A New HMM-Based Approach to Broad Phonetic Classification of Speech

Jouni Pohjalainen

Laboratory of Acoustics and Audio Signal Processing
Helsinki University of Technology, P.O.Box 3000,
FIN-02015 HUT, Finland
jpohjala@cc.hut.fi

Abstract

A novel automatic method is introduced for classifying speech segments into broad phonetic categories using one or more hidden Markov models (HMMs) on long speech utterances. The general method is based on prior analysis of the acoustic features of speech and the properties of HMMs. Three example algorithms are implemented and applied to voiced-unvoiced-silence classification. The main advantages of the approach are that it does not require a separate training phase or training data, is adaptive, and that the classification results are automatically smoothed because of the Markov assumption of successive phonetic events. The method is especially applicable to speech recognition.

1. Introduction

The problem of automatically segmenting and classifying arbitrary speech input into a small number of basic acoustic-phonetic classes has been studied for a long time. The classes usually correspond to the type, or absence, of vocal tract excitation and can be specified e.g. in one of the following ways: speech-silence (voice activity detection); voiced-unvoiced; voiced-unvoiced-silence; voiced-unvoiced-mixed; voiced-unvoiced-mixed-silence. Applications for these classification problems arise in many areas of speech processing. They include limiting the search space in speech recognition [1], selecting a proper type of excitation in speech coding and synthesis, and speech analysis applications such as automatic phoneme boundary detection.

In this paper, a method is presented that can be applied to these classification problems. We focus especially on what is perhaps the most commonly adopted approach, namely the three-way voiced-unvoiced-silence (V-U-S) classification problem. However, by a proper selection of features the proposed general method is applicable to any of the mentioned tasks. The method is based on hidden Markov models (HMMs) and their iterative estimation. It requires no separate training phase, is not speaker dependent, is adaptive in the sense that the present conditions are "learnt" directly from the test speech input, and uses simple predefined rules in combination with some fundamental properties of HMM processing. Feature selection and determination of the classification rule in each feature space is crucial to the performance. In part, this task can be based on earlier published results [2].

A HMM is used to model a speech utterance, whose length is typically a few seconds. In basic form the method is non-causal. The HMM parameters are estimated and the inferred states are associated with phonetic categories (voiced, unvoiced, and silence, in this case). Thus, sequences of phonetic events are in effect modeled as a first-order Markov chain. This approach seems to exploit certain temporal aspects of speech effectively. The idea of using a single HMM to model utterance-length portions of continuous speech and identifying the states with phonemes or longer segments is not new in itself. Poritz [3] used a low-order linear predictive HMM operating on successive blocks of speech samples and presented qualitative results suggesting that the model was capable of discriminating between phonetic categories. Levinson [4] used a large pre-trained HMM, whose states corresponded to phonetic units, for speech recognition.

The adaptive V-U-S scheme of Bruno et al. [5] also employed a Markovian assumption for the classes, but was otherwise based on a statistical decision approach. Other early speech classification methods are also based on statistical decision [2] [6], while some more recently published solutions use neural networks [7][8]. All these methods use a pattern recognition approach and rely on sufficient amount of training. More general acoustic-phonetic properties of the sound classes are typically used only in the feature selection phase. In contrast, the proposed method is primarily based on prior knowledge of interrelations between speech feature vectors in different phonetic categories. This is combined with the basic properties of HMMs and the Baum-Welch reestimation procedure. Three examples of algorithms using this approach are presented and their classification performance is evaluated experimentally.

2. The general method

The main idea is to base the classification on prior knowledge of the relative acoustic properties of the speech sound categories in a given feature space. These relations should be as general as possible and not context-dependent. For example, unvoiced speech will normally have a larger proportion of high frequency energy than voiced speech, while both types will normally have larger energy than silence (or background noise) in a silent recording environment. These rules are applied in a short-time or local context. A similar philosophy has been applied e.g. in [9] where the decision thresholds of a rule-based digit recogniser are determined from the test input itself. The method proposed here does not use thresholds, however. The classification rules are incorporated in HMM signal processing and thus used only implicitly. Incorporation of the rules takes place primarily in the model parameter initialisation phase.

The steps common to all variations of the method are as follows:

1. *(Design of the classifier)* Decide the sound categories to be separated. Denote the number of categories by K.
2. Find a feature space where discrimination according to

step 1 is possible. Each individual category should be adequately separable from the rest by a binary decision using some subset of the features and a linear decision boundary. A simple approach is to try to find features along which the desired categories lie nearest to the endpoints of the range of values. This can be done by comparing measured distributions of the individual features. Denote the length of the feature vectors by M.

3. Define a K-state continuous density ergodic HMM [10] specified by the following set of parameters: a K-dimensional vector $\boldsymbol{\rho}$ of initial state probabilities; a $(K \times K)$ state transition probability matrix $\boldsymbol{P}$; M-dimensional state-specific mean vectors $\boldsymbol{m_i}$, $i = 1, \ldots, K$; a $(M \times M)$ covariance matrix $\boldsymbol{C}$ common to all states.
4. *(Classification)* Initialise the HMM state mean vectors $\boldsymbol{m_i}$, $i = 1, \ldots, K$, with the considerations of step 2 in mind. In particular, if the category corresponding to HMM state j is minimal (maximal) among the categories with respect to some feature, then the corresponding element in the vector $\boldsymbol{m_j}$ should be initialised with the minimum (maximum) value of this feature. However, it may not be practical to use absolute minima and maxima of the possible feature values. Instead, state mean initialisation is done in the classification phase when test speech input is available. The mean vector elements are then initialised with local statistics (minimum, maximum, mean) computed from the input. Next, initialise the Markov chain parameters $\boldsymbol{\rho}$ and $\boldsymbol{P}$ with uniform probabilities. Initialise the covariance matrix $\boldsymbol{C}$ with either the covariance matrix of the test data or one determined previously. The example algorithms use the former technique.
5. Estimate the HMM parameters for the speech input using Baum-Welch reestimation [10]. A slight modification of the conventional algorithm has been used, in which the computation of the so called forward and backward (joint) probabilities is replaced by alternative formulas that rely more on conditional probabilities [11]. In all tested cases, the reestimation converged rapidly to the final parameter values after just a few iterations.
6. Obtain the inferred HMM state sequence based on the estimated parameter values. This can be done by running the Viterbi algorithm [10].
7. Convert the HMM state segmentation into a segmentation in terms of the sound categories. This is straightforward if the states are assumed to have a one-to-one correspondence with the categories, as is the case here.

The motivation for using a HMM in this manner comes from various heuristic observations. As an illustration, the three-dimensional scatter plots in Fig. 1 show the relationships of three acoustic features - log energy, linear prediction (LPC) error, and zero-crossing rate - over two sentences spoken by different speakers. Three distinct clouds can be distinguished in both cases, albeit in somewhat different locations. Each cloud can be primarily associated with a particular type of excitation. Furthermore, each cloud lies closest to one or more bounding planes of the three-dimensional feature space (each feature has a theoretical or practical upper and lower bound). It would be nice to be able to automatically segment the signal in terms of discrete states corresponding to these concentrations.

From step 3, the HMM parameters to be estimated are $\boldsymbol{\rho}$, $\boldsymbol{P}$, $\boldsymbol{C}$, and $\boldsymbol{m_1}, \ldots, \boldsymbol{m_K}$. The parameter reestimation is an iterative algorithm, in which each iteration produces new estimates. As shown by Baum et al. [12], each iteration increases the likelihood of the observations, conditional on the parameter estimates, until a local maximum on the likelihood surface is found. Intuition and experience tell us that if the input sequence is long enough, the uniform initialisation of $\boldsymbol{\rho}$ has negligible effect on the final converged estimates regardless of the properties of the input. The same has been found to hold also for $\boldsymbol{P}$ [10]. Evidently, in most cases, these parameters can even be initialised with random values subject to the constraints of a first-order Markov process. Convergence due to the covariance matrix $\boldsymbol{C}$ has not been found to be a problem here either if it is initialised e.g. as suggested in step 4. Thus, the focus is on the convergence of the state mean vectors $\boldsymbol{m_1}, \ldots, \boldsymbol{m_K}$. They should converge on points close to their initial values, corresponding to a local likelihood maximum. Whether this maximum is a desired one depends on the speech input. Each sound category should be represented in the input; otherwise the state mean initialisation in step 4 forces a distinction where there should be none and the association of the states with sound categories may not be what is assumed. If each category is represented and the feature space is adequate for discriminating between the categories, then the conditional likelihood of the input should be increased primarily by moving the state means from near the boundary regions towards the closest actual concentrations (see Fig. 1). Elements of the state mean vectors are initialised with minima and maxima precisely to make the means approach the right concentrations from approximately correct directions and minimise the risk of confusing them with each other. Of course, the phonetic content of the input is usually not known in advance and it is not certain that the input contains segments from each category. Since the sound categories discussed here are very common, this has not been found to be a problem in most cases provided the speech input is long enough.

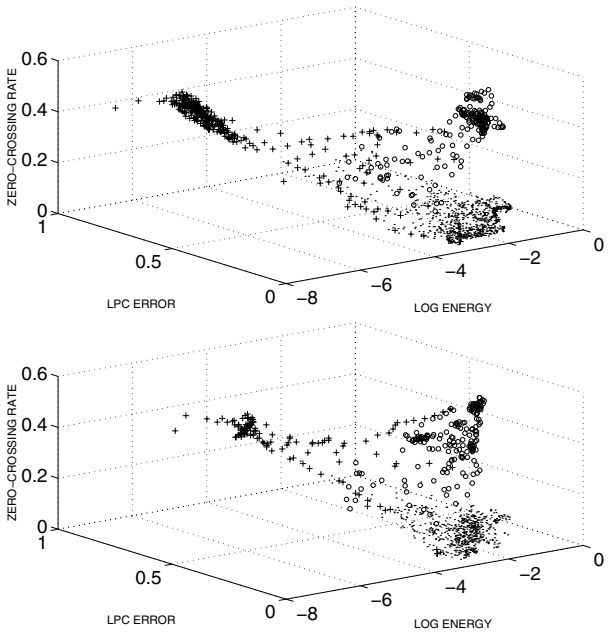

Figure 1: *Scatter plots of three features over two utterances. Phonetic category according to manual labeling is denoted by dots for voiced frames, circles for unvoiced frames, and plus signs for silent frames.*

3. Feature selection

Classification is based on a set of features that are easily computed and capable of V-U-S discrimination. The features represent some acoustic aspects of the speech signal. They are computed from speech signal frames so that each feature value only carries information from one frame. The following four features are used in the algorithms of this paper, computed from 25 ms frames with a frame shift interval of 3 ms:

1) Log energy of the Hamming windowed signal frame, f_1
2) Normalised linear prediction error energy using the autocorrelation method (with prediction order 24), f_2
3) Zero-crossing rate, f_3
4) Autocorrelation coefficient at unit sample delay, f_4

The features were selected in part by examination of their phoneme-specific distributions and also because the same or similar features have been widely used in speech classification [2] [5]-[8]. However, this particular selection is not the only possibility and statistical analysis of different features may reveal better solutions.

4. The algorithms

The V-U-S classification algorithms were allowed to use only the acoustic features f_1, f_2, f_3, and f_4. Each algorithm was allowed at most four reestimation iterations over the test speech input. Both the feature set and the iterations could be distributed between multiple HMMs as in algorithms 2 and 3.

4.1. Algorithm 1 (direct method)

This algorithm uses a single HMM with $K = 3$, $M = 4$. The states are initialised using statistics from the test input with $\boldsymbol{m_1} = (\max(f_1), \min(f_2), \min(f_3), \max(f_4))$ for voiced, $\boldsymbol{m_2} = (\max(f_1), \min(f_2), \max(f_3), \min(f_4))$ for unvoiced, and $\boldsymbol{m_3} = (\min(f_1), \max(f_2), \max(f_3), \min(f_4))$ for silence. After four reestimation iterations the most likely state sequence is found by the Viterbi algorithm and readily converted to a segmentation in terms of the phonetic categories.

4.2. Algorithm 2 (parallel method)

The algorithm uses two HMMs, both with $K = 2$, $M = 2$. The first model, denoted by A, uses features f_1 and f_2 and two reestimation iterations. Its states are initialised with $\boldsymbol{m_{A,1}} = (\max(f_1), \min(f_2))$ and $\boldsymbol{m_{A,2}} = (\min(f_1), \max(f_2))$. The second model, denoted by B, uses the remaining two features and two reestimation iterations. States are initialised with $\boldsymbol{m_{B,1}} = (\max(f_3), \min(f_4))$ and $\boldsymbol{m_{B,2}} = (\min(f_3), \max(f_4))$. The results are finally combined using the following rule: a segment is unvoiced if model A is in state 1 and model B is in state 1; voiced if model A is in state 1 and model B is in state 2; silent if model A is in state 2. Thus, model A is used for speech-silence classification and model B is used for voiced-unvoiced classification.

4.3. Algorithm 3 (hierarchical method)

This algorithm is very similar to algorithm 2 as it uses similar HMMs with the same features and initialisation. The only difference is that model B is not used on the complete speech input but instead on a concatenated version where the silence regions, found by model A, have been cut out.

5. Performance evaluation

The test material was two sets of 80 phonetically diverse Finnish sentences, each set read aloud by a different male speaker. The material was recorded with high-quality equipment in an anechoic room at a sampling rate of 22 kHz. Each utterance was manually segmented into phoneme-size units and labeled by a trained phonetician. The phonemic transcription was refined by distinguishing voiced and unvoiced /h/ and by separating the occlusion and burst segments of stop consonants. This transcription was converted to a V-U-S labeling, for use as performance evaluation reference, as follows. The voiced reference category included all vowels and voiced consonants. The unvoiced reference category included the fricative /s/ and the bursts of unvoiced stops. It was observed that the algorithms, using the features given, often misclassified low energy frication. Consequently, /f/ and unvoiced /h/ were assigned to the silence reference category together with the pauses. It should also be noted that being based on a phonemic representation of speech, the reference only approximately corresponds to an actual V-U-S segmentation. To verify the results, an independent manual true V-U-S segmentation was made for the material of one speaker.

Each of the 160 utterances was processed separately by algorithms 1, 2, and 3. Tables 1-3 show the combined category-specific frame classification results for both speakers. Only those frames were considered that, according to the reference, contained speech from just one category. The unvoiced category was the most difficult to identify correctly. This is mostly due to the mentioned inadequacy of the feature representation for correctly identifying low energy frication. Table 4 shows the overall percentage scores for each algorithm and speaker in different cases. Algorithm 2 gave the best classification accuracy. Because the scores with transition frames included were considerably lower than with only steady-state frames considered, the segment boundaries were obviously not always classified in accordance with a manual labeling made by a human. Fig. 2 shows the segmentation of one utterance using algorithm 1. It can be seen that the stop bursts can be variably classified to different categories and that the algorithm may find short additional segments near the category boundaries. These effects may still bear some phonetic relevance. The last two rows of Table 4 show overall scores using as reference the manual V-U-S labeling available for speaker 1. The scores are lower than with the phoneme-based reference. This happens mainly because the manual V-U-S labeling has a larger amount of low energy frication categorised as unvoiced speech.

The success of the classification also depended somewhat on the sentence. E.g. using algorithms 2 and 3, with speaker 1 and the manual reference, utterance-specific classification scores varied in the ranges 76 % - 100 % and 72 % - 96 % for steady-state frames and all frames, respectively. Inspection of utterances with score below 90 % showed that they were usually associated with a virtual absence of the unvoiced category and/or a combination of very short and very long segments belonging to some category.

Table 1: *Steady-state frame classification results, algorithm 1.*

		Classified as			
		V	U	S	Correct
According	V	98690	389	2354	97.30 %
to	U	847	9091	907	83.83 %
labeling	S	2786	967	24978	86.94 %

Table 2: *Steady-state frame classification results, algorithm 2.*

		Classified as			
		V	U	S	Correct
According to labeling	V	97763	315	3355	96.38 %
	U	487	9269	1089	85.47 %
	S	1081	671	26979	93.90 %

Table 3: *Steady-state frame classification results, algorithm 3.*

		Classified as			
		V	U	S	Correct
According to labeling	V	97552	526	3355	96.17 %
	U	484	9272	1089	85.50 %
	S	935	817	26979	93.90 %

6. Conclusions

A simple method capable of discriminating between predetermined phonetic categories was described and its performance was demonstrated with example implementations. Algorithm 2, where the set of four acoustic features is split between two independent binary classifications, provided the best results. Algorithm 3, using a two-stage hierarchical approach in which the voiced-unvoiced classification is made only for the concatenated non-silent segments from the first stage, also performed better than the direct three-way decision of algorithm 1, but apparently the Markov modeling was slightly less effective than in algorithm 2.

The main motivation for the development of the method is its potential use in phonemic speech recognition as a preliminary classifier that limits the search space by eliminating unlikely phoneme sequences. The main benefits of the method are that it does not require prior training and adapts well to the characteristics of the speaker and the recording environment. The performance depends very much on the selection of features chosen to represent the speech input. Feature selection affects the general structure, model definitions, and the interpretation of the classification results. Proper model initialisation is also very important as it implements the predetermined classification rules. Here, the tested features and classification rules provided good discrimination between voiced speech, silence, and high energy fricatives, but failed to discriminate between low energy frication and silence. The principle is applicable to other phonetic classification tasks if suitable feature representations are found and the models are defined and initialised appropriately.

7. Acknowledgements

This work was supported by Tekes (Finnish National Technology Agency) under the Usix STT project.

8. References

[1] O'Shaughnessy, D. and Tolba, H., "Towards a robust/fast speech recognition system using a voiced-unvoiced decision", in proc. ICASSP '99, vol.2, pp.413-416, 1999.

[2] Atal, B.S. and Rabiner, L.R., "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Trans. Acoustics, Speech, Signal Proc., 24(3):201-212, June 1976.

[3] Poritz, A.B., "Linear predictive hidden Markov models and the speech signal", in proc. ICASSP '82, pp. 1291-1294, 1982.

[4] Levinson, S.E., "Continuous speech recognition by means of acoustic-phonetic classification obtained from a hidden Markov model", in proc. ICASSP '87, pp. 93-96, 1987.

[5] Bruno, G., di Benedetto, M.D., di Benedetto, M.G., Gilio, A., and Mandarini, P., "A Bayesian-Adaptive Decision Method for the V/UV/S Classification of Segments of a Speech Signal", IEEE Trans. Acoustics, Speech, Signal Proc., 35(4):556-559, April 1987.

[6] Siegel, L.J. and Bessey, A.C., "Voiced/Unvoiced/Mixed Excitation Classification of Speech", IEEE Trans. Acoustics, Speech, Signal Proc., 30(3):451-460, June 1982.

[7] Ghiselli-Crippa, T. and El-Jaroudi, A., "A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech", in proc. ICASSP '91, pp. 441-444, 1991.

[8] Ahn, R. and Holmes, W.H., "Voiced-unvoiced-silence classification of speech using 2-stage neural networks with delayed decision input", in proc. Fourth Int. Symp. Signal Proc. Appl. (ISSPA), pp. 389-390, 1996.

[9] Rabiner, L.R. and Sambur, M.R., "Some Preliminary Experiments in the Recognition of Connected Digits", IEEE Trans. Acoustics, Speech, Signal Proc., vol. ASSP-24, pp. 170-182, April 1976.

[10] Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, 77(2):257-286, February 1989.

[11] Kim, C.-J., "Dynamic Linear Models with Markov-Switching", Journal of Econometrics, 60:1-22, 1994.

[12] Baum, L.E., Petrie, T., Soules, G., and Weiss, N., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", Annals of Mathematical Statistics, 41(1):164-171, 1970.

Table 4: *Overall classification percentages.*

	Speaker	Algorithm		
		1	2	3
Steady-state frames	1	93.83	94.59	94.51
	2	94.56	95.61	95.38
All frames	1	88.43	89.81	89.69
	2	88.13	89.68	89.41
Manual V-U-S, steady	1	93.33	94.03	93.90
Manual V-U-S, all	1	87.63	88.42	88.29

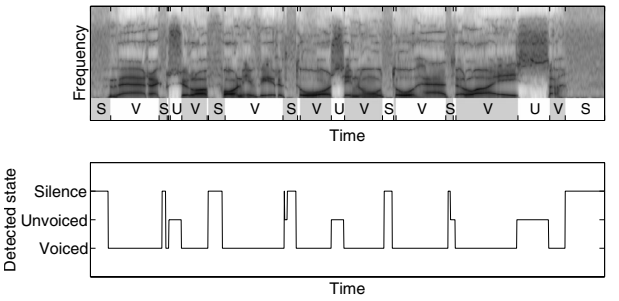

Figure 2: *Upper: Spectrogram of one utterance with V-U-S classification based on manual phonemic transcription. Lower: Automatic V-U-S classification using algorithm 1.*