Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 806186, 12 pages
doi:10.1155/2009/806186

Research Article

Compact Acoustic Models for Embedded Speech Recognition

Christophe Lévy, Georges Linarès, and Jean-François Bonastre

339 Chemin des Meinajaries, 84911 Avignon Cedex 9, France

Correspondence should be addressed to Christophe Lévy, christophe.levy@univ-avignon.fr

Received 12 March 2009; Revised 8 July 2009; Accepted 20 October 2009

Recommended by Joe Picone

Speech recognition applications are known to require a significant amount of resources. However, embedded speech recognition only authorizes few KB of memory, few MIPS, and small amount of training data. In order to fit the resource constraints of embedded applications, an approach based on a semicontinuous HMM system using state-independent acoustic modelling is proposed. A transformation is computed and applied to the global model in order to obtain each HMM state-dependent probability density functions, authorizing to store only the transformation parameters. This approach is evaluated on two tasks: digit and voice-command recognition. A fast adaptation technique of acoustic models is also proposed. In order to significantly reduce computational costs, the adaptation is performed only on the global model (using related speaker recognition adaptation techniques) with no need for state-dependent data. The whole approach results in a relative gain of more than 20% compared to a basic HMM-based system fitting the constraints.

1. Introduction

The amount and the diversity of services offered by the latest generation of mobile phones (and similar embedded devices) has increased significantly during the last decade, and these new services are considered as crucial points by the manufacturers in terms of both functionalities and marketing impact. At the same time, the size of such devices has been reduced considerably, limiting the usability of the most complex services that could be embedded. Moreover, the use of hands and/or eyes is sometimes required by classical input mechanisms, forbidding the use of a mobile device when the attention should be focused on other activities. Voice-based interfaces provide a friendly human-computer interaction medium in mobile environments, freeing hands and allowing a rich interactivity between human and compact devices.

Embedded speech processing has been largely investigated in the two last decades, both on industrial and research aspects. The major difficulties faced in an embedded implementation are caused by the limitations in the hardware-resources available, and by the variability of the contexts where the system may operate. This last issue has been tackled in the more general framework of automatic speech recognition (ASR) system robustness; most of the proposed methods operate at the signal level or at the acoustic model level. Front-end based techniques focus on the noise-reduction problem, by performing echo cancellation, noise substraction, and so forth. At the model level, the acoustic variability is considered as a more general issue, including but not limited to environmental noise, speaker variability, and speech style diversity (spontaneous and/or interactive speech). Most of the recent advances in acoustic modelling rely on the integration of sophisticated techniques such as discriminative training, vocal tract normalization, or multiple system combination. Nevertheless, the relevance of training corpora remains a key point for the accuracy of the acoustic models, and recent state-of-the-art systems generally use huge amounts of materials for acoustic training. DARPA evaluations demonstrated the efficiency of these approaches for Large Vocabulary Continuous Speech Recognition (LVCSR).

Although significant improvements can be made through use of relevant training corpora, it cannot be expected that the varying environment of a mobile device can be fully modelled by any closed corpus. A further consequence of the extensive approaches for acoustic modelling is the increase in computing resource requirements, especially memory footprint: classical LVCSR systems rely typically on acoustic

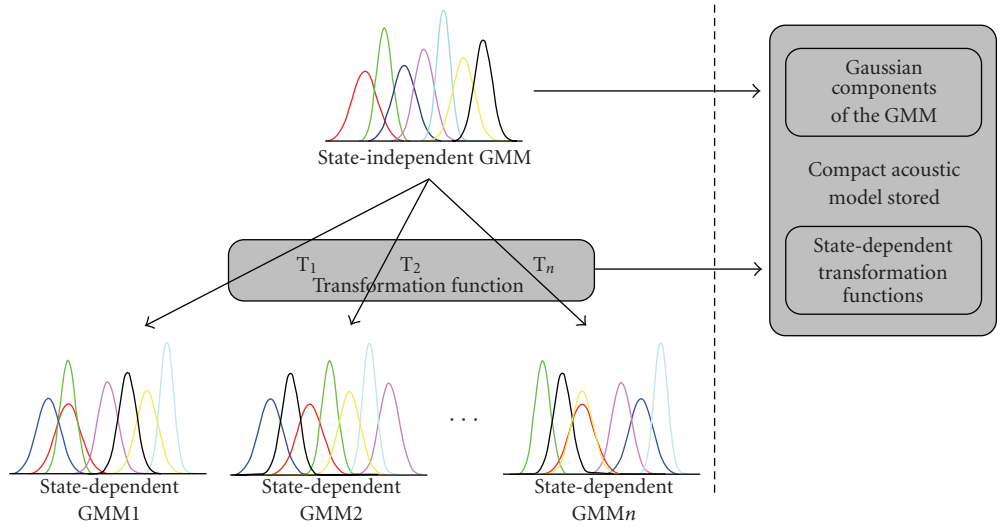

Figure 1: An overview of the proposed architecture. For state x, the state-dependent GMM (GMMx) is obtained by applying the transformation function T_x to the state-independent GMM.

models that are composed by more than 10 million free parameters and 60 K words in the lexicon. In spite of the recent advances in hardware technology, light mobile devices are not able to carry such complexity, and embedded speech-based functionalities have to be limited in order to satisfy the cost and hardware limits.

Research on embedding speech processing systems on small devices has been active for a long time. While strong advances in hardware technology have appeared, system requirements and user needs have progressed simultaneously. Therefore, hardware advances induce a scale change but fundamental issues, concerning the hardware capacities, remained.

Several architectures have been proposed for reducing the memory footprint required by the acoustic models. Vector Quantization (VQ) was introduced 25 years ago [1, 2], initially in the field of information encoding for network traffic reduction. VQ is a very low level approach. Our focus in this paper is on the modification in the modelling scheme to achieve memory footprint reduction. Moreover, VQ could be combined with the proposed modelling approach without any problem. In [3] a subspace distribution clustering method was proposed. It consists of splitting the acoustic space into streams where the distributions may be efficently clustered and tied. This method has been developed within several contexts, demonstrating a very good tradeoff between storage cost and model accuracy. Most of the recent ASR systems rely on Gaussian or state sharing, where parameter tying reduces computational time and the memory footprint, whilst providing an efficient way of estimating large context-dependent models [4–6]. In [7] a method of full Gaussian tying was proposed. It introduced Semi-continuous HMMs, for LVCSR tasks. In this architecture, all Gaussian components are grouped in a common codebook, state-dependent models being obtained by Maximum LikeLihood Estimation (MLE) based selection and weighting of the dictionary components. Numerous methods have been developped starting from this technique [8–10], mostly for hardware-limited devices.

In this paper, we present a new acoustic-model architecture where parameters are massively factored, with the purpose of reducing the memory footprint of an embedded ASR system whilst preserving the recognition accuracy. This factoring relies on a multi-level modelling scheme where a universal background model can be successively specialized to environment, speaker, and acoustic units. We propose various morphing functions for this specialization and evaluate the corresponding memory footprint reduction rates, accuracy and adaptation capacities. The performance and acoustic adaptation of the proposed approaches are investigated in various conditions within the general scheme of embedded speech recognition systems.

The next section presents an overview of our acoustic modelling architecture. Section 3 describes the corpora used for system training and testing. In Section 4, we define the application constraints targetted in this task and we present some baseline systems (obtained using classical LVCSR system). All steps of the proposed architecture are detailed in Section 5. Acoustic adaptation issues are discussed in Section 6. Finally, we conclude and we present some perspectives.

2. The Proposed Approach: Overview

HMM (Hidden Markov Model) based acoustic modelling for LVCSR usually consists in identifying and training a large set of HMMs which model various context-dependent acoustic units. This approach builds an exhaustive representation of the acoustic space, but significant amounts of information may be duplicated in overlapped state-dependent GMM (Gaussian Mixture Model). We propose to reduce significantly the memory footprint of the models

by using an acoustic model with two levels (*cf.* Figure 1). The first level attempts to represent the entire acoustic space with a unique single GMM (the state-independent GMM) shared by all HMM states (whithout considering phonetic or linguistic structures). The second level corresponds to a set of transformation functions that allows for the modelling of phone-dependent information. It is shared by all state-dependent GMMs while preserving the topology of classical HMMs.

With this architecture, the global complexity of the acoustic models depends not only on the GMM, but also on the complexity of the state-dependent transformations.

Two kinds of morphing functions where evaluated for mapping the initial word model to state-dependent ones:

(i) the first function is similar to that used in Semi-Continuous; whereas in SCHMM-based approach, one reestimates the weight with a MLE criterion, we propose two other discriminative criteria;

(ii) the second morphing function is based on a linear transformation of the mean parameters before a weight reestimation.

Both morphing functions are compared to the traditionnal HMM-based approach in Sections 5.4.1 and 5.4.2. Baseline and proposed approaches have the same memory footprint when there are compared.

To further reduce the number of parameters, a Gaussian selection for each state of the HMMs is performed. This technique is often used for embedded systems [11, 12].

More details about the proposed architecture are explained in Section 5.

3. Corpora

The availability of relevant databases for model training is a critical point for ASR systems design. Usually, application-dependent corpora are not large enough to estimate accurate models and a frequently used strategy consists in training models on a large but generic database and adapting them to the targeted context. Adapting this approach, we first use a task independent corpus, BREF [13], and two task dependent databases corresponding, respectively, to isolated digits in a clean environment (BDSON corpus [14]) and voice commands in a noisy environment (VODIS corpus [15]). These corpora are described in depth in the next section.

3.1. Application Independent Corpus

BREF. BREF [13] is a relatively large read speech corpus composed of sentences selected from the French newspaper *Le Monde*. It contains about 100 hours of speech material from 120 speakers. This corpus is considered as application-independent. It is only used for training generic models whereas BDSON and VODIS corpora are related to specific acoustic and operational environments.

3.2. Application Dependent Corpora

BDSON. BDSON [14] is a French database composed of recordings of isolated digits from 30 speakers (15 male and 15 female speakers). Recordings are performed in a clean acoustic environment. The file set was divided in two parts:

(i) one part for the application-context adaptation (BADAPT): it includes 700 digits uttered by 7 speakers (4 male and 3 female speakers); this set is used for adapting the baseline HMMs and the state-independent GMM to the application context. This phase is done once and we denote BDSON-models as the models issued from this process,

(ii) the second part for testing (BTEST): composed of 2300 digits uttered by 23 speakers (11 male and 12 female speakers).

The performance is evaluated on a digit recognition task in terms of Digit Error Rate (DER), 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).

VODIS. VODIS [15] is a French corpus dedicated to automotive applications. It includes recordings from 200 speakers. It contains a large variety of data: letters, digits, vocal commands, and 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 closed, the radio is turned on or off, the air conditioner is turned on or off). We use only the subset containing the voice commands (70 different commands are present in this subset), under the close-talk condition. This corpus was divided into two parts:

(i) one part for the application context adaptation (VADAPT): it includes 2712 commands uttered by 39 speakers;

(ii) the second part for testing (VTEST): composed of 11136 utterances of commands uttered by 160 speakers.

As we performed voice command recognition the evaluation measure used is the Command Error Rate (CER). The speakers of BADAPT and VADAPT, respectively, are different from the speakers of BTEST and VTEST (and are also different from the BREF speakers).

4. Baseline Systems

In this section, we investigate the impact of the macro-parameters on the system performance and compactness without changing the topology of the HMM. Two system profiles are defined to match the typical hardware resources available on mobile phones; a very compact model, corresponding to an upper-limit memory foot print of 6000 free parameters, and a compact model, providing 12000 free parameters. We built various models by tuning the number of Gaussian components per state and the acoustic space dimensionality.

Table 1: Evolution of DER and acoustic-model size according to the number of Gaussian components per state (context-free models). The acoustic vectors are composed of 39 coefficients (12 PLP plus energy with Δ and ΔΔ). 2300 isolated-digit recognition tests were performed on the corpus BDSON (digit/clean).

No. of gauss/state	No. of parameters	DER
2	17 064	1,48%
4	34 128	0,96%
128	1 092 096	0,96%

Table 2: Evolution of CER and acoustic-model size according to the number of Gaussian components per state (context-free models). The acoustic vectors are composed of 39 coefficients (12 PLP plus energy with Δ and ΔΔ). 11 136 tests performed on the corpus VODIS (voice command/noisy).

No. of gauss/state	No. of parameters	CER
2	17 064	5,48%
4	34 128	3,40%
128	1 092 096	1,80%

In this paper, the features extracted from the speech signal is the Perceptual Linear Predictive (PLP—[16]) coefficients. Regarding the literature (e.g., [17]), Mel Frequency Cepstral Coefficients (MFCC—[18]) are both used.

For an HMM system, the estimation of the number of parameters can be done using the equation

$$nb\_gauss * nb\_emst * (2 * nb\_param + 1), \quad (1)$$

where *nb_gauss* is the number of Gaussian in each state-GMM, *nb_emst* the number of emitting states, and *nb_param* the dimension of the acoustic parameters vectors.

4.1. Reducing the Number of Gaussians per State. Starting from a classical HMM-based model for speech, we study how the number of Gaussians impacts the system performance.

A first set of experiments is performed on the clean corpus BDSON. Table 1 presents the evolution of the Digit Error Rate (DER) according to the model size. Using 128 Gaussians per state achieves a DER of 0.96%, which corresponds to error rates reported in previous literature (see [2, 3]). Reducing the number of states results in an increase in DER to 1.48% for the smallest 2 Gaussian per state model, whilst the size of the acoustic model is decreased by a factor of 60.

In Table 2, we show the evolution of the CER according to the number of components of each emiting-state. The acoustic model is first trained with BREF and then an adaptation (MAP—[19]) is performed on the subset VADAPT of VODIS. Table 2 shows the performance on the noisy VODIS corpus. In this table, for the 2 Gaussians per state model, we observe a CER increase from 1.80% (which corresponds to the average error rate reported in the literature—[20, 21] or [22]) to 5.48% while the number of parameters is decreased by a factor of 60.

Table 3: DER and acoustic model size according to the number of Gaussian components per state (context-free models). The acoustic vectors are composed of 13 coefficients (12 PLP plus energy). 2300 isolated-digit recognition tests were performed on the corpus BDSON (digit/clean).

No. of gauss/state	No. of parameters	DER
2	5 832	4,96%
4	11 664	4,43%
128	373 248	4,52%
full	1 092 096	0,96%

Table 4: Evolution of CER and acoustic model size according to the number of Gaussian components of the emiting states (context-free models). Acoustic vectors are composed of 13 coefficients (12 PLP plus energy). 11136 voice-command recognition tests performed on the corpus VODIS (voice command/noisy).

No. of gauss/state	No. of parameters	CER
2	5 832	5,80%
4	11 664	4,80%
128	373 248	3,94%
full	1 092 096	1,80%

This first step allows to reduce the acoustic-model size by a factor of 60. Nevertheless this decrease is not enough, considering the memory space limits previously described: 6000 parameters and 12000 parameters, respectively.

4.2. Reducing the Feature-Vector Size. Starting from the 2 Gaussian-per-state models presented in the Section 4.1, further steps were taken in order to reduce the memory footprint by removing the first and second order derivatives.

Table 3 shows the influence of dynamic features (first and second order derivatives) using the clean corpus (BDSON). The DER raises from 0.96% (without any model reduction) to 4.96% for the very compact model. This 4% absolute increase leads to a reduction by a factor of 190 of the acoustic model size.

The same technique evaluated on VODIS results in similar behaviour. Since the intial model obtained 1.8% CER, the removal of first/second order (Δ and ΔΔ) derivatives leads to an absolute CER increase of about 2%. Finally, by using only static parameters (13 PLP coefficients) and 2 Gaussians (resp., 4 gaussian components) per state, the model size is divided by 180 (resp., 90) with respect to the targeted constraints and the accuracy loss is about 4% CER (resp., 3%).

The performance achieved using these reduced HMM representation act as baselines for the remains of this article. For the very compact model (5832 parameters) the baselines results are set to 5.80% with VODIS and to 4.96% with BDSON. Baselines performance obtained using the compact model (11664 parameters) are 4.80% for VODIS and 4.43% for BDSON.

Data-analysis-based methods, such as HLDA, are commonly used in LVCSR systems. However, it seems difficult to

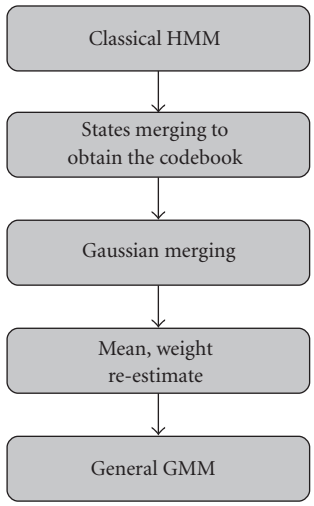

Figure 2: Process to obtain the state-independent GMM.

apply it in our experimental framework where only a small application-dependent corpus is available. We could estimate the transformation matrix on the generic corpus but we have also to adapt it to the task-dependent corpus. Some methods may be used for that, but or goal, at this point, was mainly to report baseline results of a classical method.

5. The Approach Proposed: Details

As explained in Section 2, our method is based on a two level architecture to model the acoustic units. The first level, the state-independent GMM, models the whole acoustic space. The second level consists of a set of state dependent transformation functions that model the phone dependent acoustic specifications.

The next subsections describes the method used for the state-independent GMM training and the two different classes estimating of the state-dependent morphing functions.

5.1. Training the State-Independent GMM. The state-independent GMM is derived from a classical HMM by grouping all the Gaussian components of each HMM state in a single codebook. Then, to obtain the targeted number of components, the closest Gaussians are merged. Lastly, weights are reestimated in order to get a GMM from the codebook. This sequence of steps is illustrated in Figure 2.

The first step consists of training a classical HMM. We used a set of 38 French phonemes and a classical 3-state left-right HMM topology. These HMMs are then adapted by using the appropriate adaptation subset (resp., the subset BADAPT for the BDSON corpus, and the VADAPT set for VODIS).

This inital HMM is used to build a preliminary GMM. It is obtained by grouping all the Gaussian components in a large GMM. At this point, all components are equally weighted.

Finally, this GMM is reduced by hierarchically merging the closest Gaussian pairs; we use the minimum likelihood loss criterion to identify the best Gaussian pairs. The number of expected Gaussian components is obtained using (4) and (22) according to the morphing functions used.

The distance between two components $\mathcal{N}_1(\mu_1, \Sigma_1, c_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2, c_2)$ is defined by:

$$D(\mathcal{N}_1, \mathcal{N}_2) = \frac{c_1}{c_1+c_2} \log\left(\frac{\sqrt{\Sigma}}{\sqrt{\Sigma_1}}\right) + \frac{c_2}{c_1+c_2} \log\left(\frac{\sqrt{\Sigma}}{\sqrt{\Sigma_2}}\right), \quad (2)$$

where Σ corresponds to the variance of the Gaussian component that stems from $\mathcal{N}_1$ and $\mathcal{N}_2$, as defined by (3).

The Gaussian $g'(c', \mu', \Sigma')$, results from merging $g_i(c_i, \mu_i, \Sigma_i)$ and $g_j(c_j, \mu_j, \Sigma_j)$, is defined by

$$\begin{aligned} c' &= c_i + c_j, \\ \mu' &= \frac{c_i * \mu_i + c_j * \mu_j}{c_i + c_j}, \\ \Sigma' &= \frac{c_i}{c_i + c_j}\Sigma_i + \frac{c_j}{c_i + c_j}\Sigma_j + \frac{c_i * c_j}{\left(c_i + c_j\right)^2}\left(\mu_i - \mu_j\right)\left(\mu_i - \mu_j\right)^{tr}. \end{aligned} \quad (3)$$

The last step consists of reestimating weight and mean parameters of each component, in order to obtain real GMMs and not only a codebook of Gaussians. This is achieved classically by likelihood maximization with the Expection-Maximization (EM) algorithm (see [23]).

5.2. Weight Reestimation—WRE. This approach estimates the state-dependent weight vectors from the state-independent 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. Three criteria are used for this weight reestimation:

(i) maximum Likelihood Estimation (MLE),

(ii) discriminative training by Frame Discrimination (FD),

(iii) fast Discriminative Weighting (FDW) which relies on a fast approximation of FD.

For the WRE approach the estimation of the parameters number is done using this equation:

$$\underbrace{nb\_gauss * 2 * nb\_param}_{\text{state-independent GMM}} + \underbrace{nb\_emst * nb\_sel\_gauss}_{\text{Gaussian weights}}, \quad (4)$$

where *nb_gauss* is the number of Gaussian in the state-independent GMM, *nb_param* the dimension of the acoustic parameters vectors, *nb_emst* the number of emitting states and *nb_sel_gauss* the number of selected Gaussians (Gaussian components are selected by highest weight). This last parameter is set to 20 for the very compact model and to 30 for the compact model.

In (4) the parameters *nb_param*, *nb_emst*, *nb_sel_gauss* are, respectively, set to 13 (only PLP coefficients without any delta or delta-delta parameters), 108 (due to the French set of phonemes) and 20 or 30 (depending on the required model size). So, the number of Gaussian components for the state-independent GMM is 141 for the very compact model and 324 for the compact one (in order to stay within the 6 k and 12 k limitation).

5.2.1. MLE. The estimation of weights ($\widetilde{c}_{jm}$) according to the MLE criterion is achieved by applying the updating rule:

$$\widetilde{c}_{jm} = \frac{\sum_{x_t \in \Omega_i} c_{jm} * L\left(x_t \mid G_{jm}\right)}{\sum_{x_t \in \Omega_i} \sum_{j=1}^{n_j} c_j m * L\left(x_t \mid G_{jm}\right)}, \tag{5}$$

where c_{jm} is the a priori weight of the mth Gaussian component of state j; $L(x_t \mid G_{jm})$ corresponds to the likelihood of the frame x_t knowing the Gaussian component G_{jm}, n_j the number of components of state j, and Ω_j the training corpus of state j.

Furthermore, the likelihoods of the components from the state-independent GMM are computed only once, with the state likelihoods being computed by a simple weighted combination of Gaussian-level likelihoods.

5.2.2. Discriminative Weighting. Acoustic model estimation based on the Maximum Mutual Information (MMI—[24]) criterion has been widely studied in the last decade. The general principle of this approach is to reduce the error rate by maximizing the likelihood gap between the good and the bad transcripts. The search of optimal model parameters λ is performed by maximizing the MMI objective function F_{mmie}:

$$F_{mmie}(\lambda) = \sum_{r=1}^{R} \log \frac{P_\lambda(O_r \mid M_{w_r}) P(w_r)}{\sum_{\widetilde{w}} P_\lambda(O_r \mid M\widetilde{w}) P(\widetilde{w})}, \tag{6}$$

where w_r is the correct transcript, M_w the model sequence associated with the word sequence w, $P(w)$ the linguistic probabilities and O_r an observation sequence. The denominator of the objective function sums the acoustic-linguistic probabilities of all the possible hypotheses.

One of the main difficulties in parameter estimation is the complexity of the objective function (and the derived updating rules) which requires a scoring of all the bad paths for evaluating the denominator. In order to reach a reasonable computational cost, several methods have been presented in the literature. For example, methods based on phone lattices (see [25]) or specific acoustic model topologies (see [26]).

In the particular case of our architecture, the sharing of the Gaussian components over the states could allow a direct selection of discriminant components. We highlight this point by developing, in our specific modelling framework, the *frame discrimination* method initially proposed in [26]. In this paper, the authors propose to approximate the objective function denominator by relaxing the structural constraints on the acoustic models. The resulting weight updating process consists in finding the weights $\widetilde{c}_{jm}$ that maximize the auxilary function:

$$F_c = \sum_{j,m} \left[\gamma_{jm}^{num} \log\left(\widetilde{c}_{jm}\right) - \frac{\gamma_{jm}^{den}}{c_{jm}} \widetilde{c}_{jm} \right], \tag{7}$$

where γ_{jm}^{num} and γ_{jm}^{den} are the occupancy rates estimated, respectively, on positive examples (corresponding to a correct decoding situation, noted *num*) and on negative examples (*den*); c_{jm} is the weight of the component m of state j at the previous step and $\widetilde{c}_{jm}$ is the updated weight.

By optimizing each term of this sum while fixing all other weights, the convergence can be reached in a few iterations. Each term of the previous expression is convex. Therefore, the update rule can be directly calculated using the equation:

$$\widetilde{c}_{jm} = \frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} c_{jm}, \tag{8}$$

where γ_{jm}^{k} (k can be *num* or *den*) is the probability of being in component m of state j; this probability is estimated on the corpus Ω_k that consists of all frames associated with state j.

Therefore, the occupation rate can be expressed using the likelihood functions L:

$$\begin{aligned} \gamma_{jm}^{k} &= \sum_{X \in \Omega^k} \frac{L\left(X \mid S_j\right)}{\sum_i L(X \mid S_i)} \cdot \frac{c_{jm} L\left(X \mid G_{jm}\right)}{L\left(X \mid S_j\right)}, \\ \gamma_{jm}^{k} &= \sum_{X \in \Omega^k} c_{jm} \frac{L\left(X \mid G_{jm}\right)}{\sum_i L(X \mid S_i)}. \end{aligned} \tag{9}$$

By isolating the likelihood of frame X knowing the state S_k in the denominator, we obtain:

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} c_{jm} \frac{L\left(X \mid G_{jm}\right)}{L(X \mid S_k) + \sum_{i \neq k} L(X \mid S_i)}. \tag{10}$$

In semicontinuous models, components G_{jm} are state-independent.

Let

$$\epsilon_k = \sum_{i \neq k} L(X \mid S_i), \tag{11}$$

then the occupation rate can be formulated as

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{\sum_{X \in \Omega^j} \left(L\left(X \mid G_{jm}\right) / \left(L\left(X \mid S_j\right) + \epsilon_j \right) \right)}{\sum_l \sum_{X \in \Omega^l} (L(X \mid G_{lm}) / (L(X \mid S_l) + \epsilon_l))}. \tag{12}$$

By assuming $\epsilon \simeq 0$, the numerator and the denominator of the previous rate are reduced to the update function of classical EM weight estimation. Then, the previous equation can be approximated by

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} \approx \frac{c_{jm}}{\sum_l c_{lm}}. \tag{13}$$

By combining this heuristic with (8), we obtain the weight update formula:

$$\tilde{c}_{jm} = \frac{c_{jm}^2}{\sum_l c_{lm}}. \tag{14}$$

The weight vectors are normalized (in order to obtain a sum equal to 1) after each iteration.

Thus, this training technique uses the Gaussian sharing properties of SCHMM to estimate discriminative weights directly from MLE weights, without any additional likelihood calculation. With respect to the classical MMIE training scheme, neither a search algorithm nor lattice computation is required for denominator evaluation. Hence, this method allows one to perform a model estimate at a computational cost equivalent to the one required by MLE training.

Nevertheless, this technique is based on the assumption that ϵ_i are state-independent (cf. (12)). The *a priori* validation of such an assumption seems to be difficult, especially due to the particular form of (12), where the ϵ_i quantities contribute at the same time to the numerator and to the denominator of the cost function.

5.3. Unique Linear Transformation—ULT. The method LIAMAP presented in [27] allows to adapt globally the state-independent GMM for a given state, using a unique and simple transformation. This transformation (which is common for both the mean and the variance) is a linear function:

$$\begin{aligned} \mu_{\text{state GMM}} &= \alpha \mu_{gnl\ \text{GMM}} + \beta, \\ \Sigma_{\text{state GMM}} &= \alpha^2 \Sigma_{gnl\ \text{GMM}}, \end{aligned} \tag{15}$$

where α (which is common for $\mu_{\text{state GMM}}$ and $\Sigma_{\text{state GMM}}$), a diagonal matrix, and β are estimated from a linear approximation of MAP adaptation. This adaptation (as illustrated in Figure 3) corresponds to the estimation of a linear transformation between two Gaussians obtained by

(i) merging the Gaussian components of the state-independent GMM. The final Gaussian is defined by μ and Σ, respectively the mean and the covariance matrix,

(ii) adapting the Gaussian components of the state-independent GMM to state-specific data (using MAP) and then merging adapted Gaussians into a unique Gaussian defined by $\tilde{\mu}$ and $\tilde{\Sigma}$,

(iii) computing α and β as the parameters of a linear adaptation between Gaussian $\mathcal{N}(\mu, \Sigma)$ and Gaussian $\mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$.

Each final Gaussian component (defined by its mean μ'_m and its covariance matrix Σ'_m) is computed as follows:

$$\mu'_m = \tilde{\Sigma}^{1/2} \Sigma^{-1/2} (\mu_m - \mu) + \tilde{\mu} \tag{16}$$

$$\Sigma'_m = \tilde{\Sigma} \Sigma^{-1} \Sigma_m. \tag{17}$$

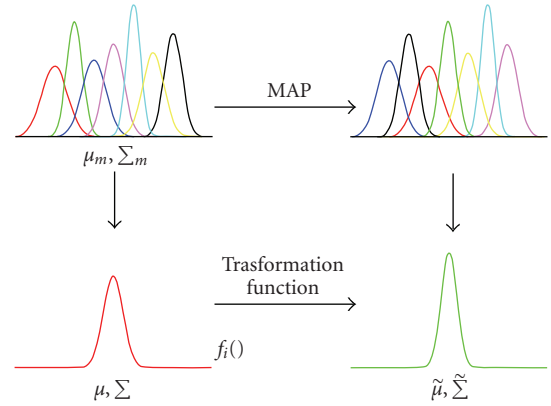

Figure 3: LIAMAP: Method to estimate a unique linear transformation for all Gaussians of a codebook.

Equation (16) can be expanded as

$$\mu'_m = \tilde{\Sigma}^{1/2} \Sigma^{-1/2} \mu_m - \tilde{\Sigma}^{1/2} \Sigma^{-1/2} \mu + \tilde{\mu} \tag{18}$$

if we set

$$\begin{aligned} \alpha &= \tilde{\Sigma}^{1/2} \Sigma^{-1/2}, \\ \beta &= -\tilde{\Sigma}^{1/2} \Sigma^{-1/2} \mu + \tilde{\mu} \end{aligned} \tag{19}$$

then (16) and (17) become

$$\mu'_m = \alpha \mu_m + \beta, \tag{20}$$

$$\Sigma'_m = \alpha^2 \Sigma_m. \tag{21}$$

Equations (20) and (21) correspond to a linear adaptation function defined only by the vectors α and β (the transformation is shared by all the Gaussian components of the state-independent GMM).

Our technique for adaptation is similar to the fMLLR (feature Maximum Likelihood Linear Regression—[28, 29]), but it has several advantages: the α parameters of (20) is a simple diagonal matrix instead of a full matrix, the criteria used are simpler (just MAP and lost-likelihood), there is no matrix inversion.

In our context, ULT is used as a first step (optional) before the weight reestimation. The WRE step (*cf.* 5.2) is always performed (using ULT or not). Figure 4 presents the complete process (ULT+WRE).

The usage of the ULT+WRE approach requires more CPU consumption compared to WRE (only) method. Indeed, during the test, before performing likelihood estimation, the ULT+WRE approach requires the estimation of the GMM parameters of each state, because only the α and β parameters of the transformation are stored. Moreover, whilst the ULT+WRE approach requires the estimation of the likelihoods for each Gaussian component of each state, the WRE (without ULT)calculates the state likelihood as a weighted sum of pre-computed Gaussian likelihoods.

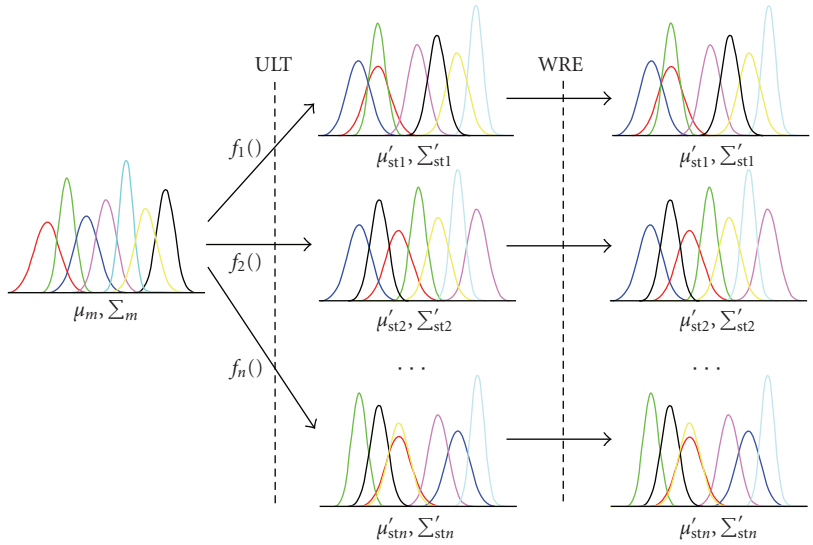

Figure 4: State-dependent transformation by applying ULT followed by WRE.

For the ULT+WRE approach the estimation of the parameters number is calculated as

$$\underbrace{nb\_gauss * (2 * nb\_param)}_{\text{state-independent GMM}} + \underbrace{nb\_emst * (2 * nb\_param + nb\_gauss\_sel)}_{\text{linear transf. \& weight}}, \quad (22)$$

where *nb_gauss* is the number of Gaussian in the state-independent GMM, *nb_param* the dimension of the acoustic parameters vectors, *nb_emst* the number of emitting states and *nb_sel_gauss* the number of selected Gaussian. This last parameters is still set to 20 for the very compact model and to 30 for the compact one.

In (22) the parameters *nb_param*, *nb_emst*, *nb_sel_gauss* are, respectively, set to 13 (only PLP coefficients without any delta or delta-delta parameters), 108 (due to the French set of phonemes) and 20 or 30 (considering the model size expected). So, the number of Gaussian components for the state-independent GMM is 33 for the very compact model and 216 for the compact one (in order to stay under the 6 k and 12 k limits, resp.).

5.4. Results. The presented approach allows state-models to be trained directly from a unique GMM (the state-independent GMM) that represents the whole acoustic space. This process consists of two steps (ULT and WRE) for which the influence is highlighted in the two next subsection.

In Tables 5 and 7, we compare the Digit Error Rate of all methods presented here with the baseline. Tables 6 and 8 present the Command Error Rate obtained on VODIS corpus (noisy conditions) and results are also compared with the baseline.

Table 5: Results obtained with WRE approach compared to the baseline system. Digit Error Rate depending on the weight reestimation rules (MLE, FDW et FD) without ULT. 2 300 tests performed on BDSON corpus (clean).

	WRE			*Baseline*
	MLE	FDW	FD	
Very compact model	3.35%	2.78%	3.13%	4.96%
Compact model	2.83%	2.17%	2.48%	4.32%

Table 6: Results obtained with WRE approach compared with the *baseline*. Command Error Rate depending on the weight reestimation rules (MLE, FFDW and FD) without ULT. 11 136 tests performed on VODIS corpus (noisy).

	WRE			*Baseline*
	MLE	FDW	FD	
Very compact model	6.05%	8.54%	5.99%	5.80%
Compact model	5.15%	7.50%	5.15%	4.80%

5.4.1. WRE Approach. With clean data (BDSON corpus), the WRE approach outperforms, in terms of Digit Error Rate, the baseline system(*cf.* Table 5). For the very compact model, the minimal DER is 2.78% (obtained with the FDW weight updating rule); to be compared to the 4.96% for the baseline system, a relative gain greater than 40% is achieved. Moreover, with the compact model, we note a decrease of the DER from 4.32% to 2.17% (always with FDW) which corresponds to a relative decrease of about 50%.

In noisy condition (with VODIS corpus), the baselines obtain a CER of 5.80% for the very compact model and of 4.80% for the compact model (cf. Table 6).

We can notice that the WRE approach alone does not allow a decrease of the CER. The best CER reaches 5.99% (WRE with FD weight updating rule) for the smallest model, whereas the CER of the baseline is 5.80%.

Table 7: Results obtained with ULT+WRE approach compared with the *baseline*. Digit Error Rate depending on the weight reestimation rules (MLE and FD) with ULT. 2 300 tests performed on BDSON corpus (clean).

	ULT+WRE		*Baseline*
	MLE	FD	
Very compact model	3.04%	3.39%	4.96%
Compact model	2.78%	2.26%	4.32%

Table 8: Results obtained with ULT+WRE approach compared with the *baseline*. Command Error Rate depending on the weight reestimation rules (MLE and FD) with ULT. 11 136 tests performed on VODIS corpus (noisy).

	ULT+WRE		*Baseline*
	MLE	FD	
Very compact model	5.25%	5.11%	5.80%
Compact model	4.01%	4.27%	4.80%

For this reason, we introduced a previous step before WRE which perform an adaptation of the state-independent GMM before applying the weight reestimation (WRE step).

5.4.2. ULT+WRE Approach. In clean conditions (refering to Table 7), we can observe that the ULT step does not allow a DER decrease superior to the WRE alone approach. Nevertheless, there is a significant decrease of DER compared to the baseline. Indeed, the DER of the very compact model is reduced more than 38% (to 3.04% with MLE weight updating rule) and more than 48% (to 2.26% with the FD weight updating rule) for the compact model.

Table 8 show results for the case of noisy condition. The ULT+WRE approach reduces the CER to 5.11% (FD weight updating rule) for the very compact model. This represents a relative reduction of around 12% compared to the baseline (CER at 5.80%). With the upper memory size constraint, the CER decreases to 4.01% (MLE weight updating rule). Compared to the 4.80% of the baseline, it corresponds to a relative reduction of about 16% while the memory footprint stays unchanged.

5.4.3. Conclusion. In conclusion, the proposed approach provides an important decrease of the error rates with clean data (BDSON), with or without ULT and whatever weight updating rule we used. For very compact model, our approach reaches a DER between 2.78% and 3.39%. With the compact model, DER is between 2.17% and 2.83%. This represents a relative decrease between 30% and 50%.

In noisy conditions, the WRE approach seems not to be sufficient. The CER obtained with our approach is slightly worse that the baseline one: the CER loss is about 0.2% (for the very compact model with FD weight updating rule), however the DER differences remain inside the confidence interval. The use of ULT (before WRE) allows Gaussian mean moving which seems to improve the model robustness.It permits to be more efficient that WRE approach which operates only on the weight vector. We noticed that it allows relative gains between 10% and 15%.

Lastly, since FDW provides great improvements on clean data, the approximation performed seems not to be robust to noise. With the VODIS corpus, the weight reestimate is always better with MLE or FD than with FDW.

6. Fast Acoustic Adaptation

Generally, for speaker/environment adaptation, speech recognition systems use MLLR [30] and/or MAP [19] methods. In the literature (e.g., [31]) we can notice that these techniques allowed an increase of accuracy of around 10%. In this section, we try to show that our approach have similar adaptation facilities.

Our architecture requires relatively amounts of data for estimate acoustic parameters, compared to the classical HMM-based models. In this approach, the standard topology of the HMM models is preserved but all the states are sharing a state-independent GMM that represents the common acoustic features. This specific model structure could lead to a new adaptation scheme where state-dependent and state-independent features could be separately adapted. Considering the very low amount of data available for training, state-dependent adaptation seems to be untractable. However, the shared GMM could be adapted by using the full adaptation data set. This global adaptation is based on the following idea: if there is a discrepancy between a state model and the same state model adapted to a speaker, then the same discrepancy probably exists between all the state-models. We will try to highlight this point by adapting the state-independent GMM without changing the transformation funtions.

This process, illustrated in Figure 5, is composed of 3 steps:

(1) training phase: the state-independent GMM and the state-transformations are trained with the development data,

(2) adaptation phase: the state-independent GMM is adapted with a small amount of few data from a speaker,

(3) testing phase: instead of applying the transformation on the state-independent GMM, they are applied to the speaker-depedent GMM.

As VODIS is the noisy corpus, we use it to test the adaptation approach. VODIS contains a subset with well-balanced phonetic sentences. Each speaker has uttered 5 sentences which will be used for adapting the state-independent GMM to a speaker. These sentences are different to the commands used for evaluating the adaptation step (VADAPT or VTEST sets).

In order to adapt the state-independent GMM we use the MAP method proposed in [32]. As is usually the case in speaker recognition, we perform this adaptation only on the mean parameters.

In Table 9, we show the results obtained with and without adaptation. Table 9(a) corresponds to the WRE approach and Table 9(b) to the ULT+WRE approach. An important gain could be noticed whatever the approach we used.

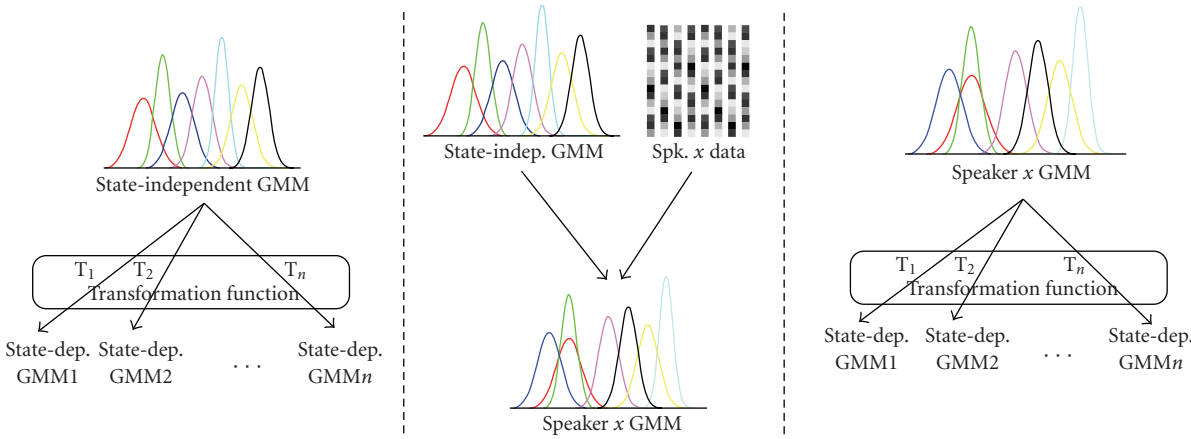

Figure 5: The proposed architecture and adaptation steps. Training phase utilises WRE or ULT+WRE without addaptation. Adaptation phase adapts the state-independent GMM using gathered speaker data. Testing phase makes use of the speaker-adapted model.

Table 9: Command Error Rate for WRE approach (9(a)) and ULT+WRE approach (9(b)) with and without state-independent GMM adaptation (adaptation performed on 5 sentences phonetically balanced). 11136 voice-command recognition tests performed on VODIS corpus (noisy).

(a) WRE approach

	Without adaptation			With adaptation		
	MLE	FDW	FD	MLE	FDW	FD
Very compact model (6 k.)	6,05%	8,54%	5,99%	5,48%	8,67%	5,36%
Compact model (11 k.)	5,15%	7,50%	5,15%	4,67%	7,28%	4,63%

(b) ULT+WRE approach

	Without adaptation		With adaptation	
	MLE	FD	MLE	FD
Very compact model (6 k.)	5,25%	5,11%	4,76%	4,48%
Compact model (11 k.)	4,01%	4,27%	3,64%	3,80%

Indeed, the WRE approach (cf. Table 9(a)) allows a relative gain of 10%. The CER of the very compact model using FD weight updating rule without adaptation is 5.99% and with adaptation it decreases to 5.36%, which represents a relative decrease of 10.52%. The gains obtained with compact models are similar (a relative decrease of 10.1%, with FD weight updating rule).

The models based on the FDW weight updating rule seem not benefit from the adaptation phase; there is no significant decrease of the CER. It results certainly from the fact that FDW is based on the hypothesis that ϵ_x (cf. (11)), which corresponds to the likelihood of non-typical Gaussians of a state, is insignificant compared to the other terms.

Table 9 shows that the models using the ULT+WRE approach are able to take more advantage of this adaptation scheme. The relative CER decrease is between 9% and 12%. For the compact model based on the MLE weight updating rule before adaptation, the CER is 4.01%. On this configuration, the adaptation allows to reach 3.64% CER (12.33% relative gain).

These results confirm the initial assumption of a relative independance between phoneme-related and speaker-related information. We obtain a relative gain between 9% and 12%, which is close to the gains typically observed in speech recognition with MAP or MLLR adaptation.

In conclusion, this approach presents several points of interests with regards to the state-free adaptation process compared to classical systems:

(i) only a small amount of data is needed to adapt efficiently the acoustic model due to the fact that all the available data are shared to adapt the state-independent GMM;

(ii) no state alignment is required because there is only one GMM to adapt (not one GMM per state and/or class);

(iii) the computational cost of this adaptation remains very low thanks to the fact there is only one GMM to adapt.

7. Conclusion

This paper deals with the issue of speech recognition in situations of limited memory resource and limited computational cost. Starting from the idea that, in classical HMM-GMM based models, Gaussian mixtures encode not only phoneme-specific information but also some general information about speech, we propose an approach that aims at limiting the redundancy in acoustic models. This is achieved by a two level architecture in which the whole acoustic space and subword units are separatly modelled. At the upper level, a general GMM models the speech signal, state-dependent models being obtained by applying compact transformations on this common GMM.

The proposed methods are evaluated in various experimental conditions. They are compared to classical HMM models with respect to the limited hardware resource typically offered by a mobile phone.

Firstly, we evaluated baseline systems that are obtained by decreasing the number of Gaussians per mixtures and by reducing the acoustic space dimensionality. Results show clearly that the classical HMM-GMM based architecture is dramatically impacted by the strong complexity reduction induced by mobile-phone hardware limits: with respects to a large acoustic model used in LVCSR tasks, the error rates are multiplied, at least, by a factor of 6 in all the test conditions.

Then, we proposed our two level architecture in various configurations. Two kinds of morphing functions were evaluated, respectively, based on weight reestimate (WRE) and a smoothed MAP adaptation (ULT).

The first approach consists of reestimating state-dependent weight vectors from the state-independent GMM. Several criteria were used, one based on likelihood maximisation (MLE) and two based on discriminative criteria (FD and FDW). Considering the CPU resources required by the frame discrimination method (FD), we introduced the FDW criterion, which is a fast approximation of FD. This approximation is restricted to semi-continuous HMMs; it allows a discriminative reestimation of the weights for a computational cost similar to the one required by MLE training.

The experimental results demonstrated the efficiency of the discriminative training of weight vectors on clean conditions: we observed a relative error rate decrease between 32% and 55%, according to the system configuration, especially with FDW, which outperforms the standard FD method. Hovewer, discriminative weighting does not provide any gain in noisy conditions. Morever, the fast approximation of frame discrimination seems to be highly sensitive to the acoustic conditions: error rates increase strongly on the VODIS corpus, compared to the standard FD method.

In order to improve the recognition rates in noisy environments, we proposed a morphing function family operating on both mean and weight parameters. This method relies on a global adaptation of the state-independent GMM by a simple linear transformation (ULT) shared by all the Gaussian components. This adaptation is performed state by state. Even if ULT does not obtain any decrease of error rate in clean conditions (compared to the WRE only approach), it provides a significant accuracy improvement in noisy conditions. In this case, WRE obtains error rates similar to the baseline and ULT+WRE allows a relative decrease of the CER between 9% and 16% (compared with the baseline as well).

Lastly, the proposed architecture offers a simple and efficient way of dealing with the speaker/environment adaptation issues under memory and CPU constraints. Assuming that speaker-related and phoneme-related information is independent, we proposed a fast adaptation scheme that is tractable in spite of the low amount of adaptation data, and under strict hardware constraints. In noisy conditions (VODIS, voice-command recognition) this adaptation scheme obtained a relative decrease of the CER between 9% and 12% compared with WRE or ULT+WRE without adaptation. Moreover, this adaptation does not require significant computing resources, nor much adaptation data.

We plan to investigate other transformation families in order to improve the discriminative capacity of the acoustic models. Moreover, subspace clustering methods have demonstrated good efficiency on embedded systems. The combination of the proposed architecture and subspace clustering could improve the tradeoff between memory footprint and model accuracy.

References

[1] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Transactions on Information Theory*, vol. 29, no. 4, pp. 473–491, 1983.

[2] R. Billi, "Vector quantization and Markov source models applied to speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '82)*, vol. 7, pp. 574–577, May 1982.

[3] E. Bocchieri and B. K.-W. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 264–275, 2001.

[4] S. J. Young, "The general use of tying in phoneme-based HMM speech recognisers," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '92)*, pp. 569–572, San Francisco, Calif, USA, March 1992.

[5] M.-Y. Hwang and X. Huang, "Shared-distribution hidden Markov models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 414–420, 1993.

[6] X. D. Huang, M.-Y. Hwang, L. Jiang, and M. Mahajan, "Deleted interpolation and density sharing for continuous hidden Markov models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 2, pp. 885–888, Atlanta, Ga, USA, May 1996.

[7] X. Huang and M. Jack, "Large-vocabulary speaker-independent continuous speech recognition with semi-continiuous hidden Markov models," in *Proceedings of the 1st European Conference on Speech Communication and Technology (Eurospeech '89)*, pp. 1163–1166, Paris, France, September 1989.

[8] J. Macas-Guarasa, A. Gallardo, J. Ferreiros, J. Pardo, and L. Villarrubia, "Initial evaluation of a preselection module for a flexible large vocabulary," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96)*, pp. 1343–1346, Philadelphia, Pa, USA, October 1996.

[9] T. Vaich and A. Cohen, "Comparison of continuous-density and semi-continuous HMM in isolated words recognition systems," in *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech '99)*, pp. 1515–1518, Budapest, Hungary, September 1999.

[10] J. Park and H. Ko, "Achieving a reliable compact acoustic model for embedded speech recognition system with high confusion frequency model handling," *Speech Communication*, vol. 48, no. 6, pp. 737–745, 2006.

[11] J. Park and H. Ko, "Compact acoustic model for embedded implementation," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP '04)*, pp. 693–696, Jeju Island, South Korea, October 2004.

[12] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 1, pp. 185–188, Toulouse, France, May 2006.

[13] L. F. Lamel, J. L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," in *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech '91)*, pp. 505–508, Genoa, Italy, September 1991.

[14] 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 the International Conference on Acoustics, Speech and Signal Processing (ICASSP '84)*, vol. 3, pp. 324–327, San Diego, Calif, USA, March 1984.

[15] 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 the 6th International Conference on Spoken Language Processing (ICSLP '00)*, pp. 378–382, Beijing, China, October 2000.

[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[17] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 1, pp. 457–460, Philadelphia, Pa, USA, March 2005.

[18] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[19] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[20] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[21] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "Speaker independent isolated digit recognition using hidden Markov models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '83)*, vol. 8, pp. 1049–1052, April 1983.

[22] A. B. Poritz and A. G. Richter, "On hidden Markov models in isolated word recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '86)*, vol. 11, pp. 705–708, April 1986.

[23] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, no. 4, pp. 289–307, 1995.

[24] 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 the International Conference on Acoustics, Speech and Signal Processing (ICASSP '86)*, pp. 49–52, Tokyo, Japan, April 1986.

[25] X. Aubert and H. Ney, "Large vocabulary continuous speech recognition using word graphs," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '95)*, vol. 1, pp. 49–52, Detroit, Mich, USA, May 1995.

[26] P. C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *Proceedings of the ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, pp. 7–16, Paris, France, 2000.

[27] D. Matrouf, O. Bellot, P. Nocera, G. Linarès, and J. F. Bonastre, "Structural linear model-space transformations for speaker adaptation," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 1625–1628, Geneva, Switzerland, September 2003.

[28] M. J. F. Gales, "Maximum likelihood linear transformations for hmmbased speech recognition," Tech. Rep., Engineering Department, Cambridge University, Cambridge, UK, May 1997.

[29] X. Lei, J. Hamaker, and X. He, "Robust feature space adaptation for telephony speech recognition," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, vol. 2, pp. 773–776, Pittsburgh, Pa, USA, September 2006.

[30] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," in *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP '94)*, pp. 451–454, Yokohama, Japan, September 1994.

[31] O. Bellot, *Adaptation au locuteur des modèles acoustiques dans le cadre de la reconnaissance automatique de la parole*, Ph.D. thesis, Université d'Avignon, LIA, Cedex, France, May 2006.

[32] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.