SPEAKER INDEPENDENT CLASSIFICATION OF VOWELS AND DIPHTHONGS IN CONTINUOUS SPEECH*

Michael S. Phillips

Computer Science Department
Carnegie Mellon University
Pittsburgh, Pennsylvania 15221
USA

INTRODUCTION

When designing a vowel recognizer for continuous speech, one must consider not only the actual recognition algorithms but also the question of what categories the recognizer should attempt to identify. Should the vowel categories be at a phonemic level or a phonetic level [1] ? For a particular level of labeling, how detailed should the categories be? For example, if a phonetic level of labeling is chosen, which allophones should be grouped together? Another important problem is to find a consistent way of labeling speech at a particular level for training and testing the system.

This paper will describe the current design of the vowel recognizer that is under development and present classification results using two sets of vowel labels. The vowel recognizer is part of the acoustic phonetic recognition module of the CMU DARPA speech understanding system (Adams and Bisiani [?]). The system consists of a signal processing module, an acoustic phonetic recognition module, a word matcher, and a sentence parser. A block diagram of this system can be seen in Figure 1. The acoustic phonetic recognition module is given various representations of the speech signal as input and produces a network representing possible phonetic transcriptions of the speech signal. The network has nodes representing possible segment boundaries and arcs that have lists of labels with associated probabilities.

1. In this paper, the terms "phonemic level" and "phoneme" refer to the speaker's internal representation of the sounds in the lexicon. The term "phonetic level" refers to the actual sound present in the speech signal. For example, speakers may produce the word "children" such that the first vowel would be perceived as an [ah] if listeners were to base their perception only on the acoustic signal (taking into account acoustic context but ignoring expectations from lexical knowledge). This vowel will be considered to be an [ih] at the phonemic level and an [ah] at the phonetic level

2. The term "segment" is used here only to refer to a portion of the speech signal. It is not meant to imply that these are phonetic segments.

3. The term "feature" is being used here as in the pattern recognition literature. It is not intended to mean phonetic feature.

*Supported by grants from DARPA and NSF

THE VOWEL RECOGNIZER

The job of the vowel recognizer is to produce a list of probabilities of vowel labels given begin and end times for a segment [2] of the speech signal. In the system, these begin and end times are produced by the segmentation algorithms. In these classification experiments, the hand transcription boundaries are used as the segment begin and end times. The segment boundaries are not considered to be the boundaries of the relevant information about the vowel since important acoustic information about the identity of the vowel may be present in the vowel's surroundings. These begin and end times are only used to define the portion of speech that the recognizer is to classify.

The vowel recognizer consists of a set of feature [3] measurement algorithms to measure the acoustic properties of the vowel and a multi-dimensional classifier to produce the label probabilities. The set of feature measurement algorithms should capture all of the relevant acoustic information. The feature measurements for the vowels consist mainly of formant measurements at various points in time, formant changes throughout the segment, spectral centers of gravity measured at various points in time, duration, and average pitch of the segment. Many measurements (for example formant frequencies) are computed in more than one way and it is left to the classifier to decide which is the most reliable way to make a particular measurement for a particular decision. The complete list consists of about 100 feature measurements.

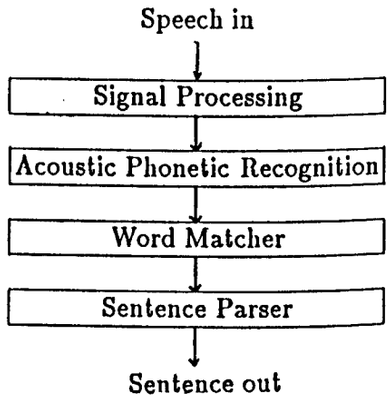

Figure 1: Block diagram of the CMU system.

Each segment is represented as a single point in a multi-dimensional space which has a dimension for each of the feature measurements. The job of the classifier is to give the probability of each of the vowel categories given a point in this space (Duda and Hart [2]). The classifier used consists of (n * (n-1))/2 pairwise classifiers where n is the number of vowel categories. The pairwise classifiers are used rather than a single classifier so that the number of dimensions can be as low as possible for each classifier. A single classifier would have to use all of the dimensions that were needed for all labels. Each pairwise classifier is able to use only the features needed for that particular pairwise decision. For example, this allows the system to use a formant tracker specifically designed for front vowels in the [iy] vs [ih] classifier, a formant tracker specifically designed for back vowels in the [aa] vs [ao] classifier, and some spectral centers of gravity for the [iy] vs [aa] classifier.

Each classifier is a two-class, n-dimensional Bayesian classifier where n is the number of feature measurements used for that particular pairwise decision. The classifier assumes a multi-variate Gaussian model of the data samples from each vowel category and given the feature measurements for a segment, assigns probabilities for each vowel category based on that model. Training consists of selecting the best features for each pairwise decision and estimating the parameters of the Gaussian model from a set of training data. Feature selection is done by performing a best-first search through all combinations of available features, using classification performance on a subset of the training data as the criterion for deciding which combination of features is best.

The vowel category probabilities from the pairwise classifiers are combined in the following manner: For each pairwise classifier, the vowel category with the highest probability is given a vote. The probability of the vowel category with the greatest number of votes is the average of the probabilities of that vowel category from each of the pairwise classifiers involving that vowel category. The probability for each other vowel category is the probability of that vowel category from the pairwise classifier for that vowel category versus the vowel category with the greatest number of votes. The probabilities are then normalized so that the sum of the probabilities for all vowel categories equals one.

VOWEL CATEGORIES

In recognition of vowels in continuous speech, the performance of the recognition system will be greatly affected by the choice of vowel categories that the recognizer is attempting to identify. Since the labels given to vowel segments in the training data define the recognizer's models of the vowel categories, the procedure used to obtain these labels is an important factor in determining the system's performance.

The goal of the vowel recognizer is that its decisions should be based on the same information that a human listener would use when making a decision about the vowel. Since the current design of the vowel recognizer works without any top down information from the higher levels of the system, this goal must be altered slightly: the vowel recognizer should use all the information that a human listener uses except any higher level language knowledge.

A listener's perception of a vowel in continuous speech is affected by many factors other than just the acoustic properties of the segment (Rudnickey and Cole [3], Jacob et al. [4]) The neighboring phonemes affect the acoustic realization of the vowel and the vowel affects the realization of the neighboring phonemes. Listeners take into account these local acoustic effects when making a judgement about the identity of the vowel. Listeners are also able to use information from a larger acoustic context (speaking rate, speaker characteristics, etc.) to make a judgement about the vowel. They will also be influenced by their expectation of what the vowel should be given their lexical and semantic knowledge.

The vowel recognition system should take into account as much of the acoustic information as possible to make a decision. This seems to mean that vowel categories on a phonetic level rather than a phonemic level should be used. The acoustic realization of a phoneme depends on higher level rules of the language. Since the vowel recognizer is only able to use acoustic information and not higher level information, it would not be able to map the varying acoustic realizations to the intended phoneme.

It's not clear how to obtain vowel labels at a phonetic level. If listeners are presented with enough of the signal to obtain the complete acoustic context, they will learn what words were spoken and be influenced by the expected phoneme. Alternatively, If short segments of speech are presented, the listeners will only be able to use the local acoustic context. Sentences composed of nonsense words could be used for training and testing the system. This would allow listeners to hear the entire utterance without being having any expectations of the intended vowels. The problem with this approach is that the speakers may have difficulty speaking the sentences in a natural manner.

Since the recognizer must map acoustic information

onto probabilities of vowel labels, the labels used for training and testing the system should have as consistent as possible a relationship to the acoustic information. There is ambiguity in phonetic labels (Church [5]). Even if listeners were able to hear vowels in their full acoustic context without being influenced by their phonemic expectation, they would not always agree. Effects such as listeners being influenced by their phonemic expectation or not being presented with the full acoustic context will increase the amount of ambiguity. Since the recognition performance of the system is limited by the amount of ambiguity in the vowel categories in the training and testing data, the labeling procedure should attempt to minimize the amount of additional ambiguity.

LABELING

Two labeling procedures were tried. In one, the people doing the labeling are able to hear the entire sentence and in the other, listeners are given the vowel segment with only a small amount of acoustic context. Both labeling procedures used the same set of labels. This list can be seen in the confusion matrices in Table 3.

The first set of labels used were the hand transcriptions being done for the DARPA speech project's acoustic phonetic database. These transcriptions are made by listening to the utterance, giving a phonetic transcription. running an automatic alignment program, and correcting alignment errors. Besides hearing the whole utterence, transcribersare able to see a spectrogram and other displays and are able to play any section of the utterance. The transcriptions are intended to be phonetic transcriptions but are biased towards the expected vowel in the cases where the realized vowel was ambiguous.

The second set of labels were produced by presenting trained listeners with each vowel segment in its local acoustic context. The segment boundaries were obtained from the hand transcriptions mentioned above. Each vowel segment was first played imbedded in the section of speech starting from the beginning of the transcribed segment before the vowel to the end of the segment after the vowel. After a half second pause, the vowel segment was played in isolation. The listener was able to have these two speech tokens played as many times as necessary. The listener then gave a phonetic label to the vowel with the option of responding with "not sure".

TESTS

The training data for these tests consists of 1000 utterances from 100 (30 female and 70 male) speakers from the DARPA acoustic phonetic database. All of the utterances were labeled both by doing the hand transcriptions and the labeling by listeners described above. The labeling by listeners was done by four listeners (each listener labeled a subset of the 1000 sentences).

	Testing labels Listener labels	Transcription labels
Training labels	(top1/top2/top3)	(top1/top2/top3)
Listener labels	48.3 / 68.7 / 79.3	40.3 / 60.8 / 72.4
Transcription labels	41.4 / 64.3 / 77.1	46.2 / 68.1 / 78.3

Table 1: System performance on the four combinations of training and testing labels. The numbers given are the percent agreement to the testing labels in the top choice, the top two choices, and the top 3 choices of the vowel recognizer.

The testing data consists of 160 utterances from 20 (6 female and 14 male) speakers. The testing speakers and utterances do not overlap with the training speakers and utterances. The testing data was labeled by the hand transcription and also by three listeners. Each listener labeled all 160 sentences so that listener versus listener agreement could be tested.

The system was trained on both types of labels and tested on both types of labels. The listeners gave a "not sure" label to 3.6% of the segments. These "not sure" labels were ignored during training. For testing the system on listeners' labels, the segments that were given the "not sure" label were automatically relabeled with the label from the hand transcriptions. The results of these tests can be seen in Table 1.

The labels obtained from the three listeners were compared to each other and also to the hand labels. Two comparisons were done: In one, only segments that were not given the "not sure" label by any of the listeners were used. In the other, all segments were used and "not sure" answers were considered to be errors. A summary of These results can be seen in Table 2. Confusion matrices for average listener versus listener agreement and for average listener versus hand label agreement can be seen in Table 3.

	Listener 1	Listener 2	Listener 3	Transcriptions
Listener 1	–	64.8/65.8	69.9/65.8	63.0/59.2
Listener 2	64.8/65.8	–	66.9/62.3	59.9/55.1
Listener 3	69.9/65.8	66.9/62.3	–	64.8/62.2

Average Listener versus Listener agreement 67.2/62.7
Average Listener versus transcription agreement 62.5/58.8

Table 2: This table shows the labeling agreement between all combinations of the three listeners and the hand transcriptions. For each entry, the first number is the percentage agreement considering only the segments that no listeners gave a "not sure" label. The second number is the percentage agree for all segment counting "not sure" labels as errors

DISCUSSION

From the results in Table 2, it can be seen that the listeners agree with each other 67% of the time and they agree with the hand labels 62% of the time. It seems that there is at least a small difference in these two type of labels. This difference may be due to convention differences between the two types of labels or it may be that the relationship between the acoustic information and the hand labels is less consistent for some distinctions than for the listener's labels. For example, it may be more difficult to make a judgement about the vowel color without being biased by phonemic expectation when listening to the entire utterance. It may also be true that the listening labels are less consistent than the hand labels for some other distinctions. For example, it is likely that it is more difficult to make a decision about vowel reduction in the listening labeling procedure since the listeners were not presented with the entire utterance and do not have access to information about speaking rate and the relative amplitudes of neighboring syllables.

A

	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns
aa	337	4	37	6	1	4	1	4	.	.	.	.	1	.	.	.	.	.	39	2
ae	4	215	3	.	.	3	36	.	.	8	.	.	1	.	.	.	.	.	.	3
ah	42	3	191	34	.	4	11	.	.	1	3	.	1	2	.	.	.	6	5	1
ax	12	.	93	140	8	1	6	1	.	7	39	.	.	1	.	.	.	19	4	26
axr	1	.	1	8	62	.	2	57	.	2	2	1	.	.	.	.	.	1	1	5
ay	17	5	8	1	.	165	.	.	11	1	4	1	.	.	.	.	.	.	.	2
eh	4	48	24	16	.	.	325	.	3	28	10	.	.	2	.	.	.	3	2	4
er	6	.	1	.	61	.	.	228	.	1	1	2	.	.	.	.	6	1	2	11
ey	.	4	.	.	.	8	10	.	323	14	2	17	.	.	1	.	.	.	.	7
ih	.	2	17	15	2	.	27	.	12	453	105	55	.	.	.	2	2	8	.	3
ix	.	.	5	57	4	.	11	1	6	151	202	20	.	1	1	2	2	13	.	22
iy	.	.	.	.	.	.	1	1	6	35	7	469	.	.	1	6	.	.	.	4
aw	13	6	2	.	.	1	.	.	.	.	.	.	35	.	.	.	.	1	.	.
ow	3	.	1	3	.	.	.	.	.	.	.	.	2	32	.	.	1	1	14	2
oy	.	.	.	.	.	.	.	.	1	.	.	.	.	.	22	.	.	.	1	.
ux	.	.	.	.	.	.	.	.	.	6	.	7	.	.	.	7	39	.	.	1
uw	.	.	.	.	.	.	.	.	.	4	.	1	.	2	.	42	78	3	.	.
uh	2	.	17	43	2	.	1	1	1	8	10	.	.	.	.	5	15	63	6	.
ao	30	.	3	1	.	.	.	1	.	1	.	.	2	4	1	.	1	2	126	.
ns	15	7	19	33	21	2	21	2	11	30	39	18	3	6	.	3	7	8	7	31
	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns

B

	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns
aa	231	2	19	3	1	1	.	2	.	.	.	.	2	1	.	.	.	.	10	1
ae	3	226	2	2	.	.	54	.	4	1	.	.	5	.	.	.	.	.	.	4
ah	33	.	194	34	1	1	10	.	.	.	1	.	.	3	.	.	.	13	7	6
ax	15	2	62	163	3	1	14	.	.	16	55	.	.	10	.	.	.	2	35	14	34
axr	6	.	5	14	98	.	4	95	.	3	10	3	.	.	.	.	.	.	6	3	29
ay	29	9	7	1	.	168	.	.	3	.	.	.	1	.	.	.	.	.	.	.	4
eh	2	22	18	3	.	3	264	3	4	24	5	.	.	1	.	.	.	1	2	11
er	6	.	2	5	35	.	5	196	.	6	.	.	.	1	.	3	5	1	3	15
ey	.	6	.	.	.	20	2	1	311	12	2	8	.	.	.	.	.	.	.	5
ih	.	5	9	14	2	.	20	1	10	316	45	19	1	1	.	1	1	10	1	3
ix	.	7	32	110	11	5	82	3	19	276	297	50	.	.	2	2	1	26	3	52
iy	.	2	.	.	.	2	.	3	27	58	30	475	.	.	1	6	.	2	.	15
aw	8	3	1	.	.	.	3	.	.	.	.	.	35	1	.	.	.	.	2	1
ow	4	.	3	3	.	.	.	.	.	2	.	.	3	33	.	.	1	3	18	2
oy	.	.	.	.	.	.	.	.	1	.	.	.	.	.	22	.	.	.	2	2
ux	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	36	65	6	.	2
uw	.	.	.	.	.	.	.	2	1	9	10	6	.	.	.	16	65	14	.	9
uh	.	.	.	.	.	1	1	.	2	4	.	.	2	.	.	.	.	33	.	3
ao	124	.	3	3	1	.	.	.	.	2	.	.	.	.	.	.	.	.	1	5
ns	.	.	6	1.	.	.	1	.	.	.	.	.	4	1	.	.	.	1	124	.
	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns

Table 3: Confusion matrices for average listener versus listener comparison (a) and average listener versus hand transcription (b). In (b) the row is the hand transcription label and the column is the listener label.

When trained and tested on the hand labels, the system's first choice accuracy is 46% and when trained and tested on the listeners' labels, the accuracy is 48%. A larger difference in performance can be seen comparing testing the system on the same type of labels as it was trained on versus testing the system on the other set of labels. Again this could either be explained by a convention difference or by a difference in the types of inconsistencies in the two labeling procedures.

It certainly seems that there is a large amount of ambiguity in the vowel categories being used for the vowel recognizer. The upper limit to the performance of the vowel recognizer is the amount of ambiguity present in the mapping from the acoustic information to the vowel labels. Since the listeners only agree with each other 65% of the time, this is the upper limit for the vowel recognizer performance if it is trained and tested on these labels. Obtaining labels that have a more consistent relationship to the acoustic information either by redefining the vowel categories or by developing a better labeling procedure should directly improve the performance of the vowel recognizer. From the system performance data, it seems that the two labeling procedures investigated so far have approximately equivalent amounts of ambiguity. If some distinctions are made more consistently with one procedure than the other, perhaps a better labeling procedure would combine the best aspects of both.

REFERENCES

[1] Adams, D. and Bisiani, R., The Carnegie-Mellon University Distributed Recognition System. Speech Technology, Mar/Apr 1986, 14-23.

[2] Duda, R. and Hart, P., Pattern Classification and Scene Analysis. John Wiley and Sons. 1973.

[3] Rudnicky, A. and Cole, R., Effect of Subsequent Context on Syllable Perception. Journal of Experimental Psychology, 4(4):638-647. 1978.

[4] Jacob, B. et al., The Effect of Language Familiarity on Vowel Discrimination. Presented at 100th meeting of the Acoustical Society of America, Los Angeles California, November 1980.

[5] Church, K., Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints. PhD. Thesis, MIT, 1983.