STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **XLIX**, Number 2, 2004

SPEAKER INDEPENDENT PHONEME CLASSIFICATION IN CONTINUOUS SPEECH

MARGIT ANTAL

Abstract. This paper examines statistical models for phoneme classification. We compare the performance of our phoneme classification system using Gaussian mixture (GMM) phoneme models with systems using hidden Markov phoneme models (HMM). Measurements show that our model's performance is comparable with HMM models in context independent phoneme classification.

Key words: Phoneme classification, Gaussian mixture models, Continuous speech recognition, Unsupervised learning

1. Introduction

In order to build a continuous speech recognition system it is necessary to model sub-word units. It is impossible to model all the words even in a reduced size vocabulary. Word models will be formed from the concatenation of sub-word units. These units will be the phonemes. The phonemes are a set of base-forms for representing the sounds in a word. Replacing one phoneme in a word with another is usually enough to turn that word into a different word (or no word).

There are two general methods used for evaluating phoneme recognition systems, classification and recognition. In classification the segmentation is given and the goal is to find the most likely label for each segment of speech, given its beginning and end time. In the more general problem of recognition, on the other hand, both the labels and the segmentation are unknown.

This paper investigates the problem of phoneme classification. All the measurements were done on the DARPA TIMIT speech corpus which is a manually segmented speech database. The state-of-the-art technology in speech recognition is Hidden Markov Models [5]. Commercial speech recognition systems usually use HMM models for phonemes with continuous densities. The common model is a left-to-right HMM with three states (see Fig.1). The reason for using three states

Received by the editors: November 2004.

2000 *Mathematics Subject Classification.* 68T10, 65C40.

1998 *CR Categories and Descriptors.* I.5.1 [**Computing Methodologies**]: Pattern Recognition – *Models*; I.5.4 [**Computing Methodologies**]: Pattern Recognition – *Applications* .

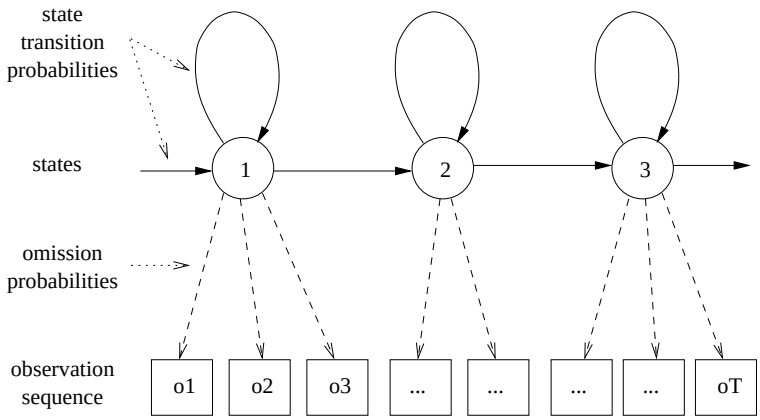

FIGURE 1. Three-state left-to-right phoneme model

is that the first part and last part of a phoneme are usually different from the middle due to co-articulation. This means that every state is responsible to model one third of a phoneme. Several authors [1] noticed that the transition probabilities have a negligible impact on the recognition accuracy and are often ignored. That is why in our system we used GMM for modeling phonemes. A GMM is a HMM with a single state, so there are no state transition probabilities, only observation emitting probabilities. Measurements show that our system's classification accuracy is close to systems using HMM for context-independent phoneme models and with Maximum-likelihood estimation of the parameters.

The structure of this paper is as follows. First we provide a short review of the phoneme classification problem and present the GMM modeling technique. Then we describe the acoustic features which were used. The final part of the paper discusses aspects of the experiments and the obtained results.

2. PHONEME CLASSIFICATION

2.1. **Phoneme Modeling.** Phoneme classification in continuous speech is a special pattern classification problem. It is easier than the phoneme recognition, because in classification phoneme boundaries are given. There are several approaches to pattern classification. Roughly we can divide these approaches into generative and discriminative modeling.

The HMM-GMM approach belongs to the generative models. Recently these models were extended and new discriminative training algorithms were proposed [6]. In the HMM-GMM approach each phoneme in the speech signal is given as a series of observation vectors $O = o_1,\ o_2, \ldots, o_T$, and each one has one model for each phoneme c. These models return a class-conditional likelihood $P(O|c)$.

The models are composed of states, and for each state we model the probability that a given observation belongs to this state. Time warping is handled by state transition probabilities, that is the probability that a certain state follows the given state. Supposing that the observations are independent (which is true only if we perform feature extraction in a very special way), the final probability for a sequence of observations can be computed using the forward algorithm [5].

HMM has several different forms like the discrete observation HMM and the continuous observation HMM. The discrete observation HMM is restricted to the production of a finite set of discrete observations. On the other hand in continuous observation HMM the observations are continuous and vector-valued. The usual way is to use a mixture of weighted Gaussian probability density function characterizing the distribution of observations within each state. The probability density function is given as

$$p(o_j) = \sum_{i=1}^{k} P_i \mathcal{N}(o_j, M_i, \Sigma_i) \tag{1}$$

where $\mathcal{N}(\ .\ , M_i, \Sigma_i)$ denotes the multidimensional normal distribution with mean M_i and covariance matrix Σ_i, k is the number of mixtures, and P_i are positive weighting factors which sum to 1. The D-dimensional normal density function has the form

$$\mathcal{N}(o_j, M_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^D det(\Sigma_i)}} e^{-\frac{1}{2}(o_j - M_i)^T \Sigma_i^{-1} (o_j - M_i)} \tag{2}$$

While HMMs are used for computing the class conditional likelihoods, discriminative models try to model the surfaces that separate the classes. For discriminative models one can use Artificial Neural Networks [7] (ANN) or a relatively new technology called Support Vector Machines [8]. A special ANN which was successfully applied for phoneme recognition is the Self-Organizing Map (SOM). This was introduced by Kohonen [9] for Finnish phoneme models and good results were also obtained for English vowels [10].

2.2. **Gaussian Mixture Models.** Our classification system was built for generative models. Instead of three states HMM we used only one state HMM which is a GMM. Our aim was to achieve the same recognition accuracy as for the classical three states HMM. We performed context independent classification (context dependent models perform better) on the DARPA TIMIT database and obtained slightly better classification accuracy than those using HMM on the same database.

Gaussian Mixture Models can be viewed as a special type of clustering. Clustering is an unsupervised classification and in the contributed papers appears as normal decomposition. Decomposition of a distribution into a finite number of normal distributions has been studied extensively. The parameters of normal distributions can be estimated using the method of moments or maximum likelihood

estimation. The maximum likelihood (ML) method is more reliable especially for high-dimensional cases. Having high dimensional feature vectors as data, we chose this one. Let us assume that $p(X)$ consists of k normal distributions as

$$p(X) = \sum_{i=1}^{k} P_i \cdot p_i(X) \tag{3}$$

where $p_i(X)$ is a normal distribution ($\mathcal{N}(X, M_i, \Sigma_i)$) with the expected vector M_i and covariance matrix Σ_i. Our problem simplifies to estimation of P_i, M_i, Σ_i ($i = 1 \dots k$) from the n available samples X_1, X_2, ..., X_n, drawn from $p(X)$. Our goal is to maximize $\prod_{j=1}^{n} p(X_j)$ with respect to P_i, M_i,and Σ_i under the constraint $\sum_{i=1}^{k} P_i = 1$. This optimization problem must be solved iteratively. One solution to this problem can be found by applying the maximum likelihood estimation technique (ML). Taking the logarithm of $\prod_{j=1}^{n} p(X_j)$, the criterion to be maximized becomes $\sum_{j=1}^{n} \ln p(X_j)$. Using the Lagrange multiplier's method the criterion to be maximized is

$$J = \sum_{j=1}^{n} \ln p(X_j) - \mu \left(\sum_{i=1}^{k} P_i - 1 \right), \tag{4}$$

where μ is a Lagrange multiplier. Computing the derivatives with respect to P_i, M_i and Σ_i and making them zeros we obtain the following formulas.

$$P_i = \frac{1}{n} \sum_{j=1}^{n} q_i(X_j) \tag{5}$$

$$M_i = \frac{1}{N_i} \sum_{j=1}^{n} q_i(X_j) X_j \tag{6}$$

$$\Sigma_i = \frac{1}{N_i} \sum_{j=1}^{n} q_i(X_j)(X_j - M_i)(X_j - M_i)^T, \tag{7}$$

where

$$q_i(X) = \frac{P_i \, p_i(X)}{\sum_{j=1}^{k} P_j \, p_j(X)} \tag{8}$$

is the a posteriori probability of X belonging to class i and satisfies $\sum_{i=1}^{k} q_i(X) = 1$. N_i is the number of samples belonging to class i.

The parameter estimation process can be described as follows.

Step 1. Choose an initial classification, $\Omega(0)$, and calculate P_i, M_i and Σ_i ($i = 1, \dots, k$).

Step 2. Having calculated $P_i^{(l)}$, $M_i^{(l)}$, and $\Sigma_i^{(l)}$, compute $P_i^{(l+1)}$, $M_i^{(l+1)}$, and $\Sigma_i^{(l+1)}$ by 5, 6 and 7 . The new $q_i^{(l+1)}(X)$ can be calculated as

$$q_i^{(l+1)}(X_j) = \frac{P_i^{(l)} \cdot p_i^{(l)}(X_j)}{\sum_{s=1}^{k} P_s^{(l)} p_s^{(l)}(X_j)} \tag{9}$$

Step 3. When $q_i^{(l+1)}(X_j) = q_i^{(l)}(X_j)$ for all $i = 1, \ldots k$ and $j = 1, \ldots n$, then stop. Otherwise, increase l by 1 and go to step (2)

In order to reduce the computation time we can force the covariance matrices to be diagonal. This assumption is justified since we can extract statistically uncorrelated features from speech. A diagonal covariance matrix allows expressing (2) as:

$$N(o^{(j)}, M^{(i)}, \Sigma^{(i)}) = \frac{1}{\sqrt{(2\pi)^D \prod_{k=1}^{D} (\sigma_k^{(i)})^2}} e^{-\frac{1}{2}\sum_{k=1}^{D} \frac{(o_k^{(j)} - M_k^{(i)})^2}{(\sigma_k^{(i)})^2}}$$

where $M^{(i)} = (M_1^{(i)},\ M_2^{(i)},\ \ldots, M_D^{(i)})$, $\sigma^{(i)} = (\sigma_1^{(i)},\ \sigma_2^{(i)},\ \ldots, \sigma_D^{(i)})$ and $o^{(j)} = (o_1^{(j)},\ o_2^{(j)}, \ldots, o_D^{(j)})$.

3. Experimental results

3.1. **The DARPA TIMIT speech corpus.** We used the TIMIT corpus for all experiments. This corpus was designed for training and testing continuous speech recognition systems. The database contains 6300 sentences, 10 sentences uttered by each of 630 speakers from 8 major dialect regions of the United States. The data were recorded at a sample rate of 16 KHz and a resolution of 16 bits. This corpus was manually segmented so the phonetic boundaries are given. The phoneme set is divided into the following 6 categories: vowels, stops, affricates, fricatives, nasals, semivowels (and glides). Other symbols are used for silence and closure intervals of stop consonants and affricates. There are 61 symbols used for labeling phonetic segments but most research papers present results on a reduced 39 symbol set. Table 1 presents the reduced 39 TIMIT symbol set.

3.2. **Features, training and testing.** We used for feature extraction 16 ms frame with 8 ms frame shift. We chose 16 ms for frame length because the phonemes belonging to the Stop category are usually very short. For example we obtained 17.93 ms average length for the b phoneme. This length was computed considering 915 occurrences of the phoneme in the speech corpus. From every frame we computed 12 mel frequency cepstrum coefficients[2, 3] and the energy of the frame. These coefficients do not incorporate any information about the way the signal changes over longer periods. However, it is well known that such information is essential in identifying sounds. In order to incorporate such

TABLE 1. TIMIT symbols

Category	Group	Category	Group	Category	Group
Vowel	ah, ax, axh	Vowel	ih, ix	Fricative	z
Vowel	iy	Semivowel	el, l	Fricative	f
Vowel	ih, ix	Semivowel	r	Fricative	th
Vowel	eh	Semivowel	w	Fricative	v
Vowel	ey	Semivowel	y	Fricative	dh
Vowel	ae	Semivowel	hh, hv	Stop	b
Vowel	aa, ao	Nasal	m, em	Stop	d
Vowel	aw	Nasal	n, en, nx	Stop	g
Vowel	ay	Nasal	ng, eng	Stop	p
Vowel	oy	Affricate	jh	Stop	t
Vowel	ow	Affricate	ch	Stop	k
Vowel	uw, ux	Fricative	s	Stop	dx
Vowel	axr, er	Fricative	sh, sz	Closure	epi,q,bcl, dcl,gcl, kcl,pcl,tcl,pau,h#

information we computed the first and second order derivatives which resulted in 39 features for a frame [4].

The 6300 utterances from TIMIT corpus were split into 4620 training utterances and 1680 testing ones. We performed three types of training. In the first type we used only 200 occurrences of every phoneme from the training set, in the second type we selected 400 occurrences and in the third one we used 1000 occurrences per phoneme. From these data we trained our phoneme models. For testing we used only the first 200 utterances from the standard 1680 set.

For every experiment we used the reduced 39-TIMIT phoneme set.

3.3. **Baseline GMM system.** In this experiment we fixed all parameters of the system except the number of mixtures. We used only diagonal covariance matrices in order to speed up the parameter estimations. Table 2 summarizes the overall classification accuracy obtained for different number of mixtures and for different numbers of phoneme occurrences used in training the models.

The best result for every type of training is in bold type. From this table we can conclude that for a fixed number of training data always exist an optimal model. If we use 200 occurrences per phonemes, the optimal model is formed by 32 normal densities. When we increased this number, there were not enough training data for accurate estimation of model parameters.

TABLE 2. Classification accuracies vs. number of Gaussians

Nr. of Gaussians	200 occ./phone	400 occ./phone	1000 occ./phone
1	41.56%	42.41%	44.68%
2	45.27%	47.77%	47.73%
4	48.54%	50.37%	51.20%
8	49.27%	52.34%	52.49%
16	51.98%	55.99%	55.68%
32	**53.55%**	57.10%	58.54%
64	52.68%	**57.77%**	58.89%
128		57.28%	60.16%
256			**60.43%**

TABLE 3. HMM

Paper	Frame size/shift	Features	Mixtures/state	Accuracy
[6]	32ms/10ms	MFCC-13+*Delta*	5	58.97%
[11]	25.6ms/unspec.	CC-12+*Delta*	8	57.10%
[12]	20ms/10ms	MFCC-18+*Delta*	unspec.	63.00%

3.4. **Baseline HMM systems.** Comparing our results with results obtained by other researchers is not an easy task, although the experiments are conducted on the same speech corpus. This difficulty is due to the incomplete presentation of the parameters of the experiments. In this section we try to summarize results obtained by others in the task of phoneme classification using HMM technology for phoneme modeling. We present results obtained by context independent phoneme models with measurements performed on TIMIT corpus. All the cited papers used 3 state hidden Markov phoneme models. Table 3 presents the results obtained by other research papers on the same task. It must be noted that [12] used full covariance matrices and the others (including our paper) diagonal covariance matrices for normal densities modeling.

More results can be found on the phoneme recognition topic, especially reported on the context dependent phoneme recognition task.

3.5. **Frames selection.** Several authors select only a few frames from the middle of each phoneme and use only these frames for training the models [9, 10].

TABLE 4. Classification accuracies vs. number of frames

Nr. of frames	Classification accuracy
3	53.07%
5	56.06%
7	56.70%
all	57.10%

Sometimes they use such a frame selection for simplifications, but it can be proved experimentally that the middle part of phonemes contain the real phoneme specific features. Our experiment was performed using as models 32 Gaussian mixtures and 200 occurrences per phonemes in training. Instead of using all frames belonging to the phoneme segment we used only a fixed number of frames and computed the recognition accuracy for the overall system. We ran this experiment for the following number of frames: 3, 5, 7 and all frames belonging to the phoneme. Table 4 summarizes the results.

3.6. **Intra-category classification.** The best classification accuracy for 400 occurrences per phoneme in training was obtained for 64 Gaussian models. We calculated the classification accuracies for the six phoneme categories. Using these models we computed the intra-category classification accuracy. Obviously the intra-category classification is higher than the all phonemes classification (see Table 5) because classification accuracy decreases with increasing the number of classes. One way to deal with this problem is to divide the entire phoneme set into phoneme categories by a category classifier and then to recognize phonemes in each category by a phoneme classifier. These categories contain all phonemes from Table 1 except silence, because this is the only category having only one phoneme group.

We computed the confusion matrix for 64 Gaussian models with 1000 occurrences per phoneme in training. The best classified phoneme was the vowel *oy* and the worst *uh*. Both are vowels. The most common errors are between symbols belonging to the same category (see Table 6).

4. CONCLUSIONS AND FUTURE WORK

The main conclusions of this paper are as follows. For a fixed number of training data always exist an optimal model. This was demonstrated experimentally. Our measurements show that using GMM as phoneme models one can reach slightly better classification accuracy than using three-states hidden Markov models. Experiments show that the middle part of the phoneme contains the most phoneme

TABLE 5. Intra category and all phonemes classification accuracies

Phoneme category	All phonemes	Intra category
Vowels	56.05%	60.70%
Nasals	48.63%	61.26%
Fricatives	68.80%	77.64%
Semivowels	59.91%	84.47%
Stops	51.19%	60.78%
Affricates	57.35%	72.80%

TABLE 6. The ten most common errors

Hand Label	Recognizer Label	Percentage of all errors
n, en, nx	ng, eng	2.7%
Closures	dx	2.6%
ih, ixx	ah,ax,ax-h	1.9%
b	g	1.9%
ih, ix	iy	1.6%
d	g	1.6%
ih, ix	ey	1.4%
r	axr, er	1.3%
Closures	dh	1.3%
Closures	g	1.3%

specific information. The confusion matrix demonstrates the viability of category classification scheme, because the most common errors are between symbols belonging to the same category.

In the future we would like to implement the category classification scheme and after that to transform our system into a phoneme recognition system. Another aim is to test our system using context dependent phoneme models and to improve the parameter estimation using other methods than the ML (Maximum Likelihood).

5. Acknowledgements

We are grateful to *Artificial Intelligence Research Group of the Hungarian University of Szeged*[1] for the possibility to use their speech database for these experiments.

References

[1] M. K. Omar, M. Hasegawna-Johnson, S. Levinson, "Gaussian Mixture Models of Phonetic Boundaries for Speech Recognition", Automatic Speech Recognition and Understanding Workshop, 2001.

[2] L.R. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", Prentice Hall, Englewood Cliffs, 1993.

[3] J.R. Deller, Jr. J. H.L. Hansen, J. G. Proakis, "Discrete-Time Processing of Speech Signals", John Wiley&Sons, 2000.

[4] X. Huang, A. Acero, H.-W. Hon, "Spoken Language Processing", Prentice Hall, 2001.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE, vol. 37, no. 2, pp. 257-86, February, 1989.

[6] R. Chengalvarayan, L. Deng, "Speech Trajectory Discrimination Using the Minimum Classification Error Learning", IEEE Transactions on Speech and audio Processing, Vol. 6, No. 6, pp. 505-515, Nov. 1998.

[7] C. M. Bishop, "Neural Networks for Pattern Recognition", Oxford University Press Inc., New-York, 1996.

[8] V. N. Vapnik, "Statistical Learning Theory", John Wiley & Sons Inc., 1998.

[9] T. Kohonen, "Self-Organizing Map", $3^r d$ edition, Springer, Berlin, 2001.

[10] N. Arous, N. Ellouze, "Cooperative supervised and unsupervised learning algorithm for phoneme recognition in continuous speech and speaker-independent context", Neurocomputing 51, pp. 225-235, 2003.

[11] B. Logan, P. Moreno, "Factorial HMMs for acoustic modeling", ICASSP, Vol. 2, pp. 813-816, 1998.

[12] R. Merwe, "Variations on Statistical Phoneme Recognition - a hybrid approach", master thesis, 1997.

Sapientia – Hungarian University of Transylvania, Faculty of Technological and Human Sciences, 540053 Târgu-Mureş, Romania

E-mail address: manyi@ms.sapientia.ro

[1] *Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged,* H-6720 Szeged, Aradi vertanuk tere 1., Hungary, http://www.inf.u-szeged.hu/speech