PERCEPTUAL AND STATISTICAL CLASSIFICATION OF VOWELS

Rolf Lindgren & Susanne Eismann, Dept of Phonetics, Stockholm University

BACKGROUND

Investigating the claim that there may exist invariant correlates in the signal for the phonemes, we have explored the vowel system in spontaneous speech, where the vowels show a large variation and overlap. The question asked is how much of the phoneme identity is hidden in the acoustics of the signal and what analysis method is optimal to describe this?

Our methodology to investigate this is to combine different acoustic analysis techniques with different classification methods and correlate this to presumed vowel identity. Eventually we may find a parameter, or combination of them, that has a strong descriptive power and is invariant. We have started in a traditional way with formant analysis. We will after complete evaluation continue with other analysis methods (FFT, cepstrum coefficients, LPC, etc).

SPEECH MATERIAL

Reference speech: REF – 308 specially designed CVC–words, read in a clear manner. Spontaneous speech: SPO–lg – 371 lexical and grammatical words; SPO–l – 234 lexical words; SPO–g – 198 grammatical words; SPO–tst – 61 vowel data from perception test (see Vowel Perception Test below). The separation into grammatical and lexical words was done for explorative purposes, in order to see if there was any systematic differences between vowels from these categories. Each data set, except data from Vowel Perception Test, was divided into two equally large training and testing parts. When training with one data type but testing against another the complete data sets were used. The formant values were extracted from the phonologically stressed syllable in the words and only long vowels were used. The input data was the mean of F1–F4, averaged over 3 consecutive analysis frames.

ARTIFICIAL NEURAL NETWORK CLASSIFIERS

In a serie of experiments two types of artificial neural networks (ANN) paradigms were used to classify formant patterns of vowels from spontaneous speech; feed–forward back–propagation, BP and Kohonen Feature Map, KFM (Eismann & Lindgren 1992a).

The calibrating data for ANN should be unambiguous and representative. Thus, we have trained the ANN's also with the speakers clear speech, i.e. every speakers spontaneous speech is tested against his own, optimal speech.

Training procedure

The BP had 4 input nodes, 6 hidden nodes and 9 output nodes. The output was a 9–bit pattern, where each bit represented a vowel type. During training each data set was iterated 500000 times or until 95% of the training data was correctly classified. The KFM had 4 input nodes and 64 output nodes. Input data were the same as for BP. Each data set was iterated 11000 times in two passes (1000 + 10000). After training the nets were used to classify data from other data sets. Each test vector were classified as the vowel phoneme corresponding to the output node with the largest activity. If one was to further qualify this principle that the winning node has to be above a certain threshold, e.g. 0.75, then the performance is substantially degraded, except for the case when the BP is both trained and tested with reference speech

RESULTS

Generally, the BP nets have been shown to have a large classification capacity, usually in the range of 90–95%. This also holds for reference speech data (97.4% correct). However, it performed poorly on vowel data from spontaneous speech (see Table 1).

If we reverse the process, training with spontaneous speech, the classification improves somewhat, but the overall performance is still rather bad. If we sharpen the demands for correct classification (output node > 0.75) the performance degrades to 15.6%. The overall low classification capability indicates that the variation is so large that it becomes ambiguous. Even after 500000 iterations the "Spont BP's" were not capable of producing an unambiguous output. Somewhat surprising is that "spont BP" performs so badly on reference speech data.

TABLE 1. *Results of experiment cross–comparing BP nets trained and tested with different types of data sets. Each cell shows the percent correct classification.*

Training	Testing data from:				
data from	**REF**	**SPO–lg**	**SPO–g**	**SPO–l**	**SPO–tst**
REF	**97.4**	36.7	25.8	46.6	37.7
SPO–lg	42.9	**57.4**	62.1	55.1	39.3
SPO–g	39.6	53.9	**54.0**	39.7	31.1
SPO–l	51.3	57.7	38.4	**52.1**	54.1

Vowel confusion matrices

In order to check for any systematic misclassifications a confusion matrix was prepared for every single test (20 tests for the BP). Table 2 shows the confusion matrix for a net trained on reference speech and tested on spontaneous speech.

TABLE 2. *Confusion matrix for BP trained on reference speech and tested with spontaneous speech. The data were manually labelled as indicated by rows and classified by the BP as indicated by columns. Each cell show the hit percentage.*

	a:	**e:**	**i:**	**u:**	**ʉ:**	**y:**	**o:**	**ɛ:**	**ø:**
a:	**57.3**	1.3	0.0	5.3	2.7	1.3	13.3	1.3	17.3
e:	5.1	**30.3**	2.0	7.1	13.1	11.1	2.0	17.2	12.1
i:	0.0	30.6	**42.9**	0.0	10.2	10.2	0.0	4.1	2.0
u:	0.0	0.0	0.0	**75.0**	0.0	0.0	25.0	0.0	0.0
ʉ:	0.0	25.0	3.1	6.3	**59.4**	3.1	3.1	0.0	0.0
y:	0.0	0.0	50.0	0.0	0.0	**50.0**	0.0	0.0	0.0
o:	30.3	1.3	0.0	36.8	7.9	1.3	**10.5**	0.0	11.8
ɛ:	38.1	4.8	4.8	0.0	0.0	4.8	14.3	**28.6**	4.8
ø:	15.4	0.0	0.0	23.1	7.7	7.7	7.7	0.0	**38.5**

Winner–takes–all hit score: 36.7 %, above 0.75 activation hit score: 25.3 %

The same pattern as with BP can be seen with KFM (Table 3). KFM trained on reference speech yields poor results on data from spontaneous speech. In fact, in this case there is a substantially large proportion of the data from spontaneous speech that is non–classified, i.e. assigned to the non–labelled cells in the map. KFM trained on data from spontaneous speech behaves approximately as BP, although the overall error rate is slightly larger. The lower number of non–classified is due to the fact that there are fewer non–labelled cells in this map because of large variation in the training data.

TABLE 3. *Results of experiment with cross–comparing KFM trained and tested with different types of data sets. In each cell is shown the percent correct classification (1st), the percent false classification (2nd) and percent non–classified vowels (3rd).*

Training data	Testing data REF	SPO–lg	SPO–g	SPO–l	SPO–tst
REF	**90–7–3**	15–12–73	11–12–77	19–11–70	13–14–73
SPO–lg	42–58–0	**40–58–2**	41–57–2	52–48–0	38–62–0
SPO–g	30–45–25	45–47–8	**37–41–22**	31–63–6	27–65–8
SPO–l	45–54–1	49–47–4	32–64–4	**54–42–4**	49–43–8

VOWEL PERCEPTION TEST

To compare the ANN's to human listeners, a vowel identification test was performed. The test consisted of two parts: I) using natural vowels, excised from spontaneous speech and II) using synthetic vowels, modelled after the vowels in I.

Perception test with natural vowels

The vowel stimuli were tokens from the spontaneous speech of one speaker. Using an euclidean distance metric in a Bark calibrated F1xF2 space and assuming normal distribution of the vowel data points, several tokens for each vowel category were selected. Using a Principal Component Analysis on the data set from each vowel category, 9 reference points were defined: the intersection between the 1st and 2nd components axis (= center of gravity for the vowel class) and the eight intersections between 1 & 2 st.dev ellipsoid and the 1st and 2nd components. The actual tokens used in the perception test were those lying closest to these reference points, thus getting a representative sample for the total variation in each vowel class. A total of 61 tokens were obtained (u: yielded only 7, y: and ɛ: were left out because of too few data points).

Perception test with synthetic vowels

The vowel stimuli were synthetic versions of the natural vowels used in the first part. The vowels were synthesized with the SUSYNT speech synthesizer (Eismann & Lindgren, 1992b) using the mean value of F1–F4 from the natural vowels middle portion. All synthetic vowels had the same F0 contour and duration (350 ms). The subjects and the procedure were the same as in the first part.

Test procedure

Each vowel token repeated three times and were presented in a randomized order. The stimuli were presented binaurally via headphones to a group of 26 speakers of Swedish. The subjects were requested to perform a 9–alternative forced–choice labelling test. The labelling alternatives included all the long vowels in Swedish, although two of them (y: and ɛ:) were not present in the test material. The subjects answered by circling the label they thought best described each of the presented stimuli.

TABLE 4. *Results of perception experiment with vowel stimuli excised from natural speech and synthetic vowels. The data were manually labelled as indicated by rows and classified by subjects as indicated by columns. Each cell shows the percentage of vowel assignment made by listeners.*

Natural stimuli

	i:	y:	ʉ:	ø:	e:	ɛ:	a:	o:	u:
i:	**45**	9	10	2	33	1	0	0	0
y:	(NOT USED AS STIMULI)								
ʉ:	3	4	**69**	17	5	1	0	0	1
ø:	3	4	17	**41**	3	6	15	9	2
e:	6	3	17	12	**45**	15	0	1	1
ɛ:	(NOT USED AS STIMULI)								
a:	0	1	6	13	3	8	**43**	21	5
o:	1	1	18	9	2	1	12	**39**	18
u:	0	0	4	1	0	0	0	5	**90**

Synthetic stimuli

	i:	y:	ʉ:	ø:	e:	ɛ:	a:	o:	u:
i:	**28**	13	15	12	26	6	0	0	0
y:	(NOT USED AS STIMULI)								
ʉ:	2	6	**40**	38	2	11	0	1	0
ø:	1	1	20	**53**	0	11	3	9	2
e:	8	3	14	29	**20**	26	0	0	0
ɛ:	(NOT USED AS STIMULI)								
a:	0	0	4	27	0	12	**42**	13	2
o:	0	0	17	27	0	1	13	**32**	10
u:	0	0	11	1	0	0	7	45	**36**

Results

The test showed that 52.1% of the natural vowels were perceived in agreement with the labelled vowel quality and in the synthetic version 35.9%. A comparison between the dispersion of of the responses for natural and synthetic stimuli reveals that listeners tend to judge the the synthetic vowels according to their location on the vowel space, whereas natural vowels to a greater extent carry their intended phonetic quality as they move in space.

The "spo–ANN" classified 37.7% of the vowels according to labelled vowel quality, but it behaves slightly different than listeners on synthetic stimuli, especially for e: and ø:.

	i:	y:	ʉ:	ø:	e:	ɛ:	a:	o:	u:
i:	**44**	11	0	0	33	11	0	0	0
y:	(NOT USED AS STIMULI)								
ʉ:	0	22	**33**	0	22	11	0	0	11
ø:	0	11	11	**11**	0	11	22	11	22
e:	11	0	0	0	**33**	22	11	11	11
ɛ:	(NOT USED AS STIMULI)								
a:	0	0	0	11	0	11	**67**	0	11
o:	0	0	0	11	0	0	11	**33**	44
u:	0	0	0	0	0	0	0	57	**43**

TABLE 4. *Confusion matrix for BP trained on spontaneous speech and tested with the stimuli used in the vowel perception test. The data were manually labelled as indicated by rows and classified by the BP as indicated by columns. Each cell show the hit percentage.*

REFERENCES

Eismann, S & Lindgren, R, 1992a. *Manual for BPSU and KFM artificial neural networks. Stockholm Universit, Dept of Phonetics.*

Eismann, S & Lindgren, R, 1992b. *Manual for SUSYNT speech synthesizer. Stockholm University, Dept of Phonetics.*