ACOUSTIC PARAMETERS OPTIMISED FOR RECOGNITION OF PHONETIC FEATURES

Anya Varnich Hansen

Center for PersonKommunikation,
Aalborg University, DK-9220 Aalborg, Denmark.
Tel. +45 96358671, FAX: +45 98151583, E-mail: avh@cpk.auc.dk

ABSTRACT

Speaker variability is a major problem in today's state-of-the-art speech recognition systems. Parameterisation of speech in terms of Acoustic Parameters (APs) motivated by phonetic feature theory has shown to be more robustness to speaker variability as compared to cepstral coefficients when tested on the task of broad-class recognition [1]. Also APs has been successfully applied for identification of semivowels [2,3].

The aim of the present study is to investigate the use of APs for phoneme recognition. An extended set of features is used to distinguish between all phonemes in the TIMIT database and APs related to the extended feature set are found in literature. A separability measure is calculated to investigate the importance of the suggested APs for the separation of phonemes and feature classes.

Results show that the APs that are the most important for separation of classes of phonetic features are also the most important for separation of phonemes classes. This indicates that phonemes can be recognised on the basis of phonetic features captured by the use of APs. However much work still needs to be done to understand and reliably extract all of the acoustic correlates of the phonetic features applied.

1. INTRODUCTION

Phonetic features [4] are the minimal units needed to distinguish among the most similar sounds in a language (for instance the feature voiced distinguishes the stops /b/ and /p/).

The acoustic properties of a given phonetic feature can be recognised from the speech signal by use of a carefully selected set of Acoustic Parameters (APs) [1,2,3,5]. Energy in different frequency bands measured relative to the energy in other frequency bands, the rapidity with which the spectrum changes and measurements of the relative placement of the formant frequencies are all examples of APs.

A representation of speech signals in terms of APs motivated by phonetic feature theory have shown to be more robust to inter-speaker variability as compared to cepstral based parameters when tested on the task of broad-class recognition [1]. The robustness to speaker variability is obtained by defining the APs in a relational manner across time and frequency as described above. A parameterisation of the speech signal that reduces the inter-speaker variability may result in Automatic Speech Recognition (ASR) systems that are more robust to speakers not included in the training data. An additional advantage may be minor demands to the amount of training data due to less variability.

Describing the speech signal in terms of APs motivated by phonetic feature theory also allows a directly incorporation of speech knowledge in the recognition and it serves as a tool for understanding acoustic phonetics and contextual variability [2].

The present study combines and elaborates on principles described in other feature based approaches. The fact that phonetic features manifest themselves in a set of APs is known from acoustic phonetics [5], and the principle of identifying features on the basis of APs is successfully applied in a broad-class recognition task in [1] and for identification of semivowels in [2,3]. The aim of the work presented in this paper is to be able to recognise phonemes on the basis of APs.

A set of phonetic features on the basis of which the phonemes in the TIMIT database are uniquely described and a set of APs related to the presence of the different phonetic features are found. A separability measure is applied to evaluate the importance of the suggested APs for the distinction between phonemes and phonetic feature classes.

2. MATERIALS AND METHODS

2.1 Phonetic feature set

A set of phonetic features that distinguish between all phonemes in the TIMIT database was designed on the basis of the theory described in [4]. The phonetic feature set contains the features: sonorant, syllabic, consonantal, high, low, front, back, anterior, coronal, round, voice, continuant, nasal and strident. Each of the phonemes in the TIMIT database are uniquely described in terms of these features. As an example the phoneme /aa/ is described as: [+sonorant, +syllabic, -consonantal, -high, +low, -front, +back, 0anterior, 0coronal, -round, 0voice, 0continuant, 0nasal, 0strident]. The phonetic features having the value 0 are only relevant for distinction of consonants, i.e. they are not relevant for vowels.

2.2 Acoustic parameters

Table 1 shows the applied phonetic features and their related APs found in literature [1,2,5,6,7,8,9]. Several other APs than the ones mentioned in Table 1 can be found. The *consonantal* feature can be captured from measuring the rapidity with which the spectrum changes. Rapid spectrum changes distinguish consonants from vowels and glides where the spectrum changes much more slowly [5]. In [2,3] the first difference of adjacent spectra based on a bank of 40 linear critical band filters are used as AP. In [1] APs for identifying non-*continuant* sounds are given and in [7] an auto-correlation function is suggested for the detection of *strident* sounds, see also [9] for details about this feature.

In the present study focus has initially been set on APs based on energies (E) in different frequency bands, voicing probability, the relative location of the first three formants (F1, F2, F3) and the formants bandwidth (bw1, bw2, bw3, bw4), to investigate the separability of this simple set of APs before more complex measures[2] are derived.

Feature	Acoustic parameter	Property / reference
sonorant	E(100-400Hz)	high [1]
	E(0-2kHz)-E(2-8kHz)	high [1][1]
	E(0-300Hz)-E(3.7-7kHz)	high [2]
	voicing probability (VP)	high [1]
syllabic	E(600-2800Hz)	high [1,2][2]
	E(2000-3000Hz)	high [1,2][2]
	E(0-8kHz)	high [5,7]
consonantal	E(0-8kHz)	low [6]
	bandwidths: bw1, bw2, bw3, bw4	broad [7]
high	E(0-500kHz)-E(500-1000)	high [5]
	F1-F0	low [2]
back	E(800-1900kHz) - E(2-3kHz)	high [5]
	F2-F1	low [2]
low	E(0-500kHz)-E(500-1000)	low [5]
front	E(800-1900kHz) - E(2-3kHz)	low [5]
	F2-F1	high [2]
round	no APs derived	
anterior	no APs derived	
coronal	no APs derived	
voice	E(0-400Hz)	high [6,7]
continuant	no APs derived	
nasal	abs(250Hz-F1), abs(2500Hz-F2), abs(3250Hz-F3)	low [7,8]
strident	E(6-8kHz)	high [6,9]

Table 1 The applied phonetic features and their related APs found in literature. All energy measures (E) are calculated relative to the maximum energy across the utterance.

For some features no APs have been extracted yet as the descriptions in literature must first be transformed to measurable APs. This is the case for the feature *round*: "as sounds become more rounded, the frequencies of the higher formants decrease. But the situation is complicated in that the effect is greater in the third formant for front vowels, and in the second formant for back vowels" [8] pp.196. In this case it would be optimal to normalise with respect to a front/back measure.

For the features *anterior* and *coronal* no description of how they manifest themselves in the acoustic signal has been found in literature by now.

2.3 Extraction of acoustic parameters

Formants and bandwidth were extracted from the acoustic speech signal by use of the 'formant' tool in the Entropic software package [10]. The energy in different frequency bands was calculated by use of the Matlab signal processing tool 'specgram' [11] that computes the windowed discrete-time Fourier transform of a signal using a sliding window. A 10 ms window and a 5 ms frame rate was applied. The logarithm was taken to transform the energy measures to dB and all energies were normalised with respect to the maximum energy across the.

2.4 The V-ratio - a separability measure

Different methods [13,14,15] can be used to determine which of the 20 APs that best distinguish between feature and phoneme classes. In the present study a seperability measure: the V-ratio that are closely related to Fishers linear discriminant function [13] has been chosen.

The V-ratio is a measure for how well an AP *i* separates the different classes *C*. The higher V-value, the better separation between classes.

$$V_i = \frac{\textit{deviation of the means (over all classes)}}{\textit{mean of the deviation (within each class)}} = \frac{\sigma_{\mu_i}}{\mu_{\sigma_i}}$$

$$V_i = \frac{\sqrt{\frac{1}{C}\sum_{k=1}^{C}\left(\mu_{i,k} - \mu_i\right)^2}}{\frac{1}{C}\sum_{k=1}^{C}\sigma_{i,k}}$$

As opposed to the methods described in [13,14] this method of selecting the most discriminative APs has the advantage that it does not require training of classifiers for the features and phonemes investigated. The drawback is that the selection of APs by use of the V-ratios does not take into account the recognition rate of the classifier.

To compare the V-ratio method to the method in [13] an experiment was performed selecting the 24 most important coefficients of a set of 12 cepstral coefficients including energy and the delta and delta-delta coefficients (in total 39 coefficients) by use of the V- ratio method. 19 of the 24 coefficients selected was the same as those selected by the method described in [13].

The ranking experiment was performed on the SISX training sentences in the TIMIT database [12]. The selection of coefficients performed in [13] was also based on the TIMIT database, but the applied dataset was not given.

2.5 Database

Data from the TIMIT database [12] sampled at 16kHz is applied for evaluation of the importance of the different APs through calculation of V-ratios. TIMIT contains 5 hours of speech and is divided into a test and a training corpus. The 3696 SISX training sentences are applied in the ranking experiments described below.

3. EXPERIMENTS

3.1 Ranking of APs according to their V-ratio

The APs are ranked according to their importance for distinction between:

1. +feature, -feature and 0feature classes
2. phonemes (each phoneme constitutes a class)

The first ranking experiment is performed for each of the 14 features. For the features sonorant, syllabic, consonantal, high, back and low there are two classes: the

[1] In [1] the AP is not normalised with the maximum energy across the utterance.

[2] In [1,2] to capture the syllabic feature an energy peak is detected and measured relative to surrounding energy deeps in the neighbouring consonant - for details see [3]. E.i. in the present study a simplified measure has been applied.

Feature → **AP** (designed to capture the feature in brackets) ↓	**son**	**syl**	**cons**	**high**	**back**	**low**	**fro**	**ant**	**cor**	**rou**	**voi**	**cont**	**nas**	**str**	**pho-neme**
1 p-voi (sonorant)	2.000 1	1.424 3	1.544 1	0.831 6	0.955 5	1.245 5	1.495 3	0.329 1	0.248 4	1.333 3	0.644 1	0.461 2	0.418 6	0.669 4	1.499 1
2 bw1 (consonantal)	0.316 13	0.246 14	0.249 13	0.097 18	0.191 14	0.285 15	0.349 13	0.124 15	0.191 5	0.319 11	0.242 10	0.197 13	0.234 12	0.376 8	0.560 14
3 bw2 (consonantal)	0.264 15	0.208 15	0.236 14	0.172 15	0.136 16	0.180 17	0.315 14	0.188 11	0.133 10	0.245 15	0.176 13	0.186 15	0.179 15	0.266 11	0.306 17
4 bw3 (consonantal)	0.097 18	0.099 19	0.104 18	0.097 19	0.078 19	0.080 19	0.103 20	0.038 19	0.022 19	0.105 18	0.040 20	0.044 20	0.035 20	0.045 20	0.150 19
5 bw4 (consonantal)	0.070 19	0.077 20	0.089 19	0.084 20	0.058 20	0.058 20	0.106 19	0.022 20	0.019 20	0.081 20	0.053 19	0.049 19	0.042 19	0.058 18	0.131 20
6 F1-F0 (high/low)	0.610 11	0.519 10	0.551 10	0.436 8	0.426 9	0.478 11	0.640 10	0.197 10	0.148 9	0.606 10	0.371 8	0.260 9	0.356 9	0.457 6	0.603 13
7 F2-F1 (back/front)	0.023 20	0.112 18	0.050 20	0.152 17	0.119 18	0.264 16	0.115 18	0.105 18	0.116 13	0.085 19	0.102 18	0.102 18	0.254 10	0.057 19	0.383 16
8 abs(250-F1) (nasal)	0.631 10	0.522 9	0.558 9	0.426 9	0.413 11	0.476 12	0.664 9	0.206 9	0.151 8	0.627 9	0.405 6	0.265 8	0.382 8	0.459 5	0.629 11
9 abs(2500-F2) (nasal)	0.296 14	0.322 12	0.275 12	0.243 12	0.294 13	0.450 13	0.356 12	0.111 17	0.130 11	0.258 12	0.152 14	0.189 14	0.242 11	0.138 17	0.552 15
10 abs(3250-F3) (nasal)	0.154 16	0.157 17	0.195 16	0.161 16	0.121 17	0.118 18	0.225 15	0.146 14	0.113 14	0.249 14	0.140 15	0.141 17	0.200 13	0.158 16	0.261 18
11 E(100-400) (sonorant)	1.620 2	1.446 1	1.473 2	1.035 2	1.086 2	1.370 2	1.518 1	0.212 8	0.107 15	1.425 1	0.463 5	0.209 12	0.448 3	0.343 10	1.353 2
12 E(0-2k)-E(2-8k) (sonorant)	0.836 7	0.607 8	0.589 8	0.303 13	0.511 8	0.702 7	0.800 7	0.254 3	0.309 3	0.735 8	0.578 3	0.377 4	0.537 2	0.882 3	1.244 4
13 E(0-300)-E(3.7-7k) (sonorant)	0.791 8	0.507 11	0.533 11	0.268 11	0.418 10	0.502 10	0.727 8	0.244 4	0.332 2	0.743 7	0.634 2	0.445 3	0.725 1	0.934 1	1.197 6
14 E(600-2800) (syllabic)	1.273 5	1.320 5	1.328 5	0.979 4	1.013 4	1.340 4	1.261 4	0.232 5	0.048 17	1.127 4	0.229 12	0.324 6	0.186 14	0.200 12	1.206 5
15 E(2-3k) (syllabic)	1.009 6	1.098 6	1.131 6	0.959 5	0.926 6	1.050 6	0.917 6	0.224 6	0.041 18	0.816 6	0.105 17	0.258 10	0.096 17	0.258 12	0.896 8
16 E(0-8k) (consonantal)	1.307 4	1.436 2	1.421 4	1.145 1	1.151 1	1.445 1	1.145 5	0.181 12	0.098 16	1.058 5	0.107 16	0.371 5	0.064 18	0.247 13	1.155 7
17 E(0-500)-E(0.5-1k) (high/low)	0.106 17	0.201 16	0.162 17	0.174 14	0.147 15	0.357 14	0.194 16	0.118 16	0.155 6	0.164 16	0.237 11	0.300 7	0.447 4	0.224 14	0.612 12
18 E(0.8-1.9)-E(2-3k) (front/back)	0.376 12	0.306 12	0.229 15	0.271 10	0.298 12	0.535 9	0.401 11	0.148 13	0.153 7	0.257 13	0.300 9	0.156 16	0.136 16	0.408 7	0.860 10
19 E(0-400) (voiced)	1.612 3	1.423 4	1.456 3	1.012 3	1.066 3	1.343 3	1.517 2	0.216 7	0.116 12	1.420 2	0.471 4	0.222 11	0.446 5	0.371 9	1.344 3
20 E(6-8k) (strident)	0.672 9	0.652 7	0.639 7	0.621 7	0.664 7	0.673 8	0.170 17	0.259 2	0.369 1	0.118 17	0.372 7	0.570 1	0.395 7	0.899 2	0.871 9

Table 2 For each of the APs suggested in literature a separability measure (the V-ratio) is calculated, to determine the importance of the AP in the separation of for instance the +sonorant class of phonemes from the -sonorant class of phonemes (see the third column). For each of the features V-ratios are given together with the resulting ranking of the APs. In the last column each phoneme constitute a class and the V-ratios describe the importance of a given AP for the separation of all phonemes. The numbers given in the left column are used when referring to the different APs in the discussion section, below the AP number is given the name of the feature it was designed to capture.

class of phonemes in which the feature is present (denoted +feature) and a class of phonemes where the feature is not present (denoted -feature). For the remaining features the phonemes for which the feature is not relevant constitutes a third class denoted the 0feature class. In the second ranking experiment each of the phonemes constitutes a phoneme class.

The first ranking experiment will show whether APs expected on the basis of acoustic phonetic knowledge to be important for the caption of a certain features are important according to V-ratios calculated on a large set of data. The second ranking will indicate which features to use when training phoneme models for speech recognition.

4. RESULTS

Results are given in Table 2. For each feature the V-ratios are given for each of the APs together with the ranking given as numbers from 1 to 20 where 1 is the most important AP for the given feature.

5. DISCUSSION

Comparing the ranking of APs for the features and phonemes it is seen, that it is the same APs (see grey shaded areas in Table 2) that are important for recognition of features and phonemes. This indicates that phonemes can be recognised on the basis of phonetic features captured in the acoustic speech signal by the use of APs.

For all features relevant for the distinction of vowels e.i. son, syl, cons, high, back, low, fro, rou the APs 1, 11, 14, 16 and 19 have high V-values. All these APs are highly related to the detection of the feature *sonorant*, which split phonemes into the two major classes: sonorants and obstruents. The presence of one phonetic feature can give information on the presence of absence of others. Thus if a phoneme is -sonorant, it is known that it is also *-syllabic, +consonantal* and *-low* therefore the same set of APs are important for these features, for the distinction into major classes, however some of the APs with lower V-values may be responsible for the fine phonetic distinctions.

The voicing probability (AP1), AP 11 and 19 capturing the strong low frequency energy resulting from voicing are important APs for the *sonorant* feature. The presence of formants in the mid frequencies (AP14) and a high overall energy (AP16) are also resulting in high V-values for these APs. AP 12 and 13 were designed to capture the sonorant feature as having a strong low-frequency energy as compared to the energy in the higher frequencies. However these APs do not have as high V-ratios as the ones first mentioned. Some of the APs such as AP1, 11 and 19 are highly correlated, therefore a constraint should be set on the maximum allowable correlation among APs when they are selected for training of for instance a HMM classifier.

AP 14 and 15 that was designed to capture *syllabic* peaks in the mid-frequency regions get relatively high V-ratios and thus the numbers 5 and 6 in the ranking. Even higher V-ratios might be obtained with a more advanced APs[2]. Though written in [7] that "phonemes possessing the *consonantal* feature are acoustically characterised by a broadening, reduction and fusion of formants and formant regions due to zeros, high damping or transient variations of formant frequencies" the bandwidths (AP2, 3, 4, 5) do not seem to be relevant APs.

Different values of the same AP can characterise different features, thus AP16 is high for sonorant sounds and low for consonantal sounds that are characterised by having a low total energy.

Of the APs designed to capture the features *high/low* AP6 is superior to AP17. While for the features *front/back* AP18 have higher V-ratios than AP7.

No APs was designed to capture the features *anterior*, *coronal*, *round* and *continuant*, so it is not surprising that the V-ratios for these features in general are low. The feature *round* is an exception. This feature is only relevant for vowels, therefore the APs that captures the characteristics of vowels (e.i. the +sonorant, +syllabic, - consonantal features) are given a high priority. The high V-ratios for these APs may express that vowels are well separated from other phonemes, thus the high V-ratios do not ensure, that rounded and unrounded vowels are well separated.

Of the APs 8, 9 and 10 suggested for capturing the feature *nasal* AP8 shows to be the most important. The very low first formant centred at about 250Hz, that characterise nasal sounds [8] is also captured by AP1, 11, 12, 13, 17 and 19. AP20 designed to capture the strident feature show to be highly relevant and gets a priority as the second most important feature.

In general the features that are not relevant for vowels (ant, cor, voi, cont, nas, str) have low V-values indicating that more efforts must be put in finding APs that better capture these features that perform the fine phonetic distinctions. As opposed to the cepstral representation of speech it is possible to put special efforts in capturing 'difficult feature'.

6. CONCLUSION

Results show that the APs that are the most important for separation of classes of phonetic features are also the most important for separation of phonemes classes. This indicates that phonemes can be recognised on the basis of phonetic features captured by the use of APs. However much work still needs to be done to understand and reliably extract all of the acoustic correlates of the phonetic features applied.

7. REFERENCES

[1] Bitar, N. and Espy-Wilson, C. (1996) "Knowledge-based parameters for HMM speech recognition", ICASSP'96, pp 29-32.

[2] Espy-Wilson, C. Y. (1994) "A feature-based semivowel recognition system", J. Acoust. Soc. Am. 96 (1), July 1994, pp.65-72.

[3] Espy-Wilson, C. Y. (1992) "Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English", J. Acoust. Soc. Am. 92 (2), Pt.1, August 1992, pp.736-757.

[4] Chomsky, N. and Halle, M. (1968) "The Sound Pattern of English", Harper and Row, New York.

[5] Stevens, K. (1980) "Acoustic correlates of some phonetic categories", J. Acoust. Soc. Am. 68 (3), pp. 836-842.

[6] Clark, J. and Yallop, C. (1991) "An introduction to phonetics and phonology", Basil Blackwell, Inc. Cambridge, Massachusetts, USA.

[7] Jakobson, R., Fant, G., Halle, M. (1951) "Preliminaries to speech analysis", The MIT Press.

[8] Ladefoged, P. (1993) "A course in phonetics", Harcourt Brace Jovanovich College Publishers.

[9] Utman, J.A. and Blumstein, S.E. (1994) "The Influence of Language on the Acoustic Properties of Phonetic Features: A Study of the Feature [Strident] in Ewe and English", Phonetica 1994; 51 pp.221-238.

[10]"Waves+ Manual" (1993) version5.0, Entropic Research laboratory, Inc.

[11]Krauss, P. (1993) "Signal Processing TOOLBOX for use with Matlab", The Math Works, Inc.

[12]Garofolo, J. et. al. (1993) "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM" U. S. Department of Commerce, National Institute of Standards and Technology, USA

[13]Bocchieri, E.L. and Wilpon, J.G.(1993) "Discriminative feature selection for speech recognition", Computer Speech and Language, 1993, 7, pp.229-246.

[14]Phillilps, M. and Zue, V. (1992) "Automatic Discovery of Acoustic Measurements for Phonetic classification", ICSLP'92, vol. 1, pp 795-798.

[15]Duda, R. and Hart, P. (1973) "Pattern Classification and Scene Analysis", John Wiley and Sons, Inc pp 114-121.