An Exploratory Experiment in Signal-to-Phoneme Classification

Harald Hammarström

Speech Tech HT2006 Term Paper

1 Introduction

The problem under study may be described as follows:

Input: A speech signal of appropriate duration representing a Swedish phoneme.

Output: The phoneme-class of the phoneme that the signal represents

It is important to note that the signals are assumed to be segmented into phoneme-duration already.

The Waxholm-database (from Speech, KTH) was available for training and testing. It contained about 60 000 phoneme-signal pairs and distinguishes between initial and non-initial phoneme occurrences.

Given the availability of this dataset, the idea was to attack the problem using supervised (and unsupervised) methods from machine learning. It was judged infeasible to attain phoneme-recognition, so phoneme-*class* recognition was chosen as the target instead.

2 Related Work

The present author is too little familiar with the area to comment on related work. This is an important weakness of the present study, especially as the problem appears to be general and interesting enough that there must be a lot of relevant work.

3 Experiments

3.1 Background

For (final) testing, 10 000 phoneme-signal pairs were selected from the database at random. This set will be refered to as the *test set*. The remaining ca 50 000 pairs was used for training and development. All subsequent mentions of training data refer to this set. The random selection of testing data ensures that testing and training data have similar distributions, and, in particular, entails that the same speaker may well be represented in both training and testing data. If phoneme-classes are thought of as something even more speaker independent than what can be inferred from the multitude of speakers in the Waxholm pairs, then this division is unfortunate – accordingly, in such a case, entirely different speakers should have been used for testing. Here however, the learning of phoneme-classes is to be understood as "learning whatever phoneme-class-like entities the annotated segments of the Waxholm speakers represent".

The Snack Sound Toolkit[1] 2.2.10 for Python was used to extract the power spectrum of a speech signal (at a given time or time-interval)[2]. The default values were taken for all parameters required – these and other acoustic data are shown in Table 1. In particular, this means that a spectrum is represented as a vector of length 256. Thus, a phoneme is represented as a series of logarithmic FFT power spectra of its speech signal, taken through its duration (in milliseconds). For some classes of phonemes, e.g. monophthong vowels, it makes sense to simplify the series into one *average* power spectrum, but for others, e.g. plosives, this is a very brutal assumption. Nevertheless, we will use average spectra in some experiments in order to get a tractable size of data to calculate with.

Somewhat unorthodoxically, a power spectrum vector (whether it is an average spectrum or taken at some instant in time) will be regarded as a function with 256 frequency values on the x-axis and the amplitude on the y-axis. (We will often refer to this kind of function as a *curve*). We will never be interested in the absolute values either of frequency or amplitude – only the general appearance of this curve. Therefore, each curve f will be

[1] `http://www.speech.kth.se/snack`

[2] In 116 cases (which have been subsequently ignored) I get an error that I don't understand when trying to extract a power spectrum "TclError: FFT window out of bounds".

Sampling Frequency	16 kHz
FFT length	512
Windowlength	128
Windowtype	Hamming
Skip	512
Preemphasis factor	0.0

Table 1: Summary of acoustic data.

normalized[3] to its average f^{avg}, or to put it mathematically:

$$f^{avg}(x) = \frac{f(x)}{avg_z f(z)} \tag{1}$$

Accordingly, the simplest kind of similarity metric between two phonemes will be the sum difference at each point of their respective normalized curves:

$$diff(f_1, f_2) = \sum_x |f_1^{avg}(x) - f_2^{avg}(x)| \tag{2}$$

The training data contains about 200 different types of phonemes, including differentiation over stressed/unstressed vowels, tonal-accent 1/2 for unstressed vowels, length (all phonemes), and also the distinction between inital and non-initial occurrences. (Most of these ca 200 types are vowels.) These can be grouped into the five phoneme-classes shown in Table 2, together with their occurrence-distribution in the training set.

Table 3 sum up the spectral heuristics I have been able to gather from overview/introductory books on acoustic phonetics (Ladefoged 2005 etc.) that were relevant for differentiating the Swedish phoneme-classes in question.

3.2 Orthodox Supervised Learning

I performed one baseline experiment using a traditional supervised learning method – to be more specific, an instance of a memory-based method (Duda,

[3]A more conventional normalization would have been to substract by average (logarithmic) spectrum. This was not used in the experiment due to a misunderstanding on the part of the author.

Plosive	18243	**37.3%**
Vowel	15054	**30.7%**
Approximant	6086	**12.4%**
Fricative	5329	**10.8%**
Nasal	4180	**8.5%**

Table 2: Distribution of phoneme-classes in the training set. There were also some 4000 annotations which do not represent phonemes of these classes, i.e. pause marks, sentence boundaries etc. Annotations of this kind will be grouped under a label 'None' subsequently.

Plosive	Weak energy/silence followed by a wide frequency band of energy
(Monophthong) Vowel	Strong Stable Voicing. (No movement of the formants over frequency.)
Approximant	Weak Stable Voicing. (No movement of the formants over frequency.)
Fricatives	High frequency-regions with random energy.
Nasal	Stable voicing. Formant structure similar to vowels but weaker and with extra resonances and anti-resonances

Table 3: Spectral heuristics as used in this study.

Class	# Cases	Accuracy
Plosive	38	**78.9%**
Vowel	20	**90.0%**
Fricative	13	**53.8%**
Approximant	12	**66.6%**
None	9	**33.3%**
Nasal	8	**100.0%**
	100	**74.0%**

Table 4: Accuracy broken up by phoneme-class for a memory based supervised learner (working on average spectra and a naïve similarity measure).

Hart, and Stork 2001). Average normalized power spectra (as above) were used as representation.

All the ca 50 000 examples of the training data were stored. In order to classify a previously unseen phoneme, all training examples are searched and the new instance is classified according to the one training example with the smallest difference to it. The measure for difference, or equivalently, highest similarity, was that of Equation 2.

Since the time for searching through all training data is prohibitive, I was only able to test this method with a random selection of a 100 cases from the test set. Identification of the phoneme could be done with 22% accuracy, whereas the accuracy for identification of phoneme-*class* is shown in Table 4.

74% accuracy is well beyond random (ca 17%) and frequency-baseline (ca 37%) and we can explain the low score for the 'None'-class as members of this "class" can't be expected to have a positive diagnostic in common. Because we use duration-averaged spectra, I expected e.g. the plosive class to be hard to distinguish from the vowel class, but such a tendency was not evidenced in the inspection of the corresponding confusion matrix. (In fact, I failed to detect any interesting patterns in a confusion matrix at all.) In no cases was there a zero-difference between a member of the test set and the training set, so true learning is going on.

3.3 Orthodox Unsupervised Learning

I performed one baseline experiment using a traditional unsupervised learning method – to be more specific, k-means clustering (Duda, Hart, and Stork

2001). Average normalized power spectra (as above) were used as representation.

The idea was the following, if the training examples are clustered into 6 classes (again, using the similarity measure of Equation 2), possibly, the six linguistically motivated phoneme-classes will emerge.

The results, given in Table 5, show that this is hardly the case. 1000 training examples (time prohibited no more) were k-means clustered into $k = 6$ classes. The six outcome clusters, when plotted on the background of the "true" classes of their members[4], turn out to have rather mixed membership. Therefore, it made little sense to try to turn the clustering outcome into a classifier of unseen phonemes.

The discrepancy between this clustering experiment and the previous supervised experiment, using the same similarity measure, is not so easy to understand. Perhaps the similarity measure is better suited to measure whether two phonemes are the same, and less well-suited to measure whether two different phonemes are of the same class.

3.4 Classification Using High-Low Sequences

One major drawback of traditional general-purpose machine learning methods is their unability to explain the results. The inner workings with all their intermediate numbers, usually as well as in the present study, shed no light on what assumptions or data items fail the intended generalizations. For this reason, usually as well as in the present study, I make experiments with hybrid methods where some assumptions are made explicit. It has an advantage in that assumptions that are interpretable can be discarded/revised/kept afterwards.

Average normalized power spectra (called curves) were used as representation in the first part of the experiment, and in the second we also look at variations over time inside the phoneme.

First, let's introduce the notion of a High-Low sequence, abbreviated HL-sequence. A HL-sequence the name for a string of H:s and L:s. Use $hl[i]$ denote the $i + 1$th character in a string hl, and $len(hl)$ to denote the length of the string. Thus, e.g. if $hl = HHHL$ then $hl[0] = H$ and $len(hl) = 4$.

[4]The true classes are known from the training data but this information is, of course, not available to the clustering algorithm.

Cluster 0		
Plosive	157	**65.4%**
None	45	**18.7%**
Fricative	27	**11.2%**
Approximant	8	**3.3%**
Vowel	2	**0.8%**
Nasal	1	**0.4%**

Cluster 1		
Plosive	84	**38.3%**
Vowel	32	**14.6%**
None	29	**13.2%**
Nasal	27	**12.3%**
Approximant	24	**10.9%**
Fricative	23	**10.5%**

Cluster 2		
Vowel	91	**46.6%**
Nasal	44	**22.5%**
Approximant	24	**12.3%**
Plosive	21	**10.7%**
Fricative	11	**5.6%**
None	4	**2.0%**

Cluster 3		
Fricative	42	**58.3%**
Plosive	28	**38.8%**
None	1	**1.3%**
Approximant	1	**1.3%**

Cluster 4		
Vowel	54	**46.1%**
Approximant	26	**22.2%**
Plosive	24	**20.5%**
Fricative	7	**5.9%**
Nasal	5	**4.2%**
None	1	**0.8%**

Cluster 5		
Vowel	108	**68.7%**
Approximant	27	**17.1%**
Nasal	10	**6.3%**
Plosive	8	**5.0%**
Fricative	4	**2.5%**

Table 5: The membership of six outcome clusters in an experiment with unsupervised learning.

A high-low sequence hl may be said to *fit* a curve f^{avg} according to the following metric.

$$fit(f, hl) = \sum_z penalty(f^{avg}(z), hl[z/r]) \tag{3}$$

$$penalty(x, v) = \begin{cases} |1 - x| & \text{if } v = H \text{ and } x < 1 \\ |1 - x| & \text{if } v = L \text{ and } 1 < x \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where / is shorthand for integer division and $r = len(f)/len(hl)$ (the $len(f)$ here means the number of z-values for which the function f is defined).

In other words, we are only concerned with locating above-average and below-average regions, not other fluctuations that do not cross the average mark. If a region is correctly matched as above/below average we do not care by how much below or above they are. If a region is incorrectly matched as above/below we penalize by how much off the mark it was. For example, a perfect fit (score 0) to the sequence LH is any function f whose $f(x)$ values are all below average for its first half, and all above average in the second half.

The intuition behind all this is to capture power spectra heuristics that are visible to the (trained) human eye in a noise-robust way. For example, (monophthong) vowels should have a two or three important H(igh):s in their curves, corresponding to their formants, whereas fricatives ought to have alternating H:s and L:s throughout. This suggests that working with HL-sequences of length 8 should be enough.[5]

Now let's look at some experimental data. Each phoneme with over 100 occurrences in the database was classified as to which of the 256 high-low sequence best described a set of 100 randomly selected occurrences of that phoneme. To be more precise, the sequence that "best describes" a set of occurrences was defined to be the sequence with smallest sum fit for all the occurrences. This was also a very time-consuming experiment so all (rather than 100) could not be used. However, it could be observed that the best-fitting high-low sequence for a phoneme appears to be consistent over several different 100 training-samples. This suggests that the high-low characterization-approach is not spurious.

[5]Though we will only use high-low sequences of length 8, some such sequences have straightforward abbreviations with each H or L are shrunk to a fix appropriate number of copies, e.g. $HL = HHLL = HHHHLLLL$ and $HLHL = HHLLHHLL$.

Sequence	Phoneme(s)
HHHHLLLH	#S
HLLH	$2S
LH	#. $'I: $E0 $g $d $"Y: $b $p: $"O: $'E: $"\3 $v #']: $‘A $'A: $"I $O $L $M $J $I $"A $G $E $A >pm $] $"] $p $pa $V $'O: $R #"]: $A: $']: $NG $[4 $'A #'I: $'[3 $"]: #'E #J #I #H OK #M $#H #L #'A $]: $'I $'E #F $'U: #p: $"E: #'] $#‘] $'U $'[$‘] $"A: #V $'\ $']
LHHH	$T $K
LHHHLLLH	$S
LLHHLHHH	$t
LLLHHHHH	$N #K #N #B $D #G $2T #D $2N #R #P $2D $P #T
LLLHLHHH	#SJ
LLLLLHHH	$F #pa #sm $k
9	87

Table 6: The best HL-sequence fit to 87 phonemes represented as normalized frequency-amplitude curves of their average (over the duration) power spectrum.

The HL-sequence categorizations for the different phonemes are shown in Table 6, using the transcription of the Waxholm data (Bertenstam, Blomberg, Carlson, Elenius, Granström, Gustafson, Hunnicutt, Högberg, Lindell, Neovius, Nord, Serpa-Leitao, and Ström 1995). (There were only 87 phonemes with 100 or more occurrences.) They are certainly not the sequences I expected them to be, but I did not investigate the reason why since there appear to be some interesting categorization in them anyway.

The natural point of continuation is, of course, to look at differentiation within the large LH class. Here we can try to undo the intuitively infelicitous simplification of only looking at the duration *average*, rather than at fluctuations throughout the duration. This would promise more as to distinguishing e.g. plosives from vowels.

The way Swedish phonemes differ duration-wise in the spectrum is basically that pressure comes and goes, e.g. for stops it goes out completely, not anything more complicated (such as diphthongs or true affricates etc.).

Thus we find ourselves in the same situation, namely, to seek a noise-robust method to tell when the pressure is high and when it is low. Thus, perhaps we can use HL-sequences for this too.

To be more precise, let's construct a function approximating a phoneme's duration-wise differences the following way. For each phoneme, pick 256 evenly spaced points on its duration on the x-axis and take the *average* (over 256 points of frequencies) amplitude on the y-axis for each of these points. The choice of 256 duration points, rather than each millisecond, was entirely due to computational limitations (and even so, selecting 256 spectra for each phoneme in the database takes half a day on my laptop). The decision for an average amplitude did not seem dangerous (after all, silence is silence at any frequency) and the HL-sequence approach requires a scalar value of some kind.

Again, let's look at the experimental data. As above, I decided to use high-low sequences of length 8. Intuitively, perhaps 4 or even 2 should be enough, but I also have poor intuitions as to the possible impact of selecting only 256-duration points (rather than all of them) – there might be a significant risk that we pick too many duration points with local amplitude minima – and length 8 did not seem too far off on the safe side.[6] Each phoneme with over 100 occurrences in the database was classified as to which of the 256 high-low sequence best described a set of 100 randomly selected occurrences of that phoneme. To be more precise, the sequence that "best describes" a set of occurrences was defined to be the sequence with smallest sum fit for all the occurrences. This was also a very time-consuming experiment so all (rather than 100) could not be used. However, it could be observed that the best-fitting high-low sequence for a phoneme appears to be consistent over several different 100 training-samples. This suggests, again, that the high-low characterization-approach is not spurious for duration-wise fluctuations either.

The HL-sequence categorizations for the different phonemes are shown in Table 7. (They will be refered to as the duration-amplitude HL-sequence of the phoneme, to be distinguished from the frequency-amplitude HL-sequence that are shown in Table 6.) This time too, the sequences were not as I expected them to be, but I did not investigate the reason why since there appear to be some interesting categorization in them anyway. The important

[6]Though after seeing the results, there are good indications that 8 was too much – one would have wanted LHHHHHLL and LHHHHLLL to go in the same class as HHHHHLLL.

thing is that the stops come out quite well.

This is, so far, essentially an unsupervised approach and it is not straightforward to evaluate or to put together to produce a signal to phoneme-class classifier that can be compared with the one in Table 4. In contrast to the unsupervised approach evaluated in Table 5, we do not have exactly 6 classes – putting the two HL-characterizations together would yield a maximum of 9x19 = 171 classes. However, luckily, only 31 classes naturally suggest themselves the following way. Construct a matrix of the duration-amplitude HL-sequences of Table 7 and the frequency-amplitude HL-sequences. Let entries in the matrix be the intersection of the sets of phonemes for the HL-sequences in question. From this we happen to get 31 non-empty classes and all 87 phonemes end up in at least one class. They are shown in Table 8.

These categories could be evaluated in terms of precision and recall over the six linguistically motivated categories, and we'd get some numbers to try to interpret. A more interesting kind of evaluation seems to be to benchmark it against the clustering experiment in Table 5. However, this would be an unfair comparison as one has 31 classes whereas the other has 6. To overcome this imbalance I tried clustering the 31 classes into 6 using *only* the labels for the categories, i.e. the HL-sequences. This is intuitively justified because of the meaning that the HL-sequences are intended to carry. After k-means clustering on the labels (using the Hamming distance as the similarity measure), the six classes in Table 9 emerge. These six classes can now be evaluated on the same scale as the previous unsupervised approach, as shown in Table 10.

To my eye, the HL-approach fares significantly better on this comparison. It still remains, however, to transform the outcome into a signal to phoneme-class classifier. One other way to benchmark it against the supervised classifier is to pick n test pairs, some of them will belong to the same true class and some not, in order to check how well they predict the same-class property of the pair. But this has not been done yet due to lack of time.

References

Bertenstam, J., M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, J. Högberg, R. Lindell, L. Neovius, L. Nord, A. Serpa-Leitao, and N. Ström (1995). Spoken dialogue data collected

Sequence	Phoneme(s)
HHHHHLLL	#H #N #M #L #p: $V $#H #P #V
HHHL	#K $K $2T #B $D #G $p: #D $2N $T $2D $P #T
HHHLLLHL	#SJ
HHHLLLLL	#F
HHLLLLLH	$’]:
HL	#R $d #J
HLLLLLHH	#S $’[3 $"\3 $#‘] $’E
HLLLLLLH	$‘]
LH	#sm $’U: $M #. $"Y: $"O: $’O:
LHHHHHLL	$G
LHHHHLLL	$p
LHHL	$k
LLHHHHLH	$J
LLHHHLHH	$g
LLLH	$"I $O #I $’A $‘A #’A $"E: #"]: $’I #’E $’E: $A $]: $[4 $] $A: $’[$2S $S $’\ $"A:
LLLHHHHH	$N $R
LLLHHHHL	$t $L $NG
LLLLLHHH	$E $’I: $I $"A #pa $’U $E0 $"] #’I: $v #’] $’A: $’]
LLLLLLLH	$F $"]:
19	87

Table 7: The best HL-sequence fit to 87 phonemes represented as normalized duration to average-amplitude (average over 256 frequency points) of their power spectrum.

('HHHHHLLL', 'LH')	#H #M $#H #L #p: $V #V
('HHHHHLLL', 'LLLHHHHH')	#P #N
('HHHL', 'LH')	$p:
('HHHL', 'LHHH')	$T $K
('HHHL', 'LLLHHHHH')	#K #B $D #G $2T #D $2N $2D $P #T
('HHHLLLHL', 'LLLHLHHH')	#SJ
('HHHLLLLL', 'LH')	#F
('HHLLLLLH', 'LH')	$']:
('HL', 'LH')	#J $d
('HL', 'LLLHHHHH')	#R
('HLLLLLHH', 'HHHHLLLH')	#S
('HLLLLLHH', 'LH')	$'[3 $"\3 $#‘] $'E
('HLLLLLLH', 'LH')	$‘]
('LH', 'LH')	$'U: $M #. $"Y: $"O: $'O:
('LH', 'LLLLLHHH')	#sm
('LHHHHHLL', 'LH')	$G
('LHHHHLLL', 'LH')	$p
('LHHL', 'LLLLLHHH')	$k
('LLHHHHLH', 'LH')	$J
('LLHHHLHH', 'LH')	$g
('LLLH', 'HLLH')	$2S
('LLLH', 'LH')	$"I $O #I $'A $‘A #'A #"]: $'I #'E $[4 $A $'E: $"E: $'\ $'[$] $]: $A: $"A:
('LLLH', 'LHHHLLLH')	$S
('LLLHHHHH', 'LH')	$R
('LLLHHHHH', 'LLLHHHHH')	$N
('LLLHHHHL', 'LH')	$L $NG
('LLLHHHHL', 'LLHHLHHH')	$t
('LLLLLHHH', 'LH')	$'I: $I $"A $E #'I: $E0 $"] $'U $v #'] $'A: $']
('LLLLLHHH', 'LLLLLHHH')	#pa
('LLLLLLLH', 'LH')	$"]:
('LLLLLLLH', 'LLLLLHHH')	$F

Table 8: Outcome classes by combining the duration-amplitude HL-sequences with the frequency-amplitude HL-sequences.

Cluster	Classes Merged
0	('HHHHHHLL', 'LLHHHHHH') ('HHHHHHLL', 'LLLHHHHH') ('LHHHHLLL', 'LLLLHHHH') ('HHHHHLLL', 'LLLHHHHH') ('LHHHHHLL', 'LLLLHHHH') ('HHHHHLL', 'LLLLHHHH') ('HHHHHLLL', 'LLLLHHHH')
1	('LLLLLLLH', 'LLLLHHHH') ('LLLLLLHH', 'LLLLHHHH') ('HHLLLLLH', 'LLLLHHHH') ('HLLLLLLH', 'LLLLHHHH') ('LLLLLLLH', 'LLLLLHHH') ('HLLLLLHH', 'LLLLHHHH')
2	('LLHHHHLL', 'LLLLLHHH') ('LLLHHHHL', 'LLHHLHHH')
3	('HLLLLLHH', 'HHHHLLLH') ('LLLLLLHH', 'LHHHLLLH') ('LLLLLLHH', 'HHLLLLHH')
4	('LLLHHHHL', 'LLLLHHHH') ('LLLLHHHH', 'LLLLHHHH') ('LLLLHHHH', 'LLLLLHHH') ('LLHHHLHH', 'LLLLHHHH') ('LLLHHHHH', 'LLLHHHHH') ('LLLLLHHH', 'LLLLLHHH') ('LLHHHHLH', 'LLLLHHHH') ('LLLHHHHH', 'LLLLHHHH') ('LLLLLHHH', 'LLLLHHHH')
5	('HHHHLLLL', 'LLLLHHHH') ('HHHLLLHL', 'LLLHLHHH') ('HHHLLLLL', 'LLLHHHHH') ('HHHHLLLL', 'LLLHHHHH')

Table 9: Outcome of clustering the combined HL-classes into six larger classes, based on the hamming distance between the (names of the) combined HL-classes.

$2T $2N $#H $2D $p $K #K #H #N #M #L $G $D #G $p: #D #p: #B $V $T #P #V $P #T	0	
Plosive	16	**66.6%**
Fricative	4	**16.6%**
Nasal	3	**12.5%**
Approximant	1	**4.1%**

#"]: $#‘] $[4 $"\3 $’[3 $] $"]: $"I $O #I $’A $‘A $’E #’A $]: $’I #’E $’E: $A $F $"E: $A: $’[$‘] $’]: $’\ $"A:	1	
Vowel	26	**96.2%**
Fricative	1	**3.7%**

$t $k	2	
Plosive	2	**100.0%**

#S $2S $S	3	
Fricative	3	**100.0%**

$N $E #. $’I: $E0 $g $NG #’] $"O: #’I: $v $’A: #sm $L $M $J $"Y: $"A #pa $I $’U: $’O: $’U $"] $R $’]	4	
Vowel	15	**57.6%**
None	3	**11.5%**
Nasal	3	**11.5%**
Approximant	3	**11.5%**
Fricative	1	**3.8%**
Plosive	1	**3.8%**

#J $d #R #F #SJ	5	
Fricative	2	**40.0%**
Approximant	2	**40.0%**
Plosive	1	**20.0%**

Table 10: The membership of six outcome clusters consisting of combined HL-classes. Each double-lined box is a cluster. The membership, cluster id-number (1-6) and phoneme class composition is given.

in the waxholm project. *STH-QPSR, KTH 1*, 49–74.

Duda, R. O., P. E. Hart, and D. G. Stork (2001). *Pattern Classification* (2 ed.). Wiley, New York.

Ladefoged, P. (2005). *Vowels and Consonants* (2 ed.). Blackwell, Oxford.