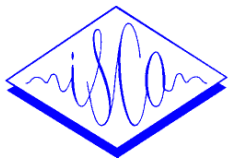
ISCA Archive
http://www.isca-speech.org/archive



4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

A SHARED-DISTRIBUTION APPROACH IN A HIDDEN MARKOV MODEL-BASED CONTINUOUS SPEECH RECOGNITION SYSTEM

Azarshid Farhat, Douglas O'Shaughnessy
e-mail: farhat@inrs-telecom.uquebec.ca
INRS-Télécommunications
16, Place du commerce, Île-des-Sœurs, Verdun,
Québec, H3E1H6
Canada

ABSTRACT

At the present time, one of the most important problem in large vocabulary continuous speech recognition is to achieve an optimum trade-off between acoustic models complexity and their trainability. In order to do so, we have defined a shared-distribution approach in our HMM-based continuous speech recognizer. In this clustering algorithm the distortion measure between two distributions is only based on the weights of gaussian mixture rather than all parameters of the distributions. Experimental results on the ATIS task show that our shared-distribution approach increased by 6% the word accuracy rate in comparison with our baseline system.

10.21437/Eurospeech.1995-368

1. INTRODUCTION

Over the past few years, hidden Markov models (HMM) have been successfully used in isolated word or continuous speech recognition. Even though word modelling leads to high performance in some applications, this approach becomes difficult for large or very large vocabulary. In this case subword units like phoneme or phone models are used. To take account of contextual variations of phonetic units, generally biphones or triphones substitute for context independent phonetic units. Due to their large number and the limited amount of training data, some triphone models are undertrained or even untrained. One of the most important challenges in statistical modelling is to estimate a very large number of parameters with a finite amount of training data. In other words, we shall achieve the optimum trade-off between triphone models complexity and their trainability.

At the present time, the most popular solutions used to resolve this dilemma are smoothing and sharing the parameters. The sharing can be achieved at the model level by clustering similar contextual models according to phonological knowledge [Wong-94] or by generalized-triphone models that are defined with an automatic merging algorithm [Lee-90]. We can also share the parameters at the distribution level or state level [Jouvet-94], [Hwang-93], [Digalakis-94], [Young-94].

A distribution-level or state-level clustering allows us to merge only the similar parts (or distributions) of models. Unlike model-level merging, with distribution-level sharing, the number of triphones remains the same; only the number of output densities is reduced. Similar distributions which are merged together can share the same set of training data.

[Jouvet-94] defined only one model per phone, named allophonic model, which integrates all contextual realizations of a sound into one complex unit. He used the context clustering trees to estimate and represent contextual parameters of the allophonic models.

[Hwang-93], [Digalakis-94] and [Young-94] used different clustering techniques for merging similar distributions. The clustered distributions or gaussian mixture constitute basic acoustic prototypes, named *senone* or *genone*.

Our distribution clustering approach has some features in common with different approaches

developped independently by [Hwang-93], [Digalakis-94] and [Young-94]. But in our approach, the distortion measure between two distributions is only based on the weights of gaussian mixtures rather than all the parameters of the distributions.

We have experimented with our approach on the INRS HMM-based, large vocabulary, speaker-independent, continuous speech recognition system [Kenny-92]. The shared-distribution model increased by 6% the word accuracy rate on the ATIS (Air Travel Information System) task in comparison with our baseline system.

2. SYSTEM OVERVIEW AND SPEECH DATABASE

In the INRS speech recognizer [Kenny-92], speech data are divided into temporal blocks. This block processing brings an efficient solution to both real-time issue and memoryless problems associated with classical forward-backward algorithm. The recognition process is based on a multi-pass search technique on current block. The results of each block are passed to the next one for a new multi-pass search until no more data is available. The acoustical vectors have 29 components. Every 10 ms, the acoustic analysis computes a set of 14 Mel Frequecy Cepstral Coefficients; 14 first order derivatives and energy derivative are computed using a window of 5 frames centered on the current frame. The acoustic-phonetic models used in both first and second passes are three-state left-to-right HMMs with no skip transitions. The output distributions were modelled by tied-mixtures. In our system we used 40 phones with a total of 1600 right context-dependent models. The total number of distributions is 4800. Language models used in our system are bigrams and trigrams.

The speech corpus used in these experiments came from ATIS task, with a vocabulary of 1087 words. 285 speakers with a total of 9269 sentences are used for the training. The tests are achieved with 33 other speakers with a total of 803 sentences. Male and female are present in both training and testing sets. Exeprimental results with shared-distribution approach are compared with the baseline system results: 82% of word accuracy when both coarse and fine language models are bigrams, and 86.2% of word accuracy when we use bigrams for coarse language model, and trigrams for fine language model.

3. SHARED-DISTRIBUTION APPROACH

Let two weight vectors α and β denoted as $\alpha : a_1, a_2, \ldots$ and $\beta : b_1, b_2, \ldots$

$$\Sigma_i a_i = A$$

and

$$\Sigma_i b_i = B$$

When two distributions are merged, the cross entropy of the resulting cluster is given by (1):

$$f_{\alpha+\beta} = -\Sigma_i (a_i+b_i/A+B) \log(a_i+b_i/A+B) \quad (1)$$

The cross entropy's variation when we merge α et β (weighted by their occurrence counts) is computed by:

$$\psi(\alpha,\beta) = (A+B)f_{\alpha+\beta} - Af_{\alpha} - Bf_{\beta} \quad (2)$$

ψ is the distortion measure used in our approach. Weighting the cross entropy by the occurrence count of each distribution let us take into account how each distribution is well-trained. Thus, distributions that occur rarely are merged before those well-trained. This distortion measure is only based on the weights of gaussian mixture. The most similar pair of clusters is the one who gives the least cross entropy's variation.

The clustering algorithm that we have defined is the generalized Lloyd algorithm, used also in vector quantization (LBG algorithm) [Linde-80]. We start with N clusters (N: numbre of distributions), each cluster contains only one

distribution. In each iteration we find the most similar pair of clusters and we merge them together. In order to improve the final clustering results, like the generalized triphone algorithm of K. F. Lee [Lee-90], after each clustering iteration we permit a number of element shifts from one cluster to another if each movement results an improvement (entropy reduction). The number of movements in each iteration is limited by a threshold of the entropy reduction and a maximum number of movement. The convergence criteria of the clustering algorithm are a threshold of the distortion measure and a minimum number of clusters.

The clustering algorithm is achieved on the contextual model's distributions of the same phone. Thus, we carried out 40 sets of clustering algorithm (1 set per phone). Before clustering, the total number of distributions per phone is 120. In each set of clustering process we reduced the final number of distributions approximatively by 40%.

The distribution-shared approach provides a good compromise between the trainability of the models and their complexity. With this technique and from a limited amount of training data, we can produce an acceptable set of well-trained distributions with a great number of contextual models.

4. COMPARATIVE RESULTS

We have experimented with our approach on the INRS HMM-based, large vocabulary, speaker-independent, continuous speech recognition system [Kenny-92]. In our experiments we used the ATIS spontaneous speech corpora.

The baseline system uses 1600 right-context dependent phonetic models with 4800 distributions. After the clustering the total number of distributions is reduced to 2927 (an average reduction of 40%).

TABLE 1. Comparative results obtained by shared-distributions approach with bigrams language model.

	word accuracy rate (%)
B0: baseline system (bigram)	82%
B_SH: distribution-sharing with bigrams	83.3%

TABLE 2. Comparative results obtained by shared-distributions approach with trigrams language model.

	word accuracy rate (%)
T0: trigram language models (baseline system)	86.2%
T_SH: distribution-sharing with trigrams	87%

Baseline accuracy with non-sharing distributions is about 82% and 86.2% with bigrams and trigrams language models, respectively. The preliminary results show a relative improvement about 6% by distribution-sharing approach (83.2% and 87% with bigrams and trigrams respectively).

TABLE 3. Comparison of word error's details for shared-distribution approach with bigrams language model.

	ins %	del %	sub %	word error %
B0	3.0	4.7	10.3	18
B_SH	2.8	4.0	9.9	16.7

Tables 3 and 4 show word error's details for shared-distribution approach in comparison with baseline system. For all types of errors

(insertion, deletion or substitution) we notice a relative reduction errors about 5% to 7%.

TABLE 4. Comparison of word error's details for shared-distribution approach with trigrams language model.

	ins %	del %	sub %	word error %
T0	2.8	3.1	7.9	13.8
T_SH	2.6	3.0	7.4	13.0

5. CONCLUSION AND FUTURE WORKS

The distribution-sharing approach is one of the most powerful solutions to the dilemma of complexity-trainability of the HMMs. Our clustering technique based on weight similarity provide very encouraging results.

With our shared-distribution approach we reduced by 40% the number of distributions in our contextual models (biphones) with a relative word accuracy increase of 6%.

In the near futur, we will use triphone models in our recognizer. In this case, the great number of distributions (192.000 in our system) will make shared-distribution approach essential.

At the present time, in the first and second pass of our system we use the same coarse and fine acoustic-phonetic models (right context-dependent models). In the futur, we would differentiate these coarse and fine acoustic models. To do so, we would use our present biphones in the first pass and the triphones in the second pass.

With the shared-distribution approach and in our multi-pass search technique, we will automatically define a high degree of tying among mixtures for the first pass (coarse acoustic-phonetic models) and a small degree of tying in other passes (fine acoustic phonetic models).

REFERENCES

[Digalakis-94] V. Digalakis, H. Murveit, "Genones: Optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer", *Proc. of International Conference on Acoustics Speech and Signal Processing*, pp. 537-547, Adelaide, April 1994.

[Hwang-93] M. Hwang, X. Huang, "Shared-Distribution Hidden Markov Models for Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 1, n° 4, pp. 414-420, October 1993.

[Jouvet-94] D. Jouvet, K. Bartkova, A. Stouff, "Structure of allophonic models and reliable estimation of the contextual parameters", *International Conference on Spoken Language Processing*, pp. 283-286, Yokohama, septembre 1994.

[Kenny-92] P. Kenny, R. Hollan, G. Boulianne, H. Garudadri, Y.M. Cheng, M. Lennig, D. O'Shaughnessy, "Experiments in Continuous Speech Recognition with a 60,000 Word vocabulary", *International Conference on Spoken Language Processing*, pp. 225-228, Banff, 1992.

[Lee-90] K.F. Lee, "Context-dependent phonetic hidden Markov models for continuous speech recognition", *IEEE Trans. on Acoustic Speech and Signal Processing*, pp. 599-609, April 1990.

[Linde-80] Y. Linde, A. Buzo, R.M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. on Communications*, pp. 84-95, January 1980.

[Wong-94] M.K. Wong, "Clustering triphones by phonological mapping", *International Conference on Spoken Language Processing*, pp. 1939-1942, Yokohama, septembre 1994.

[Young-94] S.J. Young, P.C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition", *Computer Speech and Language*, vol. 8, pp. 369-383, october 1994.