Proceedings of The Institute of Acoustics

MID-CLASS PHONETIC ANALYSIS FOR A CONTINUOUS SPEECH RECOGNITION SYSTEM

J. Dalby, J. Laver, and S.M. Hiller

Centre for Speech Technology Research
University of Edinburgh

INTRODUCTION

A hierarchical approach to acoustic phonetic analysis of speech for the automatic segmentation and labelling component of a connected speech recognition system is motivated by two characteristics of spoken English. The first is that the phonological grammar of syllable and word structure of the language sharply constrains the allowable sequences of segments in words. The second is that for some classes of speech sounds, at least, relatively coarse and hence presumably robust identification and classification techniques can be applied to partition the speech waveform into segments which correspond to sets of phonemes.

In a series of studies conducted at MIT, V. Zue and his colleagues [1, 2, 3] have shown that even large lexicons can be partitioned into rather small sets of words with equivalent spellings in broad phonological class terms when only six phoneme sets, stop, fricative, nasal, liquid, glide, and vowel are defined. For example Shipman and Zue [1] report that the average frequency normalized equivalence class size of words in a 20,000 word lexicon is less than 40 and that over one third of the entries have spellings which are unique in broad class terms. On the assumption that coarse classification of speech into segments made up of these broad class categories by a feature-based automatic acoustic analysis system can be performed more reliably than phoneme identification, the reduction in the lexical hypothesis space provided by knowledge of the broad class membership of segments in the input provides useful back-up information to a lexical access component which is designed to take advantage of it.

In the acoustic-phonetic analysis component of the automatic speech recognition system being developed at CSTR we have taken this idea a step further and have defined a set of MidClass phoneme categories, such as voiced stops, voiceless stops, strong and weak fricatives, nasals, several vowel classes, etc. We believe that these MidClass phoneme sets will prove to be easier to identify reliably than (at least some of) the individual phonemes in the input speech and can show that extracting this phonological MidClass information from the speech signal limits the lexical search space of our system in a useful way. For the past several months we have been developing a feature-based acoustic phonetic rule set for automatic MidClass segmentation and labelling of speech which, for some of the MidClass categories shows promise for future development.

MID-CLASS PHONETIC ANALYSIS

SYSTEM OVERVIEW

To provide an existence proof, as it were, that feature-based acoustic phonetic analysis of continuous speech can be done and that a phoneme hypothesis lattice could be produced which could be used by an appropriately designed lexical access component, a simple acoustic phonetic segmentation system was developed. The data for development of the segmentation and labelling rules came from a small set of 16 utterances produced by one speaker. The sentences were chosen from a larger set of 'phonemically dense sentences' [4] which were devised to contain various classes of phonemes in several positions in the sentence. For example, the sentence 'Three chefs face a thief' contains strong and weak voiceless fricatives,front vowels, and /r/. These sentences contained examples of all of the MidClass phonemes defined in our system but obviously only a few of the possible segmental or prosodic environments for the different MidClasses are represented in such a small sample of speech. The evaluation data base consisted of these same 16 sentences spoken on a different occasion by the original talker, the same 16 produced by a second talker, and a further set of 16 different sentences which were similar in their phonological content produced by the original talker. The 64 sentences were low pass filtered at 8 KHz, sampled at 12 bit resolution at 16 KHz, and, after time-aligned phonemic-level transcriptions were made by a phonetician, the sentences were analyzed using the StarPak batch signal processing software developed at CSTR [5].

The signal processing algorithms employed in the current system derive a set of acoustic parameter vectors for each of the analyzed utterances. These parameter vectors, produced in a C computing environment are then passed to a Lisp processing environment where the acoustic phonetic MidClass analysis rules are applied using the CSTR SegLab software [6].

Two types of acoustic-phonetic rules, threshold rules (TRules) and sequence rules (SRules) are invoked to detect and classify the MidClass segments in the input data. A third set of rules, which we call APRules is used to provide the analysis system with derived acoustic parameters by computing simple functions such as the first or second-order difference function of the input acoustic parameter vectors. Threshold rules test the acoustic parameters (or derived parameters) against criterion values and create feature segments for the threshold category wherever the relevant acoustic parameter meets or exceeds the threshold. SRules take these feature segments as input and set minimum and maximum duration criteria for a specified feature or set of features, allow for the specification of a sequence of features (with minimum and maximum durations), and specify a left and right context for the feature sequence. SRules allow for both positive and negative values of the features or feature sets to be specified and there is no limit to the number of features which can be specified in an SRule. The output of the SRules is a set of MidClass segmentation labels. Multiple and overlapping MidClass segmentation hypotheses are generated in this component and passed to the lexical access component of the system. A sample of the output of the acoustic-phonetic rules set of the current system is shown in Figure 1, which shows the segmented human transcription for the sentence 'Three chefs face a thief' at the bottom of the display and the MidClass hypotheses generated by the rules above the transcription. Each MidClass category is represented by a capital letter abbreviation such as B for voiced stops, S for strong voiceless fricatives, FV for front vowels, D for diphthongs, G for glides, etc.

MID-CLASS PHONETIC ANALYSIS

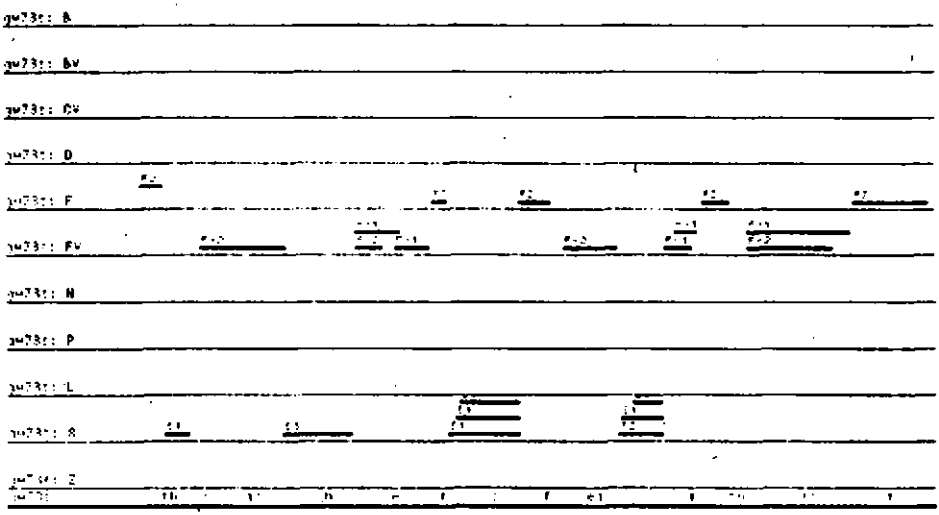

Figure 1: Sample output of the phonetic rule component

SYSTEM DESCRIPTION

Acoustic Parameters

The acoustic parameters of the current system were chosen on the basis of their usefulness in discriminating between different classes of sounds based on reports of existing feature-based speech recognizers [7, 8] or because of their known importance in human speech perception (see e.g., [9] for a survey). Several channel energies are used for making coarse category decisions such as discriminating between 'silence' (stop consonant closures) where very little energy is present except at very low frequencies, frication (fricatives or the aspiration following stop consonant release) where most of the energy in the waveform is at relatively high frequencies, and sonorant segments (vowels and the sonorant consonants) which have most of their energy in a mid-frequency region. Very low frequency energy is used for making voiced/voiceless decisions for some of the MidClasses.

Center frequencies, amplitudes, and bandwidths for the first 3 formants are estimated using an implementation of an algorithm developed by McCandless [10]. These parameters are included in the analysis files for their usefulness in detecting and classifying sonorant consonants and the vowels. An estimate of fundamental frequency [11] is included in the analysis package for its usefulness in detecting voiced vs. voiceless speech segments. Zero crossing rate and the normalized first autocorrelation coefficient are included as potential frication detectors, and frame

amplitude range as an energy measure which proves useful for fricative classification.

Threshold Rules

The strategy of progressively fine partitioning of the speech waveform is for the most part implemented in the TRules which detect and classify the acoustic features which the segment hypothesizing rules, the SRules, refer to. Three of these TRules, 'Silence', 'Frication', and 'Sonorance' are central to the strategy since many other rules refer to their output via logical operators. For example, the criterion for labelling a segment as 'silence' is that the energy in the signal above 300 Hz be below an empirically determined threshold. Once a speech segment has been labelled as 'silence', i.e., a stop consonant closure or pause, further TRules classify the silence as voiced or voiceless by applying a threshold test to a measure of very low frequency energy. This rule only applies to the feature segments created by the silence rule. Similarly, frication is detected by applying a threshold test to a low frequency/high frequency energy ratio and by a threshold test applied to the normalized first auto-correlation coefficient. Once detected as frication, the speech segment is subjected to further tests for voiced/voiceless and strong/weak classification. But the voicing measure for frication segments is not the same as that for stop closure or 'silence' segments. Sonorant segments are created wherever the input speech is not silence, not frication, and where either the pitch tracking algorithm had a non zero value or there was sufficient energy in a mid-frequency energy channel to indicate either a vowel, diphthong, or sonorant consonant. Once a sonorant feature segment has been created, additional rules look at rates of change in energy in the mid-frequency region, location of the center of spectral mass, and the output of the formant estimation algorithm to subdivide it into potential liquid, nasal, front and back vowel segments, and so forth.

Sequence Rules

Since it is the case that many of the phonemes of English are composed of sequences of acoustically very dissimilar events, the closure-burst-aspiration of the voiceless stops or the closure-frication of the affricates being the most obvious examples, a mechanism for concatenating the feature segments output by the TRules is needed in order to generate MidClass phoneme hypotheses. In the current rule set voiceless stops are hypothesized whenever a silence segment which is also voiceless is followed by a frication segment. But since it is often the case that vocal fold vibration persists into the closure phase of the phonologically 'voiceless' stops, there is also an SRule which allows a short period of voiced silence to precede the voiceless silence segment. Similarly, the SRules create voiced stop hypotheses from silence segments which also pass the 'voiced silence' threshold test but also allow a segment of voiceless silence to follow since, again, the phonologically voiced stops are not necessarily characterized by evident vocal fold vibration throughout the closure phase. The SRules which create voiced and voiceless fricative segments treat the voicing distinction in a similar fashion.

EVALUATION

As mentioned above, segmentation and labelling rules for all of the MidClass segments (except the glides /y, w/) were developed on a small set of sentences (16) from one speaker. No effort was spent on attempting to achieve 100% correct

MID-CLASS PHONETIC ANALYSIS

classification of the segments in this data base, however. Instead, the development sentences were used to extract threshold and feature segment sequence information that we expected would apply to any input utterance of similar phonological content by any male talker so long as the tempo (fairly slow) and style (normal reading style) were similar. Although the overall 'hit-rate' in terms of percent correct for the development data base at 63% is not particularly high, it appears that the modest success of the rules generalizes to the three sets of test data in a reasonably successful way. For Test Set 1, the second reading of the original speaker, the overall score is 57%. For Test Set 2, the same sentences read by a second male talker, the overall score is just as good at 58% and for Test Set 3, 16 different sentences read by the original talker the overall score is 49%. In order for a MidClass hypothesis to be counted as correct, the hypothesized segment was required to overlap the hand-placed segment boundaries by at least 50% after a normalizing procedure which ensured that all possible paths through the segment lattice contained abutting segments was applied.

The breakdown of the performance of the rules on the different mid-class categories for all 4 data sets is shown in Figure2.

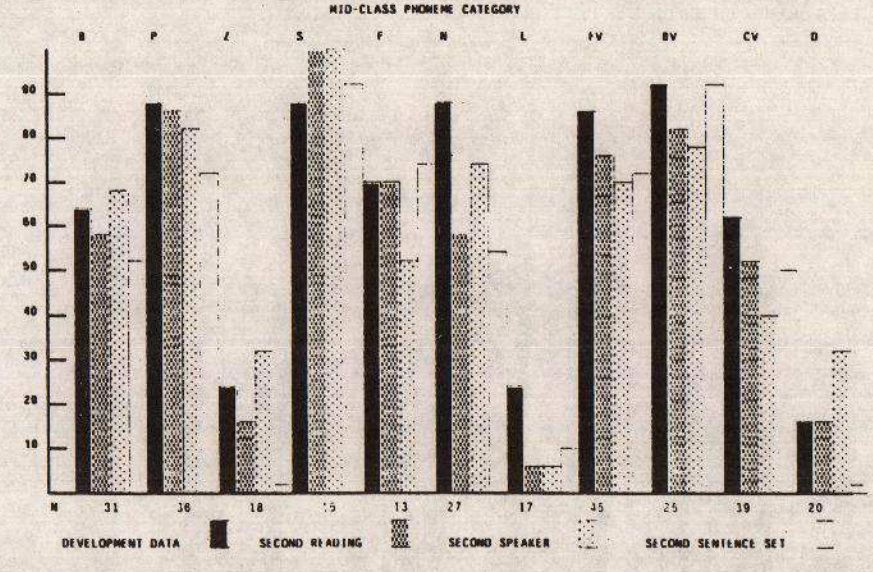

Figure 2: Correct Hypotheses/Number of Phonemes x 100, by Category

From the figure it is clear that for some of the MidClass segments the identification rates are fairly high while for others the scores are clearly unacceptable. For the rules that are working at all well, there is quite good consistency between the data sets in the display. For the voiceless stops (abbreviated 'P' in the display), strong voiceless fricatives (S), front vowels (FV), and back vowels (BV), correct hypotheses were generated by the rules for over 70% of the segments in all four data sets. Voiced stops (B), weakvoiceless fricatives (F), and nasals (N) show scores of at least 50% for

MID-CLASS PHONETIC ANALYSIS

all data sets and in some cases the scores are substantially higher. The figure also shows that the rules fail dramatically for some of the MidClass categories. Scores for strong voiced fricatives (Z), liquids (L), and diphthongs (D) are very bad indeed, and the scores for central vowels (CV) are poor. None of the weak voiced fricatives (/v, dh/) in the data were detected as fricatives (they were often labelled as either stops or as part of adjacent vowels).

These percentages show only the rate at which correct MidClass hypotheses were generated by the acoustic-phonetic rule set. The success of the rule component in generating accurate phoneme hypotheses must not be interpreted simply in these terms, however, since a trivial and useless way of scoring %100 on this measure would be to generate all of the possible MidClass labels for each segment that is detected. Since the acoustic-phonetic rule component generates the input to a lexical access component for which multiple phoneme hypotheses are acceptable (at the expense of processing time) so long as the correct one is present in the input lattice, a measure of the number of wrong segment hypotheses generated by the rules is needed. In Figure 3 we show the number of correct phoneme hypotheses expressed as a percentage of the total number of hypotheses generated by the rules for each of the MidClass categories and the four data sets.

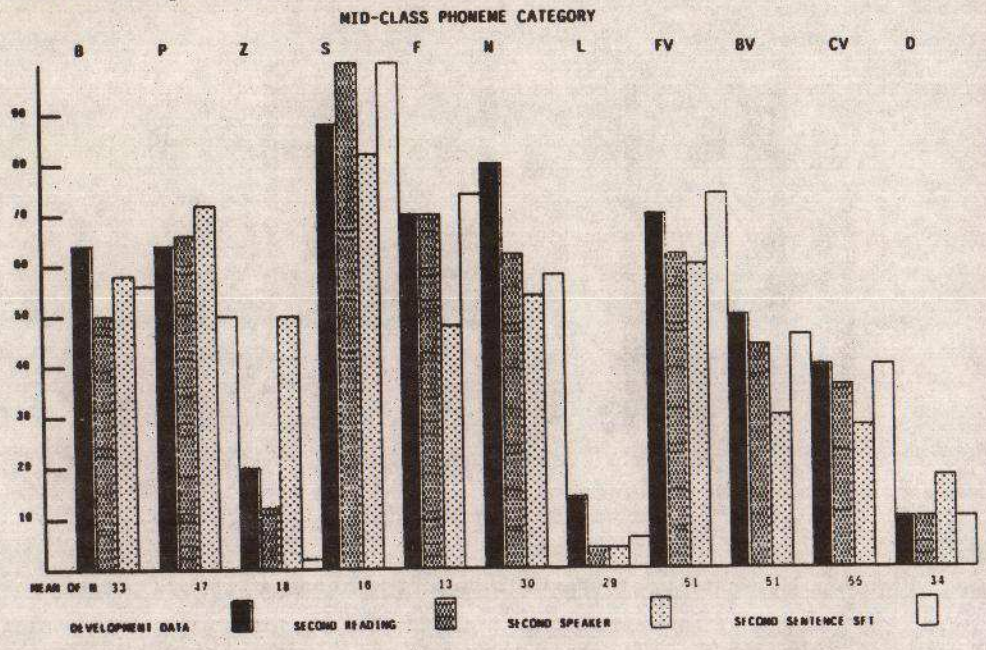

Figure 3: Correct Hypotheses/Total Hypotheses x 100 by Category

These data can be regarded as the 'confusion metric' for the system and the amount of overgeneration of segment hypotheses by the rule set can be estimated by comparing the scores in this figure with those in Figure 2.

DISCUSSION

MID-CLASS PHONETIC ANALYSIS

In spite of its obvious shortcomings, we believe that this prototype acoustic-phonetic rule set has confirmed the claim of Cole et al. [12] and others that feature-based phoneme recognition is possible, and that we now have a framework within which the development of a powerful speaker-adaptive acoustic-phonetic processor can be pursued. Some of the segment classes were identified with an accuracy which, given the simplicity of the rules, shows promise for generalizing to a larger data base and for modification toward a multi-speaker task.

No phonetician will be surprised by the distribution of the scores for the MidClass segments in our current system. The major weaknesses are just where one would expect them; for example, identification of liquids, glides, and diphthongs, where fine-grained accuracy of formant tracking is essential, is very poor, as are the scores for the voiced fricatives, where the varying proportion of voice/noise excitation of the vocal tract causes simple identification techniques such as the ones attempted in the current system to fail.

Nevertheless we have found the exercise of implementing the system a useful one since both its modest successes and obvious failures point the way forward for our research effort. Four areas which must be addressed in the next phase of our workhave been identified. The first is the development of a speaker adaptation component which can serve as a preprocessor to the acoustic-phonetic rule component. Most, if not all of the criterion values used in the rules must be speaker dependent and derived automatically from samples of speech collected at enrollment time. Temporal as well as spectral normalization procedures must be implemented in the adaptation phase. Second, computational mechanisms for implementing segmental and prosodic context sensitivity for the rules must be developed since it is clear that the acoustic information which specifies phoneme identity is encoded in units of at least syllable size [13]. Third, mechanisms for weighting evidence from multiple acoustic features must be implemented in the phonetic rules if they are to be robust in the face of the variability they must cope with. If powerful mechanisms for building in context insensitivity are not discovered and implemented, the elaboration of the rule set to account for each possible context will surely result in the generation of a segment lattice with so many hypotheses in it that even the most powerful parsing mechanism will bog down. Fourth, a great deal of effort must go into refining the set of acoustic features we extract from the waveform and into the development of better signal processing techniques for extracting them.

REFERENCES

[1] D.W. Shipman and V.W. Zue, 'Properties of large lexicons: implications for advanced isolated word recognition systems', IEE ICASSP, pp 546-549, (1982).

[2] D.P. Huttenlocher and V.W. Zue, 'A model of lexical access from partial phonetic information' IEE ICASSP, pp 26.4.1-26.4.4, (1984).

[3] A.M.Aull and V.W. Zue, 'Lexical stress determination and its application to large vocabulary speech recognition', IEEE ICASSP, pp 41.1.1-41.1.4, (1985).

[4] A.W.F. Huggins and R.S. Nickerson, 'Speech quality evaluation using "phoneme-specific" sentences', J. Acoust. Soc. Amer., vol 77 (5), pp. 1896-1906.

[5] G. Duncan, J. Dalby, and M.A. Jack, 'STAR-PAK: a signal processing package for acoustic-phonetic analysis of speech', Proceedings of the Institute of Acoustics, Vol8, (1986).

[6] A. Blokland, G. Watson, J. Dalby, and H. Thompson, 'SEGLAB: an interactive environment for phonetic segmentation and labelling of speech', Proceedings of the Institute of Acoustics, Vol8 (1986).

[7] C.J. Weinstein, S. McCandless, F. Mondshein, and V. Zue, 'A System for Acoustic-Phonetic Analysis of Continuous Speech', IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP 23 (1), (1975).

[8] M.F. Medress, 'The Sperry Univac System for Continuous Speech Recognition', in W.A. Lea, ed., Trends in Speech Recognition (Englewood Cliffs, NJ: Prentice Hall, 1980).

[9] M. Studdert-Kennedy, 'Speech perception' in N. J. Lass, ed., Contemporary Issues in Experimental Phonetics, pp 243-293, (New York: Academic Press, 1976).

[10] S. McCandless, 'An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra', IEEE Transactions on Acoustics, Speech, and Signal Processing, vol ASSP-22, pp. 135-141 (1974).

[11] W. Tucker, and R.H.T. Bates, 'A Pitch Estimation Algorithm for Speech and Music', IEEE Transactions on Acoustics, Speech, and Signal Processing, vol ASSP 26 (6), pp. 597-604 (1978).

[12] R.A. Cole, R.M.Stern, and M.J.Lasry, 'Performing fine phonetic distinctions: templates vs. features',in J.S. Perkell and D.H. Klatt, eds., Invariance and Variability in Speech Processes, (Hillsdale, New Jersey: Lawrence Erlbaum, 1986).

[13] A. Liberman, F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy, 'Perception of the Speech Code', Psychological Review, vol 74 (6), pp. 431-461 (1967).