Proceedings of the Institute of Acoustics

PHONEME MATCHING TECHNIQUES FOR SPEECH IDENTIFICATION PROBLEMS.

M. J. Carey and E. S. Parris

Ensigma Ltd, Turing House, Station Road, Chepstow, Gwent, NP6 5PB, U.K.

mike@ensigma.com, eluned@ensigma.com

1. INTRODUCTION

Many workers are now working on the problem of large vocabulary continuous speech recognition, often attempting to transcribe text of speech derived from vocabularies in excess of 20,000 words. However, in this paper we show that several important problems in speech processing can be solved without transcribing the speech into text, or indeed without knowing the vocabulary from which the speech was derived. The sort of problem that can be solved is usually that of identification of some attribute of the speech. Examples of this are the speaker's identity, the language spoken, the gender of the speaker or the topic under discussion. The latter problem has been discussed elsewhere[1,2,3], and so in this paper we will consider the first three problems, that is, speaker, language and gender identification. In Section 2 we will describe the system we use to match the incoming speech to a set of sub-word models, and in Section 3 we present the theory which underlines our approach to identification. In Section 4 we present recent results we have achieved in each of the application areas under discussion.

2. MATCHING PHONEME SEQUENCES

The phoneme matching system comprises the first two stages shown in Figure 1. It takes as its input speech utterance and seeks to provide as its output phonetic transcription of the input utterance, together with acoustic likelihoods of the occurrence of each of the phonemes. Since we restrict our interest to telephone quality speech the speech is pre-filtered in the feature extractor to limit its value to 3.4 kHz, and then sampled at 8 k samples per second. Speech is then analysed using a nineteenth order mel scale filterbank which produces estimates of the spectrum of the speech signal at a 10 ms frame rate. A discrete cosine transformation is then used to decorrelate the outputs of the filterbank, and hence provide the mel cepstrum of the speech signal. The twelve cepstral coefficients are augmented with the corresponding twelve estimates of the trend of the cepstra, the delta cepstra, to give twenty four components. Estimates of the energy and its differential are added to this to give a twenty sixth order feature vector. This in its turn is subjected to a second transformation which is itself calculated by Linear Discriminant Analysis[4,5]. The 10 ms frames of linear discriminant features are then forwarded to the pattern matching algorithm.

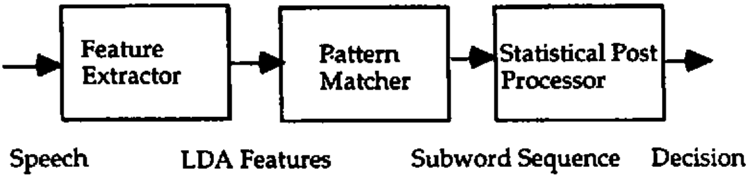

Figure 1. The Identification System

In the pattern matching algorithm, dynamic programming is used to find the sequence of sub-word Hidden Markov Models which best explains the input sequence of feature vectors. In our system each sub-word unit, which approximates to a phoneme, is represented by a three state left to right continuous density Hidden Markov Model. When the model is derived from speech from a single speaker the omission probability distribution is modelled by a multivariate unimodal Gaussian. However, when speaker independent models are required the omission probability distribution is modelled by a multivariate mixture density where the mixtures typically have five modes. We have also found it advantageous to build separate models for male and female speakers. Duration is modelled explicitly in the system by the gamma distribution, which associates a durational penalty with the final states of each model. The pattern matching process is constrained to match plausible sequences of models by a bigram grammar. The output of the pattern matcher is a time aligned estimate of the sequence of phonemes which occurred, together with an estimate of the likelihood of the sequence of acoustic vectors, given the model set. The performance of the system on real conversational speech is only a phoneme accuracy of 30 - 40%. However, we shall see in subsequent sections that when this is combined with the correct statistical post-processing the results of the identification tasks we are considering can be surprisingly good.

3. THEORY

3.1 Acoustic Matching

The subword matching system produces the probability of the sequence of acoustic feature vectors a given the most likely sequence of models S from set j, $p(a|S_j)$ However we require $p(a|S_j)$ the probability of the sequence of models given the acoustics. Bayes' theorem provides this, i.e.

$$p(S_j|a) = \frac{p(a|S_j)p(a)}{p(S_j)}.$$

Since all sequences over all sets are assumed equiprobable then

$$p(S_j|a) = p(a|S_j)p(a).$$

That is the output of the matching system is the probability of the models given the acoustics. weighted by the probability that we observed the acoustic sequence. We can apply this result to identification problems by matching the acoustics to different sets of models for example a male set and a female set of models for gender identification assigning the speaker of the utterance to the set with the greatest likelihood.

3.2 Phonetic Matching

The above analysis assumes that the utterance must match only models for one set or the other. However there may be some benefit in allowing the subword matching algorithm to choose the most appropriate models from all the sets available. The decision process is then to assign the utterance to the class from which the largest number of best matching models is drawn. However benefit can be derived from weighting the choices according to the discriminating ability of the models for the particular task. Assigning an appropriate weighting to each class can be carried out as follows. Initially viewing the problem as a simple decision between two classes Bayes rule is applied, that is if $p(C_r|M) > p(C_{\bar{r}}|M)$ then select class r else select class $\bar{r}$. Since $p(M|C_r)$ and $p(M|C_{\bar{r}})$ can be estimated Bayes' theorem gives

$$\frac{p(M|C_r)p(C_r)}{p(M)} > \frac{p(M|C_{\bar{r}})p(C_{\bar{r}})}{p(M)} \text{.......}(1)$$

where in this case, the message M consists of subword models $w_1 \ldots w_{N_m}$ from a vocabulary K, the most probable sequence as given by the Viterbi alignment. If the probabilities of subwords are independent $p(M|C_r) = p(w_1|C_r) \ldots . p(w_{N_m}|C_r)$ If each subword in the vocabulary occurs n_k times and $p(w_k|M) = n_k / N_m$ then

$$\log p(C_r|M) = N_m \sum_{k=1}^{K} p(w_k|M) \log p(w_k|C_r)$$

and similarly for $\log p(M|C_{\bar{r}})$. Furthermore if $p(C_r|M)$ and $p(C_{\bar{r}}|M)$ are equiprobable then the inequality of (1) becomes

$$N_m \sum_{k=1}^{K} p(w_k|M) \log \frac{p(w_k|C_r)}{p(w_k|C_{\bar{r}})} > 0$$

In the training set we assume that $p(w_k|C_r) = p(w_k|M)$. Hence the contributicn of each subword to the discrimination between the classes is given by

$$p(w_k|C_r)\log\frac{p(w_k|C_r)}{p(w_k|C_{\bar{r}})}..........(2)$$

We refer to this measure as the 'usefulness' of the subword. The discrimination contributed by each subword therefore depends not only on the relative frequency with which it is matched in true classes and other classes speech but also its absolute frequency of occurrence in the true classes speech. Not surprisingly subword models of frequently occurring phonemes are of more value in the decision process than those of infrequently occurring phonemes. This term (2) can be modified to take into account the imperfect recognition and false alarms

$\hat{p}(w_k|C_r) = p_r p(w_k|C_r) + p_f$ and $\hat{p}(w_k|C_{\bar{r}}) = p_r p(w_k|C_{\bar{r}}) + p_f$

where p_r is the recognition probability of subword k, and p_f is the false alarm probability of subword k.

The test for the occurrence of a single class is carried out by accumulating the likelihood ratio scores $\log\frac{p(w_k|C_r)}{p(w_k|C_{\bar{r}})}$ for each of the subwords as they occur in the test utterance M containing N_m subwords. Where we wish to discriminate between multiple classs this process is repeated for each of the classes' sets of models and the true class is deemed to be the one with the highest likelihood ratio score.

3.3.Combining Algorithms

Several techniques exist for combining different measures to achieve an overall score. We use logistic regression, a technique related to linear regression and LDA. It seeks to combine a set of attributes, these being the output of the algorithms, to produce a value lying between 0 and 1 which can be treated as a probability estimate for the classification task. This is achieved by linearly combining the attributes and passing the results through a sigmoid function.

$$q(\mathbf{x}) = \frac{1}{1 + \exp - \{a_o + \sum_{i=1}^{I} a_i x_i\}}$$

Where x_i are the measures to be combined and a_i are the weightings given to each measure and a_o is a threshold. If the distributions formed by the output variables of the algorithm is multivariate Gaussian then the optimal set of weights are given by $\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$ where $\mathbf{S}$ is the within class covariance matrix of the variables, and $\bar{\mathbf{x}}_0$ and $\bar{\mathbf{x}}_1$ are the mean vectors for each class. The threshold weight, a_o ,is given by $a_0 = -\mathbf{a}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_0)$

4. APPLICATIONS

4.1 Gender Identification

In this experiment in gender identification we matched the features from the unknown speaker's utterance to two sets of speaker independent subword models, $M_m$and M_w,built from the men and women speakers of the combined Oregon Graduate Institute Multiple Language Database described below. If in the traceback path the optimal sequence of models is through the female subword models then $p(M_m|a) < p(M_w|a)$ and we assume the speaker is a woman, otherwise the speaker isidentified as a man. The results of this experiment are shown in Table 1. While the speech files used in this experiment are typically 40s long the model set corresponding to the optimal state sequence is determined after one or two seconds of the utterance. While this technique works equally well in English, Spanish and German it is interesting to note that attempting to classify the gender of a German speaker using models based on the 'wrong' language, English, also works quite well.

Speech	Models	Correct
English	English	94%
Spanish	Spanish	95%
German	German	95%
German	English	89%

Table 1. Gender Identification Performance

4.2 Speaker Identification

In this experiment which is described more fully in [6] speaker dependent models were estimated for each of the ten speakers in the test set using two minutes of training speech. The system was then tested with each of the ten speakers 30-second test files, i.e., 100 test files in total. Each test consisted of running the subword matching system with a combination of models from a single speaker and a set of speaker independent models. This was repeated for all the speakers in the test set. Two methods of scoring the phoneme outputs were applied to decide the identity of the true speaker. The first method 'subword count' sums the number of times the speaker's subword models were matched instead of the speaker independent models for each speaker.The speaker with the highest count is deemed to be the true talker. The second method, as described in Section 3 weights each phoneme by the appropriate log likelihood ratio. The likelihood scores are accumulated for each speaker and the speaker with the highest score is declared as the true talker.

Speaker	Correct Identifications	Correct Identifications
	Phoneme Count	Likelihood Ratio
1	10	10
2	7	8
3	5	7
4	10	10
5	9	9
6	10	10
7	10	10
8	10	10
9	9	10
10	10	10
Total	90%	94%

Table 2 Speaker Identification Performance.

Table 2 shows the results achieved using the above two methods. It can be seen that by using the different discriminatory abilities of phonemes via their usefulness scores, gives an improvement in identification rate from 90% to 94%. Speaker identification of 64% can be achieved by just using the most useful phoneme `z' for discrimination. A rate of 90% is achieved by using the seven most useful phonemes z, i, e, m, v, ng and {. In contrast when the least useful phonemes are used first they givea markedly lower accuracy, 40%, for seven subword models.

4.3 Language Identification

The language identification experiment was carried out on part of the Oregon Graduate Institute eleven language database[7]. The data for each language is divided into three sections, a training set containing fifty speakers, a development test set with about twenty speakers and a final test set with another twenty speakers. Subword models for each language were built using the training set and the development test set was used to estimate the usefulness of each of the models. some models had zero or negative usefulness and were discounted from the scoring process. The development test set was also used to estimate the logistic regression coefficients. The results of

Table 3 shows that for a representative group of language pairs the phonetic scoring algorithm using the technique described in Section 3.2 outperformed the acoustic matching process of Section 3.1. However these results also show the benefits of combining the results of the two algorithms using logistic regression and compare well with those achieved by other workers e.g.[8].

Algorithm	English - German	English - German	English - Spanish	English - Spanish	English - Mandarin	English - Mandarin
	Dev Test	Final Test	Dev Test	Final Test	Dev Test	Final Test
Acoustics	86	84	86	89	89	84
Phonetics	92	92	94	92	91	84
Combined	97	95	97	94	94	95

Table 3. Language Identification Performance (%)

5. CONCLUSIONS

We have shown in this paper that subword matching techniques can be combined with statistical pattern processing to address a number of problems in speech research. Specifically they can be used in speaker, gender and language identification. The results achieved do not require word level transcription of the utterance which greatly reduces the computation required and obviates the need for a definition of the vocabulary of the utterances.

6. REFERENCES

[1]J. Baker et al. "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification", Proc ICASSP 93

[2]R. Rose et al. "Technique for Information Retrieval from Voice Messages", Proc ICASSP 91

[3] E. Parris and M. Carey, "Topic Spotting with Task Independent Models", submitted to ICASSP95

[4]M. Hunt et al. "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination", Proc ICASSP 91

[5]E. Parris and M. Carey, "Estimating Linear Discriminant Parameters for Continuous Density Hidden Markov Models", Proc ICSLP 94

[[6]E. Parris and M. Carey, "Discriminative Phonemes for Speaker Identification", Proc ICSLP 94

[7]Y. K. Muthusamy, R. A. Cole and B. T. Oshika. "The OGI Multi-Language Telephone Speech Corpus", Center for Spoken Language Understanding, OGI.

[8]M. A. Zissman and E. Singer. "Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modelling", Proc. ICASSP94, Adelaide, 1994.