GMM-BASED ACOUSTIC MODELING FOR EMBEDDED SPEECH RECOGNITION

Christophe Lévy, Georges Linarès, Jean-François Bonastre

Laboratoire d'Informatique d'Avignon (France)

{christophe.levy, georges.linares, jean-francois.bonastre}@univ-avignon.fr

ABSTRACT

Speech recognition applications are known to require a significant amount of resources (training data, memory, computing power). However, the targeted context of this work - mobile phone embedded speech recognition system - only authorizes few KB of memory, few MIPS and usually small amount of training data.

In order to fit the resource constraints, an approach based on a semi-continuous HMM system using a GMM-based state-independent acoustic modeling is proposed in this paper. A transformation is computed and applied to the global GMM in order to obtain each of the HMM state-dependent probability density functions. This strategy aims at storing only the transformation function parameters for each state and authorizes to decrease the amount of computing power needed for the likelihood computation.

The proposed approach is evaluated on two tasks: a digit recognition task using the French corpus BDSON (which allows a Digit Error Rate of 2.5%) and a voice command task using French corpus VODIS (the Command Error Rate leads around 4.1%).

Index Terms: embedded speech recognition, acoustic modeling.

1. INTRODUCTION

The amount of services offered by the last generation of mobile phones has significantly increased compared to regular mobile phones. Nowadays, phones propose new kind of services like organizer, phone book, e-mail/fax, or games. During the same time, the mobile phone size has been largely reduced. Both these evolutions raise an important question: "How could we use a mobile phone with all its services without a large keyboard ?". Voice based human-to-computer interfaces supply a friendly solution to this problem but require to embed a speech recognizer into the mobile phone.

Since the last decade, performance of Automatic Speech Recognition (ASR) systems has been improved and nowadays authorizes to build efficient vocal human-to-computer interfaces. Moreover, even if scientific progresses could be noticed, the gain (in performance) remains linked to the computer resource: a last generation computer with a lot of memory is generally required. The main problem to embed ASR in a mobile phone is the low level of resource available in this context which classically consists of a 50/100 MHz processor, a 50/100 MHz DSP, and less than 100KB of memory.

State-of-the-art speech recognition systems are mainly related to statistical methods like Hidden Markov Model. For this kind of systems, a large training data set is required in order to reach good performance. The training data should be as close as possible to the targeted application. For mobile phone embedded speech processing, few speech corpora are available (moreover the main part of collected speech material is not directly recorded in a mobile phone which adds coding and transmission problems). In order to cope with this problem, the acoustic models are generally trained using large available corpora recorded in different conditions before being adapted to the targeted context using the limited amount of data available.

Mobile phone context involves a large environment variability as clients use their mobile phone in several locations (office, car, street,...). In order to improve the speech robustness in these adverse conditions, large acoustic models (trained with enough data) and/or acoustic-model adaptation are needed. Nevertheless, mobile phone resource constraints emphasize the need of new solutions.

In this paper, we mainly focus on the memory constraints even if the proposed solution allows a significant gain in terms of computational cost and reduces the needs in training data. Our approach consists in modeling the acoustic space by a unique GMM, which is derived for obtaining each HMM-state probability density function by applying a simple transformation (*cf.* figure 1). In this context, only the transformation parameters need to be stored for a given state. This approach also allows a gain in terms of computation power as a part of the likelihood computation is shared between the states. Our approach, firstly proposed in [1], is technically close to the semi-continuous HMM [2], or to [3, 4], the main difference concerns the view and the training of the acoustic model: as a GMM or as a HMM with tied components. Other approaches for embedding classical HMM recognizers in mobile phone are also present in the literature, like in [5].

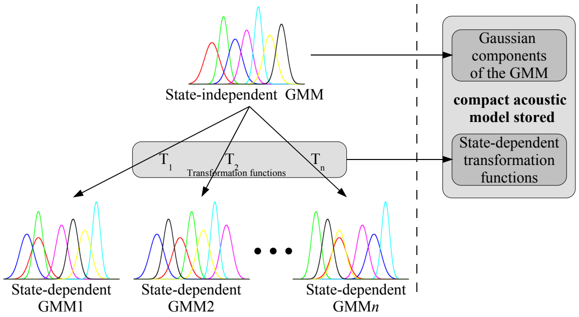

Fig. 1. *General schema of the proposed approach.*

In section 2, we present the corpora used for the experiments. Then, a baseline HMM system is presented in section 3. Section 4 describes the proposed approach. Section 5 shows some experimental results and finally, some conclusion and perspectives are provided in section 6.

2. CORPORA

In this paper, two databases are used to evaluate the proposed approach: BDSON and VODIS. A third database, BREF, is used only for the training of the state-independent GMM model. The databases are French corpora. BREF and BDSON are collected from clean acoustic environments and VODIS is recorded in a more realistic environment (records are made into a car).

2.1. BREF

BREF [6] is a large read-speech corpus composed of sentences selected from the French newspaper "*Le Monde*". This corpus contains about 100 hours of speech material from 120 speakers.

This corpus is used to train the baseline HMM system and to estimate the state-independent GMM (in both cases, the models are trained using BREF and then adapted to BDSON as explained in the next paragraph).

2.2. BDSON

BDSON [7] includes speech material (in French) recorded from 30 speakers (15 male and 15 female speakers).

BDSON was divided into two parts:

- one for the application-context adaptation (BADAPT_SET): it includes 700 digits pronounced by 7 speakers (4 male and 3 female speakers). This set is used to adapt the baseline HMM and the state-independent GMM to the application context. This phase is done once and the rest of the paper will refer to these adapted models;
- one for testing (BTEST_SET): composed of 2300 utterances of digits pronounced by 23 speakers (11 male and 12 female speakers).

The task targeted on this corpus is isolated word recognition, embedded in a mobile phone. Due to the database, the performance is evaluated thanks to a digit recognition task, where the digits are considered as words (*i.e.* no specific adaptation of the system is done, like reduction of the number of phoneme models). As the vocabulary is composed of the ten French digits, the performance measure is denoted in this paper Digit Error Rate (DER).

2.3. VODIS

VODIS [8] is a French corpus dedicated to car embedded applications. It includes recordings from 200 speakers. It contains a large variety of data: letters, digits, vocal commands, spelled words... Recordings are made with close-talk and far-talk microphones. The acoustic environment varies for every recording session (three cars, the window is opened or not, the radio is turned on or not, the AC is turned on or not). The Speech/Noise Ratio (SNR), estimated with the system presented in [9], is around 3.5dB[1].

We use only the subset containing the voice commands, under the close-talk condition. It was divided into two parts:

- one for the application context adaptation (VADAPT_SET): it includes 2712 commands pronounced by 39 speakers;
- one for testing (VTEST_SET): composed of 11136 utterances of commands pronounced by 160 speakers.

As we performed voice command recognition the evaluation measure used is the Command Error Rate (CER).

The speakers of BADAPT_SET and VADAPT_SET are different from the speakers of BTEST_SET and VTEST_SET (and are also different from the BREF speakers).

3. BASELINE HMM SYSTEM

The training process is composed of two successive stages. The models are firstly trained using the BREF corpus. Then, the second stage consists in adapting these models (weight, mean and variance) on xADAPT_SET[2] using the MAP approach [10].

The models are composed of 38 phonemic models and 108 emitting states (context-independent models). The number of Gaussian components per state is set to 128. 39 PLP coefficients per frame (13 static and first and second derivative) are used. The complexity of these models is about 1 million parameters.

Table 1 shows a DER about 0.96% when using the BDSON corpus and 1.80% with VODIS corpus.

Table 1. *Digit Error Rate (1(a)) and Command Error Rate (1(b)) for baseline HMM.*

(a) BDSON corpus: digit recognition task (2300 tests).

	DER	#parameters
128 gauss. (39 coef.)	0.96 %	1092k

(b) VODIS corpus: voice command recognition task (11136 tests).

	CER	#parameters
128 gauss. (39 coef.)	1.80 %	1092k

4. PROPOSED GMM-BASED APPROACH

The proposed approach consists in modeling the acoustic space using one unique GMM and then in deriving the state dependent Probability Density Functions (PDF) from it. The basic transformation function to obtain the state-dependent GMM is a MAP adaptation of the weight parameters following by a top-component selection: only the weights of the winning components are memorized for a given state-GMM (Weight Re-Estimation – WRE). An optional phase is also proposed, where the general GMM model is firstly adapted using the same transformation function for all the components (Unique Linear Transformation – ULT). Figure 2 presents the complete process.

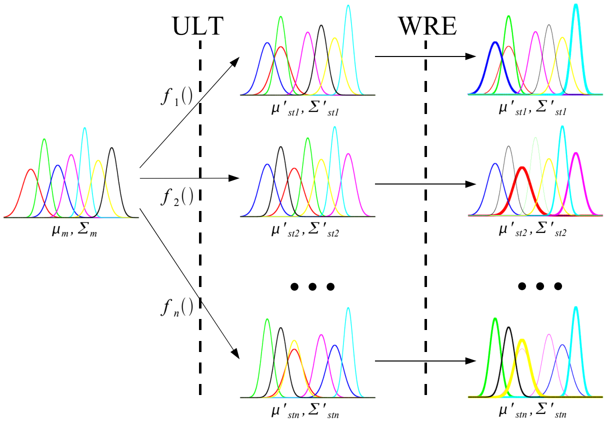
ULT
WRE
$f_1()$
$f_2()$
$f_n()$
μ_m, Σ_m
$\mu'_{st1}, \Sigma'_{st1}$
$\mu'_{st2}, \Sigma'_{st2}$
$\mu'_{stn}, \Sigma'_{stn}$
$\mu'_{st1}, \Sigma'_{st1}$
$\mu'_{st2}, \Sigma'_{st2}$
$\mu'_{stn}, \Sigma'_{stn}$
• • •
• • •

Fig. 2. *State-dependent transformation by applying ULT following by WRE.*

[1] for comparison the SNR of French-corpus BREF ([6]) is around 15dB.

[2] BADAPT_SET is used for digit recognition task and VADAPT_SET for voice command recognition task

4.1. Weight Re-Estimation (WRE)

This approach consists in estimating state-dependent weight vector from the initial GMM and an HMM-based frame alignment. Then, each state is represented by the state-independent GMM component set and by its specific weight vector. Two criteria are used for the weight re-estimation :

- Maximum Likelihood Estimation (MLE),
- Maximum Mutual Information Estimation (MMIE).

4.1.1. MLE

The Gaussian weights (w_i) are re-estimated using a MLE criterion defined by:

$$w'_i = \frac{w_i * L(fr|g_i)}{\sum_{g_j=1}^{nb_g} w_j * L(fr|g_j)} \qquad (1)$$

where w_x is the *a priori* weight of the x^{th} Gaussian component, and $L(fr|g_x)$ corresponds to the likelihood of frames fr for a given state g_x.

After this weight adaptation, only the $NBest$ Gaussian components are stored in order to decrease the memory occupation.

Furthermore, likelihood for each component of the global GMM is computed only once and then all the state likelihoods are easily computed thanks to a simple weighted combination of individual-component likelihoods.

4.1.2. MMIE

The HMM-training using a discriminant criterion (like MMIE) has largely been studied ([11]). The aim is to maximize the mutual information between the training word sequences and the observation sequences. The MMIE criterion increases the *a posteriori* probability of the word sequence corresponding to the training data given the training data.

For a training observation sequences $\{O_1, \ldots, O_R\}$ (of size R) with corresponding transcriptions w_r, the MMIE objective function to maximize is given by:

$$\mathcal{F}_{\text{MMIE}}(\lambda) = \sum_{r=1}^{R} log \left(\frac{P_\lambda(O_r|M_{w_r})P(w_r)}{\sum_{\tilde{w}} P_\lambda(O_r|M\tilde{w})} \right) \qquad (2)$$

where M_w is the model corresponding to the word sequence w and $P(w)$ is the linguistic probability. The denominator sums all possible word sequences $\hat{w}$ allowed in the task.

4.2. Unique Linear Transformation (ULT)

The method LIAMAP presented in [12] allows to adapt globally the initial GMM to a given state, using a unique and simple transformation. This transformation (applied both on the mean and the variance) is a linear adaptation:

$$\mu_{StateGMM} = \alpha * \mu_{GaussianCodebook} + \beta \qquad (3)$$

$$\Sigma_{StateGMM} = \alpha^2 * \Sigma_{GaussianCodebook} \qquad (4)$$

where α (which is common for $\mu_{StateGMM}$ and $\sigma_{StateGMM}$) and β are given as follows:

$$\alpha = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2} \qquad (5)$$

and

$$\beta = -\widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu + \widetilde{\mu} \qquad (6)$$

More details could be found in [13].

4.3. Evaluation of memory occupation

For each approach - HMM baseline system, WRE and WRE+ULT - we estimate the acoustic model sizes (in terms of number of parameters) and select the number of Gaussian components for each system in order to have the same number of parameters in the models. Two limits are used corresponding to realistic memory available in mobile phone.

For the largest models, the number of Gaussian components is limited to 282 components for the WRE method and to 174 components for the ULT+WRE approach. For the smallest models, the number of Gaussian components becomes 141 for the WRE approach and 35 for the ULT+WRE (for the global GMM). Acoustic vectors are composed of 12 static PLP coefficient added energy. Compared to the baseline system, we use only the static parameters (no first and/or second derivative).

5. RESULTS

Experiments on isolated-digit recognition task were carried out with BDSON corpus while others are performed with VODIS for voice command task.

In order to fit the memory constraint, only the $NBest$ top-components are stored. This selection is dynamically made state by state (the same total number of parameters is preserved, the average number of components is set to 20 for smallest model and to 30 for the largest).

Regarding table 2(a), with the smallest acoustic model, the DER varies between 3.22% (ULT+WRE/MMIE) and 4.65% (WRE/MMIE). The DER increases largely while the acoustic model size is reduced from 1092k parameters to 6k parameters. With the largest model the DER is 2.52% (WRE/MMIE) while the acoustic model size is divided by a factor 100.

Table 2(b) shows the results for voice command recognition task with a noisy corpus. The CER is between 4.14% (largest model and ULT+WRE/MLE) and 6.09% (smallest model and WRE/MLE). Compared to the 1.80% for the HMM-system without constraint, it represents an increase but the acoustic model size reduced by a factor between 100 and 200.

ULT+WRE system shows an improvement compared to WRE alone for both the model sizes (which correspond to the targeted application). Nevertheless, WRE allows a large decrease in terms of computational resource compared to ULT system (for ULT approach, the component likelihoods are computed for each state).

6. CONCLUSION AND PERSPECTIVES

In this paper, a solution was proposed for embedding automatic speech recognition in a mobile phone. The proposed technique is based on a HMM with a global, state-independent, GMM modeling of the acoustic space and a set of transformation functions able to adapt this GMM to obtain each state-dependent probability density function. This approach reduces drastically the memory size of the models and the computation time needed to compute the likelihoods (even if, in this paper, the focus was mainly put on the memory occupation).

Two different techniques were proposed for the transformation functions. The first one, WRE, consists in adapting by MAP the state-independent GMM and in selecting and storing only the top-component weights. WRE allows to save the memory without reducing the global GMM size and to reduce the likelihood computation time since the component likelihoods are calculated only once per frame. The second one, ULT+WRE, transforms the mean and

Table 2. *Digit Error Rate / Command Error Rate for WRE/MLE, WRE/MMIE, ULT+WRE/MLE and ULT+WRE/MMIE.*

(a) BDSON corpus: digit recognition task (2300 tests).

Model size	WRE/MLE	WRE/MMIE	ULT+WRE/MLE	ULT+WRE/MMIE
6k	4.17 %	4.65 %	3.39 %	3.22 %
11k	3.09 %	2.52 %	3.00 %	2.70 %

(b) VODIS corpus: voice command recognition task (11136 tests).

Model size	WRE/MLE	WRE/MMIE	ULT+WRE/MLE	ULT+WRE/MMIE
6k	6.09 %	6.05 %	5.01 %	5.10 %
11k	5.03 %	5.35 %	4.14 %	4.40 %

variance parameters before applying WRE. It uses a unique linear transformation for all the components. ULT+WRE allows a better modeling of state-dependent PDFs (and adds a cost in terms of computing time).

The method presented in this paper allows to embbed speech recognition system into mobile phone. For voice command recognition task a CER of 4.14% (*cf.* table 2(b)) when the ULT+WRE/MLE method is applied. A system without memory constraint obtains a CER of 1.80%, but requires 100 times more memory. For digit recognition task, the best DER obtain is 2.52% (*cf.* table 2(a)) with WRE/MMIE approach. A system without constraint allows a DER less than 1% but the acoustic model is composed of more than 1 million parameters while our model is about 11k parameters.

In this work, no adaptation was done, on the speaker or on the environment. Thanks to the structure of our models, an interesting way for these adaptations could be to adapt only the state-independent GMM, assuming that the state-dependent transformations could remain unchanged when the state-independent model is adapted. A possible strategy for this adaptation could be to use the test data to adapt the common GMM parameters (or a part of the parameter set) before decoding. Even if few frames are available, only one model has to be adapted instead of one model per state for a classical HMM system. Furthermore, this approach doesn't need a decoding before the adaptation since all the frames are related to only one state. This characteristic suppresses the influence of decoding errors during the adaptation step and the additional computing cost remains very low. These adaptations schemes seem very promising and we will focus on in further work.

7. REFERENCES

[1] C. Lévy, G. Linarès, and J.F. Bonastre, "Mobile phone embedded digit-recognition," in *Workshop on DSP in Mobile and Vehicular Systems*, Sesimbra, Portugal, September 2005.

[2] X. D. Huang, K.F. Lee, and H. Hon, "On Semi-Continuous Hidden Markov Modeling," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1990)*, Albuquerque, New Mexico, USA, April 1990, pp. 689–692.

[3] S.J. Young, "The general use of tying in phoneme-based HMM speech recognisers," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1992)*, San Francisco, California, USA, March 1992, pp. 569–572.

[4] J. Park and H. Ko, "Compact acoustic model for embedded implementation," in *Proceedings of International Conference on Spoken Language Processing (ICSLP'2004)*, Jeju Island, Korea, October 2004, pp. 693–696.

[5] S. Astrov, J.G. Bauer, and S. Stan, "High performance speaker and vocabulary independent ASR technology for mobile phones," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2003)*, Hong Kong, April 2003, pp. 281–284.

[6] L.F. Lamel, J.L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'1991)*, Gênes, Italie, September 1991, pp. 505–508.

[7] R. Carré, R. Descout, M. Eskénazi, J. Mariani, and M. Rossi, "The French language database: defining, planning and recording a large database," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1984)*, San Diego, California, USA, March 1984, pp. 324–327.

[8] P. Geutner, L. Arevalo, and J. Breuninger, "VODIS - voice-operated driver information systems: a usability study on advanced speech technologies for car environments," in *Proceedings of International Conference on Spoken Language Processing (ICSLP'2000)*, Beijing, China, October 2000, pp. 378–382.

[9] H.G. Hirsh, "Estimation of noise spectrum and its applications to SNR-estimation and speech enhancement," Tech. Rep. Technical report tr-93-012, 1993.

[10] J.L. Gauvain and C.H. Lee, "Maximum A Posteriori estimation for multivariate gaussian mixture observations of Markov chains," in *IEEE Transactions on Speech and Audio Processing*, April 1994, vol. 2-2, pp. 291–298.

[11] L.R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model parameters for speech recognition," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, Tokyo, Japan, April 1986, pp. 49–52.

[12] D. Matrouf, O. Bellot, P. Nocera, Linarès, and J. F. Bonastre, "Structural linear model-space transformations for speaker adaptation," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'2003)*, Geneva, Switzerland, September 2003, pp. 1625–1628.

[13] C. Lévy, G. Linarès, P. Nocera, and J.F. Bonastre, *Embedded mobile phone digit-recognition*, chapter 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2, Springer Science, H. Abut, J.H.L. Hansen and K. Takeda edition, 2006.