vowel identification.

7. ACKNOWLEDGEMENTS

We would like to thank Christian Geng, ZAS, Berlin, for providing us with the Hungarian data.

8. REFERENCES

[1] Braunschweiler, N. 1997. Integrated cues of voicing and vowel length in German: A production study. *Language and Speech* 40(4), 353–376.

[2] Doughty, J., Garner, W. 1948. Pitch characteristics of short tones. II. Pitch as a function of tonal duration. *J. Experimental Psychology* 38, 478–494.

[3] Gósy, M. 2002. Magánhangzók változása az idő függvényében. In: Hunyadi, , (ed), *Kísérleti fonetika – laboratóriumi fonológia*. Debrecen: Debreceni Egyetem Kossuth Egyetemi Kiadója 7–20.

[4] Kovács, M. 2002. *Tendenciák és szabályszerűségek a magánhangzó-időtartamok produkciójában és percepciójában*. PhD thesis Debreceni Egyetem.

[5] Mády, K., Reichel, U. D. 2007. Quantity distinction in the Hungarian vowel system – just theory or also reality? *Proc. ICPhS* Saarbrücken.

[6] Peterson, G. E., Lehiste, I. 1960. Duration of syllable nuclei in English. *J. of the Acoustical Society of America* 32(3), 693–703.

[7] Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

[8] Syrdal, A. K., Gopal, H. S. 1986. A perceptual model of vowel recognition based on the auditory representation of American vowels. *J. of the Acoustical Society of America* 79(4), 1086–1100.