IDIAP RESEARCH REPORT

Speaker Normalization using HMM2

Shajith Ikbal [a,b] Katrin Weber [a,b]
Hervé Bourlard [a,b]

IDIAP–RR 02-15

April 24, 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny, Switzerland.
[b] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

IDIAP Research Report 02-15

Speaker Normalization using HMM2

Shajith Ikbal Katrin Weber Hervé Bourlard

April 24, 2002

Abstract. In this paper, we present an HMM2 based method for speaker normalization. Introduced as an extension of Hidden Markov Model (HMM), HMM2 [2, 3] differentiates itself from the regular HMM in terms of the emission density modeling, which is done by a set of state-dependent HMMs working in the feature vector space. The emission modeling HMM aims at maximizing the likelihood through optimal alignment of its states across the feature components. This property makes it potentially useful to speaker normalization, when applied to spectrum. With the alignment information we get, it is possible to normalize the speaker related variations through piecewise linear warping of frequency axis of the spectrum. In our case, (emission modeling) HMM based spectral warping is employed in the feature extraction block of regular HMM framework for normalizing the speaker related variabilities. After a brief description of HMM2, we present the general approach towards HMM2-based speaker normalization and show, through preliminary experiments, the pertinence of the approach.

Acknowledgements: The authors Shajith Ikbal and Katrin Weber are supported for their work by the Swiss National Science Foundation project grants FN 2001-061325.00/1 and FN 2000-059169.99/1, respectively.

1 Introduction

State-of-the-art speech recognition systems developed using Hidden Markov Models (HMMs) [1] suffer from excessive sensitivity to various kinds of variabilities generally observed in the feature vectors. One of the major sources of such variabilities is the speaker differences. Speaker differences arise because of the differences in vocal tract shapes for different speakers, which basically result in differences in the formant structure of the spectrum. Techniques developed to handle these speaker related variabilities fall into one of the two classes, namely *adaptation techniques* and *normalization techniques*. In adaptation techniques the idea is to adjust the parameters of speaker independent models using some adaptation data so that the models would match better with the speaker of the test utterance. Maximum Likelihood Linear Regression (MLLR) [9, 10] is an example for the adaptation techniques. Normalization techniques try to remove the speaker related variations from the feature vectors so that during training the models generated would be sharper, and during recognition test utterances would better match with the models. Linear vocal tract normalization (LVTN) [7, 8] is a particular case of normalization technique. As discussed in the present paper, HMM2 [2, 3] can also be used as a normalization technique.

The effect of speaker differences in the spectral domain is the rescaling of the frequency axis, which stated in other words is the warping of the spectrum along the frequency axis. The LVTN method assumes that such warping is linear, by assuming that the speakers differ mainly by their vocal tract lengths, and tries to normalize the speaker differences by warping the frequency axis linearly. If $S(w)$ represents the unwrapped spectrum, a single warping factor α is used to obtain linearly warped spectrum $S(\alpha w)$. The warping factor α for each utterance is estimated through a maximum likelihood procedure [7], which is described mathematically as follows: Let λ denote the parameter set of speaker normalized model and W denote the transcription of utterance for which optimal α is to be found out. If $S_t(\alpha w)$ denotes the linearly warped power spectrum estimated from t^{th} time frame of the utterance and $\mathbf{x}_t^\alpha$ denotes the cepstral vector derived from $S_t(\alpha w)$, then the optimal warping factor $\hat{\alpha}$ is estimated as,

$$\hat{\alpha} = \underset{\alpha}{\text{argmax}} \; P(\mathbf{x}_0^\alpha, \mathbf{x}_1^\alpha, ..., \mathbf{x}_{T-1}^\alpha | \lambda, W) \tag{1}$$

where T denotes the length of the utterance.

But, linear warping of the spectrum is a suboptimal solution. A better solution would be to perform nonlinear warping. HMM2 provides good scope for performing this when the emission modeling HMMs (originally employed to compute the emission probabilities) are applied to the spectrum. They tend to optimally align the similar parts of the spectrum to maximize the likelihood, implicitly yielding an optimal warping function $f(.)$ as,

$$f^*(.) = \underset{f(.)}{\text{argmax}} \; P(S(f(w))|\lambda_i) \tag{2}$$

where λ_i denote the parameters of the emission modeling HMM. This warping function can be used to nonlinearly warp the spectrum into $S(f^*(w))$. Actually, the function $f^*(.)$ obtained using HMM2 is a piecewise linear warping function, with variable warping complexity.

In the next section, we give a brief introduction to the HMM2 formalism. In section 3, we discuss how HMM2 could be used to perform speaker normalization. In Section 4, we discuss preliminary experiments and observations made at the intermediate stages of the proposed system. In Section 5, we further illustrate the approach by discussing a previous work done using HMM2 for formant-like feature extraction. In this case, HMM2 has been employed in a manner similar to that of the present approach, for extracting formant-like features, and has been shown to yield impressive performance.

2 HMM2

Introduced as an alternative to the regular HMM, HMM2 [2, 3] is basically a statistical approach for acoustic modeling of the speech signal, aiming at simultaneously modeling the temporal and frequency structures of the signal.

As illustrated in Figure 1, HMM2 is built up from conventional HMMs where the usual multi-Gaussian densities associated with each HMM state are replaced by frequency-based HMMs, called *frequency HMMs*, to model the emission density[1].

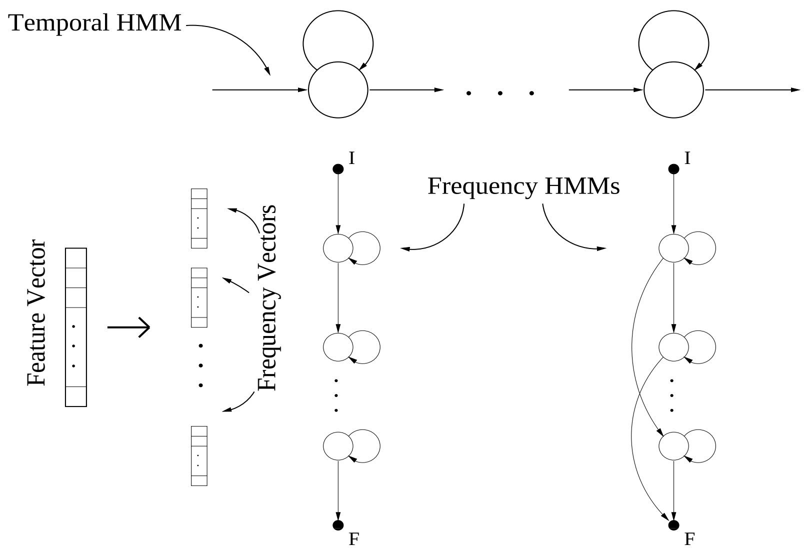

Figure 1: Typical architecture of HMM2.

The frequency HMMs treat feature vectors as sequences and estimate the emission probability by calculating the likelihood of feature vectors being generated by them. For this purpose, each feature vector is converted into a sequence of smaller vectors called *frequency vectors*, as illustrated in the Figure 1, and the frequency states belonging to the frequency HMM are assumed to have emitted those vectors. In the simplest case, these *frequency vectors* could simply be the frequency components. As will be explained later, these frequency vectors are usually appended with the corresponding feature component index to restrict the frequency HMM states from behaving in an unconstrained manner [6]. Let $\mathbf{x}_t$ and q_t denote respectively the feature vector and temporal state at time t. If $\{\mathbf{x}_{t,0}, \mathbf{x}_{t,1}, ..., \mathbf{x}_{t,s}, ...\mathbf{x}_{t,S-1}\}$ denotes the frequency vector sequence derived from $\mathbf{x}_t$, then the likelihood of a sample frequency state sequence $R = \{r_0, r_1, ..., r_s, ..., r_{S-1}\}$ of the frequency HMM belonging to the temporal state q_t, generating the vector sequence is

$$p(\mathbf{x}_t, R|q_t) = p(r_0|I, q_t)p(\mathbf{x}_{t,0}|r_0, q_t) \prod_{s=1}^{S-1} p(r_s|r_{s-1}, q_t)p(\mathbf{x}_{t,s}|r_s, q_t) \tag{3}$$

where $p(r_0|I, q_t)$ denotes the initial probability of frequency state r_0 belonging to the temporal state q_t, $p(r_s|r_{s-1}, q_t)$ denotes the probability of performing transition from the frequency state r_{s-1} to the state r_s at the temporal state q_t, and $p(\mathbf{x}_{t,s}|r_s, q_t)$ denotes the probability of frequency state r_s belonging to the temporal state q_t emitting the frequency vector $\mathbf{x}_{t,s}$. The probability of the frequency states

[1]In this paper, we call the emission modeling HMMs with name *frequency HMMs* as it suits the present context well, though names such as *internal HMMs* or *feature HMMs* have been used in the previous works.

emitting the *frequency vectors* is modeled by lower dimensional Gaussian Mixture Models (GMM). Based on the topology of the frequency HMM for temporal state q_t, if $\mathcal{D}_t$ represents the set of all possible frequency state sequences that could have generated the frequency vector sequence, then the emission probability calculated using the frequency HMM is

$$p(\mathbf{x}_t|q_t) = \sum_{R \in \mathcal{D}_t} p(\mathbf{x}_t, R|q_t) \tag{4}$$

Alternatively, the emission probability can also be computed using the well known Viterbi approximation as,

$$p(\mathbf{x}_t|q_t) = \max_{R \in \mathcal{D}_t} \; p(\mathbf{x}_t, R|q_t) \tag{5}$$

The parameter set of HMM2 contains, the transition probabilities of all the temporal states, transition probabilities of all the frequency states belonging to all temporal states, and the parameters of GMMs assigned to each frequency states of each temporal state. A derivation of the EM algorithm to estimate these parameters is given in [2]. An explanation of HMM2 from the implementation point of view, including its training and recognition, is given in [4]. Other than speaker normalization, which is the topic of discussion in the present paper, HMM2 has recently been shown to be useful for extracting formant-like features [5].

3 Speaker Normalization using HMM2

Assume that we have a top-down frequency HMM which is trained with spectra obtained from speech signals of different speakers uttering the same sound. If this frequency HMM is Viterbi aligned against a spectrum, which again corresponds to the same sound but obtained from the speech signal of a new speaker, each frequency state would get aligned to those regions of the spectrum which it has learned during the training. Let S denote the length of the spectrum, and $R = \{r_0, r_1, ..., r_s, ..., r_{S-1}\}$ denote the frequency HMM state sequence obtained as a result of the Viterbi alignment. Let us also assume that there is a speaker normalized spectrum for the same sound, whose Viterbi alignment against the frequency HMM would have yielded the state sequence $R^{'} = \{r^{'}_0, r^{'}_1, ..., r^{'}_s, ..., r^{'}_{S-1}\}$. The difference between the sequences R and $R^{'}$ basically corresponds to the speaker related differences of the speaker-dependent spectrum with respect to the speaker normalized spectrum. Actually, the difference gives information about the warping function $f(.)$ required to transform the speaker-dependent spectrum to the speaker-normalized spectrum. For example, if the states $\{r_s, r_{s+1}, ..., r_{s+n}\}$ and $\{r^{'}_{s'}, r^{'}_{s'+1}, ..., r^{'}_{s'+n'}\}$ from the respective sequences R and $R^{'}$ correspond to the same frequency state in the frequency HMM, then the spectral coefficients in the frequency range $\{s, s+1, ..., s+n\}$ need to get warped into the range $\{s^{'}, s^{'}+1, ..., s^{'}+n^{'}\}$.

The proposed method uses frequency HMMs at the feature extraction stage of the regular HMM framework, to perform state-dependent speaker normalization of the feature vectors. Each temporal state of the HMM is assigned a frequency HMM to normalize the feature vectors extracted from the frames corresponding to that state. The feature extraction, training, and recognition stages of the proposed system are explained in the following subsections.

3.1 Feature Extraction

Speaker normalized Mel Frequency Cepstral Coefficients (MFCC) are used as the feature vectors. These MFCC parameters are extracted from the speaker normalized power spectrum, by first mel-windowing the spectrum to get Filter Bank Coefficients (FBC) and then transforming the FBCs using Discrete Cosine Transform (DCT) to obtain MFCCs. The speaker normalized power spectrum is obtained by warping the frequency axis of the unnormalized spectrum piecewise linearly, using the frequency HMMs. As explained earlier, state sequences R and $R^{'}$, are used to perform the warping. As it is difficult to perform the frequency warping directly on the power spectrum, which involves interpolation of the energy values for the intermediate frequencies, it is implemented in an indirect manner at the FBC computation stage, as depicted in the Figure 2. The mel-windows are altered using the sequences R and $R^{'}$ and are employed on the unnormalized power spectrum to obtain the FBCs, which is equivalent to warping the power spectrum and using the unaltered mel-windows. For example, if the states $\{r_s, r_{s+1}, ..., r_{s+n}\}$ and $\{r^{'}_{s^{'}}, r^{'}_{s^{'}+1}, ..., r^{'}_{s^{'}+n^{'}}\}$ in the respective sequences R and $R^{'}$ correspond to the same state in the frequency HMM, then the mel-windows falling in the range of $\{s^{'}, s^{'}+1, ..., s^{'}+n^{'}\}$ of the frequency axis should be mapped to the range $\{s, s+1, ..., s+n\}$.

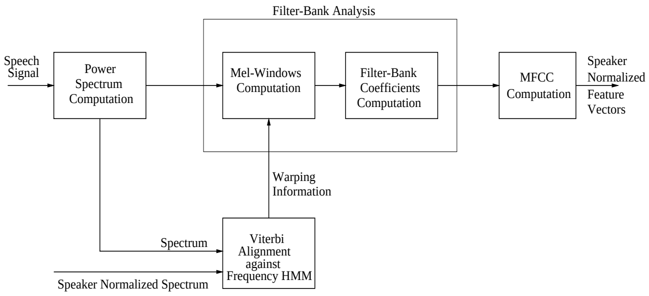

Figure 2: Implementation of HMM2 based Speaker Normalized Feature Extraction.

As we have seen, the main idea behind the proposed method is to align the similar regions of the spectrum by performing Viterbi alignment, and then to use the alignment information to warp the spectrum. For this matter, first of all it is desirable to have a definition of different regions in the spectrum that will be most suitable for the speaker normalization problem. As the locations of peaks and valleys of the spectrum are considered to be important for speech recognition, we designed the topology of the frequency HMM (top-down) and number of states to facilitate segmentation in terms of those peaks and valleys. In order to have Viterbi alignment work properly, these regions should be distinguishable in terms of the energy, where as in the raw spectrum they are highly indistinguishable in terms of the energy. For this reason, a modified version of the spectrum called 'smoothed-differenced' spectrum [11] is used instead of the raw spectrum. In 'smoothed-differenced' spectrum the energy values corresponding to each frequency index are replaced by an estimate of slope at that point. This makes the spectral regions separated in terms of their peaks and valleys to have either positive or negative values, i.e., piecewise stationary.

3.2 Training

Training is a three step procedure. The first step involves training of the a regular HMM system without any speaker normalization done at the feature extraction stage, the second step the training of the frequency HMMs, and third step the training of the HMM system with speaker normalization done at the feature extraction stage.

The models generated during the first step are used mainly for obtaining the state-level segmentation of the training utterances, that is needed during the second step. In addition they also serve as the baseline system.

In the second step, frames from all the training utterances corresponding to each temporal state of the HMM are collected using the segmentation information, and the spectra derived from them are used to train frequency HMM corresponding to that temporal state. This results in piecewise segmentation of all the spectra and the learning of frequency state density functions.

These frequency HMMs are then used in the third step for performing state-specific speaker normalization (2) at the feature extraction stage. The state sequence $R^{'}$ for a particular temporal state, which is also needed along with R to perform the warping, is used as the mean of all the R obtained from frames corresponding to that state.

3.3 Recognition

Recognition involves state-specific spectral warping, done at the feature extraction stage for all the frames in the test utterance, to obtain state-specific feature vectors. Taking a closer look, there is a good possibility for spectral warping resulting in inter-state confusion, though the actual goal of its use is to normalize the speaker related variations. For example, the spectra corresponding to two different vowels also differ mainly in terms of the formant frequencies. This means that it is possible to warp the spectrum corresponding to one vowel to the spectrum corresponding to other. As the frequency HMM, employed to perform state-specific spectral warping, acts in an unconstrained manner, it may very well transform the spectrum of some state to the spectrum of the state for which it is used. To avoid this, during training, from all the sequences R obtained using the training utterances, mean and variance of the frequency indices [6] corresponding to each frequency state in the frequency HMM are computed. Then those values are used in recognition to constrain the frequency states to stay within a particular range of frequency indices during Viterbi alignment. This should avoid the inter-state confusion as the amount warping required to transform the spectrum of one speaker to the other is less as compared to the amount warping required to transform spectrum of one sound to the other.

4 Study of the System

The proposed method achieves speaker normalization through piecewise linear warping of the spectrum. The warping is based on the sequence R which is obtained as a result of Viterbi alignment of the unnormalized spectrum against the frequency HMM. In order for this to work reliably, the frequency HMM should be able to segment the test spectrum reliably into defined regions, during the Viterbi alignment. So when a 'smoothed-differenced' spectrum is Viterbi aligned against frequency HMM, the segmentation obtained should be in terms of the peaks and valleys of the corresponding spectrum. To confirm this, we have checked out several Viterbi alignments done by the frequency HMMs and have seen that it is indeed happening. Figure 3 shows an example segmentation obtained as a result of alignment of a frequency HMM belonging to certain state against a spectrum which also is from the same state. The spikes in the figure show the segmentation obtained.

During recognition, we do not know a priori the state which a particular frame belongs to. As a result, as explained earlier, state-dependent feature vectors are extracted from each frame for all the states. This requires Viterbi alignment of the test spectrum against the frequency HMMs of all the states. In this case, as only one alignment is genuine alignment, and all other alignments should yield improper segmentation. Figure 4 shows an example of this. The frequency HMM of Figure 3, is aligned against a spectrum which belongs to a different state. As we can see from the figure, the segmentation obtained is improper, which in turn would affect the feature extraction and may improve the discrimination.

5 Formant-like Feature Extraction using HMM2

In this section, we discuss a previous work [6] where HMM2 has been used for extracting formant-like features and shown to yield impressive results. We have chosen to discuss this here because it is closely related to what we are doing now for the speaker normalization, and further illustrate

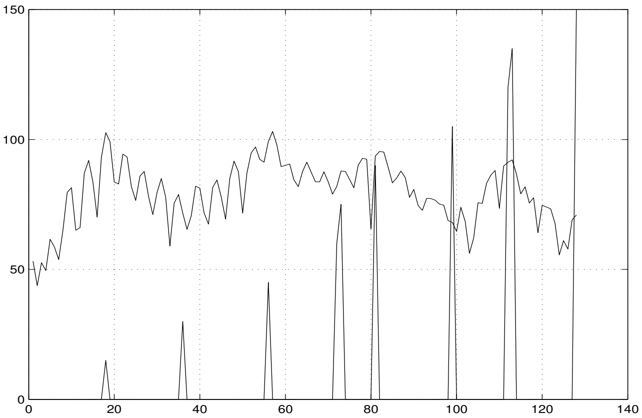

Figure 3: Segmentation done by the frequency HMM.

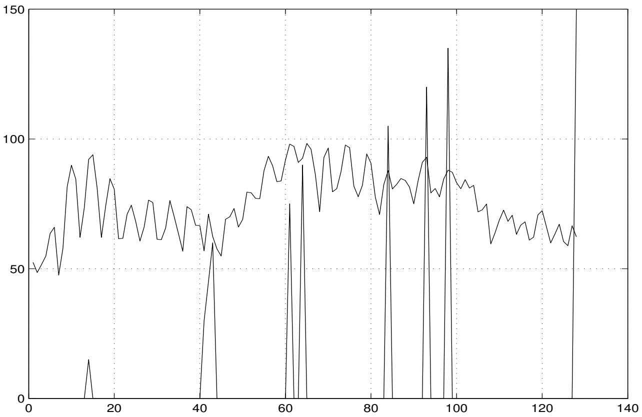

Figure 4: Improper segmentation done by the frequency HMM, when presented with a spectrum belonging to a state other than the state for which it is trained.

the potential of HMM2 based speaker normalization. Both the approaches employ frequency HMMs to perform state-dependent segmentation of the spectrum. In the formant-like feature extraction problem, the segmentation obtained is used to estimate the formant-like frequencies, where as in the speaker normalization the segmentation is used to perform piecewise linear warping of the spectrum and hence to normalize the speaker related variations.

In the formant-like feature extraction work, filter banked spectrum is used instead of the raw spectrum [6]. The segmentation obtained as a result of Viterbi alignment of the differenced filter banked spectrum against the frequency HMM are used to estimate the formant values. In fact, the boundaries between the segmented regions are used directly as the formant-like frequencies. Figure 5 shows an example of the formant tracking done on a test utterance. The vertical lines indicate the state-level segmentation of test utterance. The formant-like frequencies corresponding to frame of a particular temporal state is obtained using the frequency HMM of the state. The horizontal lines show the segmentation tracks extracted using the state-dependent frequency HMMs.

The over all reliability of the formant-like frequencies estimated has been checked by using them

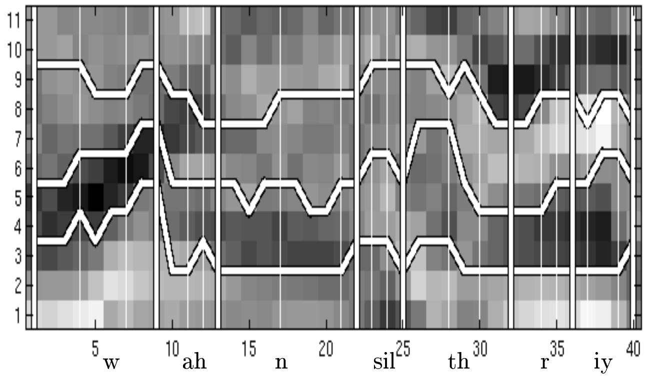

Figure 5: An illustration of formant tracking done by the state-dependent frequency HMMs.

as feature vectors in the regular HMM system. A speech database, called Numbers 95, containing speaker-independent free format numbers spoken over telephone is used for this purpose. Using 3 formant-like frequencies extracted from each frame as a feature vector, a recognition performance of **81.4%** is achieved on the Numbers 95 database. This is reasonably comparable to the performance of state-of-the-art systems, as dimension of the feature vectors used is only 3. This result basically shows that the segmentation obtained using the frequency HMMs carries reliable and meaningful information.

6 Conclusion

In this paper, we have presented a new approach for speaker normalization using HMM2. The state-specific feature HMMs, actually used to compute the emission probabilities in the HMM2, are employed at the feature extraction stage of the regular HMM framework to perform state-dependent speaker normalization. The speaker normalization is done by warping the spectrum piecewise linearly based on the frequency state sequence R, which is obtained as a result of Viterbi alignment of the speaker-dependent spectrum against the frequency HMM. A study of intermediate stages of the resulting system is presented. While HMM2 has already been used quite successfully in other frameworks, the present work shows its potential to further improvements.

References

[1] Bourlard H. and Morgan N. (1993). "Connectionist Speech Recognition: A Hybrid Approach". *The Kluwer International Series in Engineering and Computer Science*, Kluwer Academic Publishers, Boston, USA. Vol. 247, 1993.

[2] Bengio S., Bourlard H., and Weber K. (2000). "An EM Algorithm for HMMs with Emission Distributions Represented by HMMs". *IDIAP Research Report*, IDIAP, Martigny, Switzerland. IDIAP-RR 00-11, May 2000. ftp://ftp.idiap.ch/pub/reports/2000/rr00-11.ps.gz.

[3] Weber K., Bengio S., and Bourlard H. (2000). "HMM2 - A Novel Approach to HMM Emission Probability Estimation". in *Proc. of ICSLP*. Vol. 3, 147-150. Oct. 2000. ftp://ftp.idiap.ch/pub/reports/2000/rr00-30.ps.gz.

[4] Ikbal S., Bourlard H., Bengio S., and Weber K. (2001). "IDIAP HMM/HMM2 System: Theoretical Basis and Software Specifications", *IDIAP Research Report*, IDIAP, Martigny, Switzerland. IDIAP-RR 01-27, Oct. 2001. ftp://ftp.idiap.ch/pub/reports/2001/rr01-27.ps.gz

[5] Weber K., Bengio S., and Bourlard H. (2001). "HMM2 - Extraction of Formant Structures and Their Use for Robust ASR". in *Proc. of Eurospeech*, Aalborg, Denmark. 607-610, Sep. 2001.

[6] Weber K., Bengio S., and Bourlard H. (2001). "Speech Recognition using Advanced HMM2 Features". in *Proc. of IEEE ASRU Workshop*. Dec. 2001. ftp://ftp.idiap.ch/pub/reports/2001/rr01-24.ps.gz.

[7] Lee L., and Rose R.C. (1996). "Speaker Normalization Using Efficient Frequency Warping Procedures". in *Proc. of ICASSP*, Atlanta, Georgia, USA. vol. 1, 353-356, 1996.

[8] Wegmann S., McAllaster D., Orioff J., and Perkin B. (1996). "Speaker Normalization in Conversational Telephone Speech". in **Proc. of ICASSP**, Atlanta, Georgia, USA. vol. 1, 339-341, 1996.

[9] Leggetter C. J., and Woodland P. G. (1994). "Speaker Adaptation of Continuous Density HMMs Using Linear Regression". in **Proc. of ICSLP**, Yokohama, Japan. vol. 2, 451-454, 1994.

[10] Dijalakis V. V., Rtischev D., and NeuMeyer L. G. (1995). "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures". IEEE Trans. on Speech and Audio Processing. vol. 3, No:5, 357-365, 1995.

[11] Nadeu C. (1999). "On the Filter-Bank based Parameterization Front-End for Robust HMM Speech Recognition". in **Proc. of Robust'99**. 235-238, May 1999.