ACOUSTIC MODEL ADAPTATION BASED ON COARSE/FINE TRAINING OF TRANSFER VECTORS USING DIRECTIONAL STATISTICS

Shinji Watanabe and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

In this paper, we reformulate an adaptation scheme of Coarse/Fine Training (CFT) of transfer vectors in acoustic modeling by using *directional statistics*. In CFT, the transfer vector is decomposed into a unit direction vector and a scaling factor. By using coarse tied Gaussian class (coarse class) estimation for the unit direction vector, and by using fine tied Gaussian class (fine class) estimation for the scaling factor, we can obtain accurate transfer vectors with a small number of free parameters. Directional statistics is a method for analyzing geometric parameters (e.g. angle and unit vector) using directional data, and is suited for the analysis of the CFT representation. Using directional statistics as a basis, we construct expectation-maximization algorithms for CFT parameters analytically using the maximum likelihood and Bayesian (maximum a posteriori) approaches. In particular, with the Bayesian approach, prior and posterior distributions for unit direction vectors are represented with a von Mises distribution, a representative distribution in directional statistics. Speaker adaptation experiments show that our proposal improves the performance of large vocabulary continuous speech recognition due to the efficient coarse/fine representation of transfer vectors, compared with the conventional transfer vector adaptation.

1. INTRODUCTION

Speaker adaptation techniques are aimed at improving speech recognition performance solely by using a small amount of adaptation data. Under such conditions, common Maximum Likelihood (ML) approaches often have a detrimental effect on performance due to over-training, and conventional adaptation techniques possess mechanisms for suppressing such over-training. Transformative adaptation approaches, as typified by the Maximum Likelihood Linear Regression (MLLR) adaptation, do not estimate the target model directly, but estimate mapping or transformation from initial to target models [1–5]. Model parameters are usually grouped into classes in advance, and we estimate a set of transformation parameters for each class, so that a reasonable amount of data is available for estimating each transformation. Bayesian adaptation approaches, as typified by the Maximum A Posteriori (MAP) adaptation, directly estimate individual parameters in the target model, taking into account both data and prior distributions [6]. When there is a shortage of data for a model parameter, the prior distribution becomes dominant in the estimation and prevents the parameter from being estimated solely relying on the data. An advantage of transformative adaptation over Bayesian adaptation is the quick effect of adaptation for a small amount of data. This is because there are fewer free parameters to be estimated, due to the use of parameter classes where model parameters in the same class are commonly transformed. On the other hand, an advantage of Bayesian adaptation over transformative adaptation is its asymptotic property, where the performance of an adapted model comes close to that of a speaker-dependent model when a large amount of data is available. This is because the Bayesian adaptation estimates model parameters individually instead of using parameter classes, and theoretically this becomes equivalent to ML estimation for an infinite amount of data. Thus, we can also view the differences between adaptation approaches in terms of the resolution of model parameter classes estimated directly or indirectly through adaptation. In the above description, the transformative adaptation is based on a coarse resolution class (coarse class), while the Bayesian adaptation is based on individual parameters, which offer extremely fine resolution (fine class). We recently proposed a new adaptation technique Coarse/Fine Training (CFT), which optimally combines both coarse and fine estimation to utilize their advantages for any amount of data [7]. The CFT approach focuses on the transfer vector estimation of a Gaussian mean from an initial model to a target model. The transfer vector is decomposed into a direction vector and a scaling factor. By using coarse classes and fine classes to estimate direction vectors and scaling factors, respectively, we are able to estimate accurate transfer vectors with a small number of free parameters. In this paper, we introduce a new formulation of CFT by imposing the geometrical constraint of the unit length on the direction vector in the feature dimension space. This approach is inspired by the statistical analysis of directional data, which is known as directional statistics [8]. Using directional statistics as a basis and by introducing a von Mises distribution in particular, we can successfully derive an analytical solution for the transfer vectors of ML and MAP frameworks. Thus, our adaptation scheme provides a new paradigm for acoustic modeling based on directional statistics. We also apply a tree structure, which organizes hierarchical Gaussian classes [2], in the CFT implementation, and represents the coarse/fine classes efficiently. We demonstrate the effectiveness of our proposal in a supervised speaker adaptation task in large vocabulary continuous speech recognition experiments.

2. COARSE/FINE REPRESENTATION OF TRANSFER VECTORS

In this section, we introduce a new coarse/fine representation of transfer vectors for mean vector parameters in acoustic modeling. With conventional transfer vector approaches, to reduce the number of free parameters, transfer vector $\boldsymbol{\mu}_k^{new}$ is shared by several Gaussians as follows:

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{ini} + \boldsymbol{\Delta}_{l(k)}, \tag{1}$$

where $\boldsymbol{\mu}_k^{ini}$ denotes a mean vector parameter of the initial data in Gaussian k. $\boldsymbol{\Delta}_{l(k)}$ denotes a transfer vector where class $l(k)$ is a set of Gaussians including Gaussian k [1, 2].

In the coarse/fine representation, we estimate $\boldsymbol{\Delta}_{i(k)}$ and scaling factor $g_{j(k)}$ using different classes $i(k)$ and $j(k)$ from Gaussian class k [7], as follows:

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{ini} + g_{j(k)}\boldsymbol{\Delta}_{i(k)}, \tag{2}$$

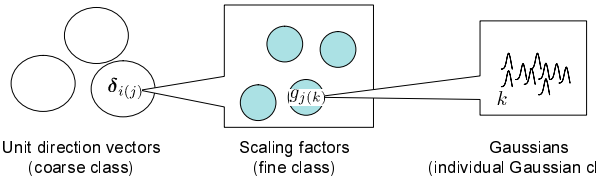

Fig. 1. Hierarchical inclusion relation between unit direction vector $\boldsymbol{\delta}_{i(j)}$, scaling factor $g_{j(k)}$, and Gaussian k.

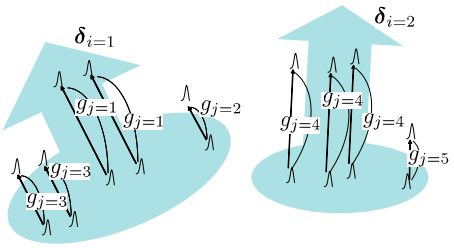

Fig. 2. Behavior of transfer vectors represented by $\boldsymbol{\delta}_{i(j)}$ and $g_{j(k)}$ in a variance-normalized feature dimension space.

where classes $i(k)$ and $j(k)$ are different sets of Gaussians that both include Gaussian k. That is to say, *the direction vector and the scaling factor are tied across distinct sets of Gaussians*. There is only one parameter for the scaling factor $g_{j(k)}$, and it is much smaller than that for the direction vector $\boldsymbol{\Delta}_{i(k)}$, which is the same as the number of feature dimensions. This means that the estimation of the scaling factor $g_{j(k)}$ requires a much smaller amount of data than the estimation of the direction vector $\boldsymbol{\Delta}_{i(k)}$. Therefore, we can estimate the transfer vector for Gaussians even with a small amount of adaptation data by (i) estimating the direction vector from the large fraction of adaptation data assigned to coarse tied Gaussians (coarse class estimation), and (ii) estimating the scaling factor from the small fraction of adaptation data assigned to fine tied Gaussians (fine class estimation).

In this paper, we further impose three constraints in Eq. (2) based on directional statistics. Namely, we adopt *unit vectors* for directional vectors and leave the magnitude estimation of transfer vectors to scaling factors. Then, the unit direction vector has the constraint that $|\boldsymbol{\delta}_i| = 1$. Moreover, we adopt a hierarchical inclusion relation whereby coarse class i is a set of fine class j, and fine class j is a set of individual Gaussian class k, i.e., $k \in j(k) \in i(j)$, as shown in Figure 1. We also adopt a representation of variance-normalized transfer vectors instead of a direct representation of transfer vectors. These simplify the derivation of the EM algorithm in the next section. Using the unit vector constraint, the hierarchical inclusion relation and the variance-normalized representation as a basis, we obtain a new coarse/fine representation as follows:

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{ini} + g_{j(k)} \Sigma_k^{ini} \boldsymbol{\delta}_{i(j)}, \tag{3}$$

where Σ_k^{ini} denotes a covariance matrix parameter of the initial data in Gaussian k. We call this "directional coarse/fine representation", which efficiently represents the transfer vectors of mean parameters in a variance-normalized feature dimension space, as shown in Figure 2. In the directional coarse/fine representation, unit direction vectors and scaling factors are statistically estimated based on directional statistics.

3. ANALYTIC SOLUTIONS OF CFT BASED ON DIRECTIONAL STATISTICS

In this section we discuss the analytic solutions we obtain when estimating coarse/fine parameters $\boldsymbol{\delta}_{i(j)}$ and $g_{j(k)}$. To avoid complicated indexes, we simplify indexes $i(j)$ and $j(k)$ to i and j in this section. The auxiliary function (known as the Q function) in the EM algorithm is obtained by calculating the expectation of complete data likelihood with respect to a posterior distribution for latent variables:

$$\sum_{k,t} \zeta_k^t \log \mathcal{N}(\boldsymbol{o}^t | \boldsymbol{\mu}_k^{new}, \Sigma_k^{ini}), \tag{4}$$

where k denotes all Gaussian indexes over all HMM states in all phoneme categories, ζ_k^t denotes the occupancy count of frame t assigned to Gaussian k, and $\mathcal{N}(\boldsymbol{o}^t | \boldsymbol{\mu}_k^{new}, \Sigma_k^{ini})$ denotes the Gaussian for feature vector $\boldsymbol{o}^t$ of frame t. In Eq. (4), we omit HMM state transition and mixture weight factors because we only discuss the estimation of $\boldsymbol{\mu}_k^{new}$. By substituting the mean vector representation (Eq. (3)) into Eq. (4), we can obtain the concrete form of Eq. (4). Although, in general, it is difficult to calculate the summation of variables in tied Gaussian classes i and j with respect to Gaussian k, by utilizing the hierarchical inclusion relation and the variance-normalized representation, we can derive the simple form of the auxiliary function as:

$$\begin{aligned} &\sum_{k,t} \zeta_k^t \log \mathcal{N}(\boldsymbol{o}^t | \boldsymbol{\mu}_k^{ini} + g_j \Sigma_k^{ini} \boldsymbol{\delta}_i, \Sigma_k^{ini}) \\ &\quad \propto -\frac{1}{2} \sum_i \sum_{j \in i} \zeta_j (g_j)^2 + \sum_i \left(\boldsymbol{\delta}_i \cdot \sum_{j \in i} g_j \zeta_j \boldsymbol{\rho}_j \right), \end{aligned} \tag{5}$$

where $(\boldsymbol{a} \cdot \boldsymbol{b})$ denotes the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and $j \in i$ denotes the summation with respect to fine class j included in coarse class i. We omit terms that do not depend on $\boldsymbol{\delta}_i$ and g_j. In addition to coarse/fine parameters g_j and $\boldsymbol{\delta}_i$, Eq. (5) is expressed by two other statistics ζ_j and $\boldsymbol{\rho}_j$, defined as:

$$\begin{aligned} \zeta_j &\equiv \sum_{k \in j} \zeta_k \\ \boldsymbol{\rho}_j &\equiv \frac{1}{\zeta_j} \sum_{k \in j} (\Sigma_k^{ini})^{-1} \zeta_k \left(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^{ini} \right). \end{aligned} \tag{6}$$

where ζ_j is the sum of the occupancy counts over a set of Gaussians k in fine class j. $\widehat{\boldsymbol{\mu}}_k$ is an ML estimate of mean parameter $\boldsymbol{\mu}_k$, i.e., $\widehat{\boldsymbol{\mu}}_k = \sum_t \zeta_k^t \boldsymbol{o}^t / \zeta_k$ in HMMs. $\boldsymbol{\rho}_j$ is the occupancy weighted average of variance-normalized transfer vectors in Gaussians k included in fine class j. Therefore, $\boldsymbol{\rho}_j$ substantially corresponds to a variance-normalized transfer vector in fine class j, and we call it the "averaged transfer vector". These two statistics are sufficient statistics for estimating g_j and $\boldsymbol{\delta}_i$ in directional statistics, and we can obtain these statistics after the E-step in the EM algorithm, similar to the case of usual mean parameter estimation in HMMs. The E-step can be performed by substituting g_j and $\boldsymbol{\delta}_i$ estimated in the M-step into the output distributions. Therefore, in the next section, we introduce M-step solutions for CFT parameters using ML and Bayesian approaches.

3.1. ML solution (CFT-ML) for M-step

The ML solution is obtained by differentiating the auxiliary function from Eq. (5) with respect to $\boldsymbol{\delta}_i$ and g_j under the constraint that $|\boldsymbol{\delta}_i| = 1$, as follows:

$$\boldsymbol{\delta}_i^{ML} = \frac{\sum_{j \in i} g_j^{ML} \zeta_j \boldsymbol{\rho}_j}{\left| \sum_{j \in i} g_j^{ML} \zeta_j \boldsymbol{\rho}_j \right|}. \tag{7}$$

$$g_j^{ML} = \left(\boldsymbol{\delta}_i^{ML} \cdot \boldsymbol{\rho}_j \right). \tag{8}$$

Equation (7) shows that the estimated unit direction vector retains the constraint, which guarantees stable training so firmly that the directional coarse/fine representation is retained even after the EM step. Equation (8) shows that g_j^{ML} corresponds to the magnitude when averaged transfer vector $\boldsymbol{\rho}_j$ is projected onto unit direction vector $\boldsymbol{\delta}_i^{ML}$. Note that if we use individual Gaussian classes for i and j (i.e. $i \rightarrow k$ and $j \rightarrow k$), mean vectors obtained by $\boldsymbol{\delta}_k^{ML}$ and g_k^{ML} are equivalent to ML estimate ($\widehat{\boldsymbol{\mu}}_k$) in HMMs.

3.2. Bayesian solution (CFT-MAP) for M-step

We introduce the exact posterior distribution needed to obtain the MAP estimates of the CFT parameters. With the Bayesian approach, the key is to set appropriate prior distributions where their probabilistic variables satisfy the parameter constraint. We set Gaussian $\mathcal{N}(g_j|u_j^0,(v_j^0)^{-1/2})$ for a prior distribution of g_j, which is a no-constraint continuous value, where u_j^0 and v_j^0 are prior parameters. $\boldsymbol{\delta}_i$ has the constraint that $|\boldsymbol{\delta}_i| = 1$, and we set von Mises distribution $\mathcal{M}(\boldsymbol{\delta}_i|\boldsymbol{\nu}_i^0,\kappa_i^0)$ for a prior distribution of $\boldsymbol{\delta}_i$, where $\boldsymbol{\nu}_i^0$ and κ_i^0 are prior parameters. The von Mises distribution is widely used in directional statistics to represent the probabilistic distribution of a unit direction vector [8]. The von Mises and Gaussian distributions belong to the exponential family, and the posterior distributions can be analytically solved as having the same function distribution as the prior distributions. Therefore, MAP estimates are obtained by extracting the maximum probabilistic values for the obtained posterior distributions. The analytical solution is as follows:

$$\boldsymbol{\delta}_i^{MAP} = \frac{\kappa_i^0\boldsymbol{\nu}_i^0 + \sum_{j\in i} g_j^{MAP}\zeta_j\boldsymbol{\rho}_j}{\left|\kappa_i^0\boldsymbol{\nu}_i^0 + \sum_{j\in i} g_j^{MAP}\zeta_j\boldsymbol{\rho}_j\right|}. \quad (9)$$

$$g_j^{MAP} = \frac{u_j^0 v_j^0 + \zeta_j\left(\boldsymbol{\delta}_i^{MAP}\cdot\boldsymbol{\rho}_j\right)}{v_j^0+\zeta_j}. \quad (10)$$

These appear very similar to the general MAP solutions (ex. [6]), which interpolate uncertain ML estimates for small amounts of data by using prior parameters. That is to say, when ζ_j becomes small, MAP estimates $\boldsymbol{\delta}_i^{MAP}$ and g_j^{MAP} approach prior parameters $\boldsymbol{\nu}_i^0$ and u_j^0, respectively. Therefore, if we set $\boldsymbol{\nu}_i^0$ and u_j^0 from the parameters obtained by using a sufficient amount of data, $\boldsymbol{\nu}_i^0$ and u_j^0 mitigate the uncertain estimation effect caused by small amounts of data. Conversely, when ζ_j becomes large, MAP estimates $\boldsymbol{\delta}_i^{MAP}$ and g_j^{MAP} approach ML estimates $\boldsymbol{\delta}_i^{ML}$ and g_j^{ML}, respectively, which guarantees the appropriate estimation for large amounts of data. These limits for large and small amounts of data show the validity of the MAP solution.

4. IMPLEMENTATION USING TREE STRUCTURE

In CFT, the most important issue is how to appropriately provide a tied Gaussian structure for the coarse and fine classes according to the amount of data. In this paper we adopt a tree structure, which organizes hierarchical Gaussian classes [2], in the CFT implementation, and represents the coarse/fine classes efficiently. In [2], they construct a binary tree, whose nodes hold several Gaussians from an initial acoustic model in advance. Then, from the adaptation data, they obtain an occupancy count of a node as the sum of the all occupancy counts assigned to Gaussians in the node by using the Viterbi algorithm. By pruning the child nodes if their occupancies are less than a manually set occupancy threshold, they obtain a set of leaf nodes, and regard them as classes of transfer vectors in Eq. (1). Thus, they realize the transfer vector adaptation called Autonomous Model Complexity Control (AMCC), which is effective for any amount of adaptation data.

Table 1. Experimental conditions for speaker adaptation

Sampling rate/quantization	16 kHz / 16 bit
Feature vector	12 order MFCC with energy
(39 dimensions)	+Δ+ΔΔ
Window	Hamming
Frame size/shift	25/10 ms
Num. of states	3 (Left to right)
Num. of phoneme categories	43
Num. of context-dependent HMM states	1,000
Num. of mixture components	16
Initial training data	read sentences, 10.2 hours (44 males) †
Adaptation data	1st-half lectures, 320 utterances (10 males) ‡
Test data	2nd-half lectures, 13,162 words (10 males) ‡
Language model	Standard trigram (made by CSJ transcription)
Vocabulary size	30, 000
Perplexity	82.2
OOV rate	2.1 %

† ASJ (Acoustical Society of Japan) database

‡ CSJ (Corpus of Spontaneous Japanese) database

We also use occupancy thresholds to obtain coarse and fine classes in that tree structure. A scaling factor in a fine class has only one free parameter, and is well estimated even when the amount of adaptation data is smaller than that of a direction vector in a coarse class. Therefore, we set a smaller occupancy threshold for the fine class, and a larger occupancy threshold for the coarse class. In this tree structure, it is certain that small occupancy nodes are included in large occupancy nodes. Therefore, we can satisfy the hierarchical inclusion relation for the coarse, fine, and individual Gaussian classes based on this tree structure representation.

5. SUPERVISED SPEAKER ADAPTATION EXPERIMENT

We conducted experiments to show the effectiveness of the CFT adaptation within the transfer vector adaptation scheme. We compared CFT-ML and the conventional transfer vector adaptation described in Eq. (1) (AMCC) [2] in terms of the improvement in adaptation accuracy to show the effectiveness of the directional coarse/fine representation. We also compared CFT-ML and CFT-MAP to show the effectiveness of the CFT Bayesian solution. Table 1 summarizes the experimental conditions. The initial (prior) acoustic model was constructed from read sentences and we adapted this model using 10 lectures spoken by 10 males with their transcriptions. In this task, the mismatch between the training and adaptation data is caused not only by the speakers, but also by the difference in speaking styles between read speech and a lecture. Then, the Gaussian tree structure described in Section 4 was constructed from the initial acoustic model. The total number of leaf nodes in the tree was 16,000. By setting the occupancy thresholds (=10, 30, and 50) with reference to the result reported in [2], we obtained tied Gaussian classes for each fraction of the adaptation data, and used them as coarse classes in CFT, and as transfer vector classes in AMCC. For fine classes in CFT, we always used occupancy threshold = 1 because we found that this value was not very sensitive to the recognition performance in the preliminary investigation. As regards the adaptation and test data, each lecture was divided in half based on the utterance units, and the first half of the lecture (320 utterances) was used as adaptation data and the second half (13,162 words) was used as recognition data. The total amount of adaptation data consisted of more than 32 utterances for each male, and 1, 2, 4, 8, 16, and 32 utterances were used as adaptation data. As a result, 6 sets of adapted acoustic models for several amounts of adaptation data were prepared for each male.

Figures 3 through 5 compare word error rates obtained by CFT-

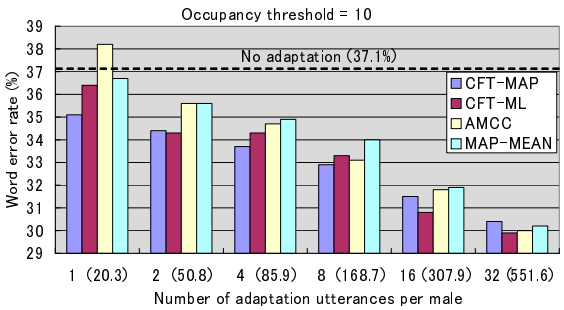

Fig. 3. Comparison of CFT-MAP, CFT-ML, and AMCC (occupancy threshold = 10).

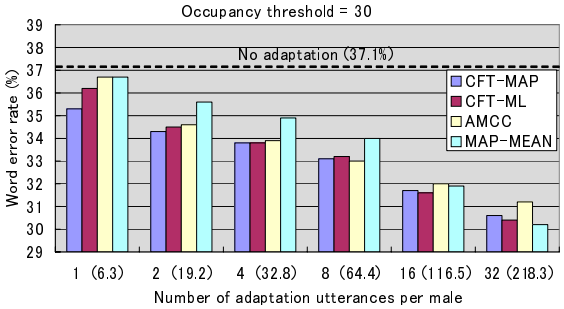

Fig. 4. Comparison of CFT-MAP, CFT-ML, and AMCC (occupancy threshold = 30).

MAP, CFT-ML, and AMCC for each occupancy threshold (10, 30, and 50). We also add the word error rates of the conventional MAP adaptation, which estimates only the mean parameters of individual Gaussians (MAP-MEAN) [6]. The average number of coarse classes (or AMCC classes) per male increased as the amount of adaptation data increased, and the performance of CFT-MAP, CFT-ML, and AMCC was almost always better than MAP-MEAN for any amount of data, which shows the effectiveness of the autonomous model complexity control, described in Section 4. For a large amount of data (more than 16 utterances) CFT-ML was always better than AMCC by up to 1.0 %. This is because CFT-ML has more free parameters than AMCC by the number of scaling factors, which provides an appropriate resolution of the parameter representation. This shows the effectiveness of the directional coarse/fine representation especially for large amounts of adaptation data. Next, we compared CFT-MAP with CFT-ML and AMCC, where prior parameters $\boldsymbol{\nu}_0$ and u_0 were set respectively by a unit direction vector and a scaling factor in the root node, which is the coarsest class in the tree structure. The other prior parameters were set at $\kappa_0 = 100$ and $v_0 = 1.0$. For a small amount of data (fewer than 2 utterances), CFT-MAP was superior to CFT-ML by up to 1.3 %, and to AMCC by up to 3.1 %. In such cases, the performance of CFT-ML and AMCC deteriorated as the occupancy threshold decreased where the amount of data per free parameter also decreased. This indicates that transfer vectors in AMCC and unit direction vectors in CFT-ML might be estimated incorrectly due to over-training. In CFT-MAP, such over-training effects were mitigated by the root node parameters based on a sufficient amount of data, which leads to the superiority of CFT-MAP. However, for the large amount of data, the performance of CFT-MAP was worse than that of CFT-ML, due to the effect of the root node parameters being estimated using too coarse a class. Therefore, we must consider an appropriate prior setting for CFT-MAP that includes the advantage of the directional coarse/fine representation even when there are large amounts of data.

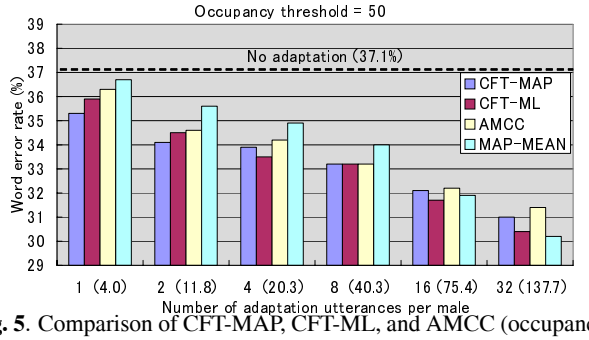

Fig. 5. Comparison of CFT-MAP, CFT-ML, and AMCC (occupancy threshold = 50).

6. SUMMARY

In this paper we introduced a new representation of the transfer vectors in acoustic model adaptation by using the Coarse/Fine Training of transfer vectors (CFT) based on directional statistics. Our adaptation scheme provides a new paradigm for acoustic modeling based on directional statistics. We showed the effectiveness of CFT in speaker adaptation experiments for large amounts of adaptation data, and of the Bayesian CFT (CFT-MAP) based on the von Mises distribution for small amounts of adaptation data. However, we found that the setting of prior parameters should be improved in CFT-MAP, and we will incorporate the structural MAP approach [9] for this purpose in the future. We will also try to apply CFT to MLLR [5], because the CFT representation can be applied to other transformative approaches in addition to the transfer vector approach presented in this paper.

7. REFERENCES

[1] M. Rahim and B-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on SAP*, vol. 4, pp. 19–30, 1996.

[2] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous control using tree structure," in *Proc. EUROSPEECH95*, 1995, pp. 1143–1146.

[3] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," in *Proc. ICASSP1995*, 1995, vol. 1, pp. 688–691.

[4] J. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation," in *Proc. ICASSP1995*, 1995, vol. 1, pp. 696–699.

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[6] J-L. Gauvain and C-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on SAP*, vol. 2, pp. 291–298, 1994.

[7] S. Watanabe and A. Nakamura, "Acoustic model adaptation based on coarse-fine training of transfer vectors and its application to speaker adaptation task," in *Proc. ICSLP2004*, 2004, vol. 4, pp. 2933–2936.

[8] K. V. Mardia, *Statistics of directional data*, Academic Press Inc., 1972.

[9] K. Shinoda and C-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. on SAP*, vol. 9, pp. 276–287, 2001.