A New HMM Adaptation Approach for the Case of a Hands-free Speech Input in Reverberant Rooms

Hans-Günter Hirsch, Harald Finster

Niederrhein University of Applied Sciences, Krefeld, Germany
hans-guenter.hirsch@hs-niederrhein.de

ABSTRACT

A new method is presented for adapting the HMMs of a speech recognition system to the condition of a hands-free speech input in a room environment. The reverberation in a room usually has a bad effect on the performance of a recognition system. Reverberation causes an artificial extension of acoustic excitations what gets visible as so called reverberation tail when looking at the envelope of the short-term energy over the whole frequency range or in subbands.

The approach is based on the assumption that the acoustic excitation of a speech segment, as modeled by an HMM state, will be seen as attenuated versions at successive HMM states. Adding this attenuated excitations in the spectral domain at each HMM state leads to a considerable improvement of the recognition performance.

Furthermore a new approach is presented to adapt the Delta parameters that are usually taken as additional acoustic features. The efficiency of both new techniques has been proved by some experiments on isolated and connected word recognition with the TIDigits speech data base.

Index Terms: robust speech recognition, HMM adaptation, hands-free speech input, reverberation

1. INTRODUCTION

The main effort for improving the performance and the robustness of speech recognition systems in real application scenarios is spent on the development of techniques to compensate the influence of background noise and of an unknown frequency weighting due to e.g. the characteristics of the microphone. This compensation approach is either be realized as a robust feature extraction or by an adaptation of HMMs (Hidden Markov Models).

In many applications it would be desirable to allow a hands-free speech input without the need of wearing a close-talking microphone. This would make the use of recognition systems more comfortable. But the hands-free speech input leads to a major deterioration of the recognition performance due to the influence of the room acoustics. Only a few investigations (e.g. [1],[2],[3]) have been carried out on compensating the influence of a hands-free speech input in rooms so far.

A new technique will be described in the next section to adapt the spectral and energy parameters of HMMs to speech data, that have been recorded in a reverberant environment. This new approach can be applied to whole-word HMMs as well as to triphone models. Furthermore the technique can easily be combined with an existing approach [4] for adapting HMMs to stationary background noise and unknown frequency characteristics. Results of some first recognition experiments will be presented to show the gain in recognition performance.

2. NEW ADAPTATION APPROACH

The influence of a hands-free speech input in a room can be modeled as a superposition of the original signal and delayed and attenuated versions of this signal. These delayed versions are caused by multiple reflections at the walls or at any object in the room. The effect is called reverberation. The transmission between a speaker and a microphone in a room can be modeled by convolving the speech signal with a room impulse response. But this impulse response is usually time variant in case the speaker moves or the conditions in the room change due to e.g. opening a door or a window or other people moving in the room. The estimation of the room impulse response is a quite difficult task because of the fairly high length of the impulse response and the time variant behavior.

To derive the new approach, the modeling of speech segments is considered as it is done with a single state of a HMM. A speech segment is usually described by a distribution density function for the spectrum and the frame energy. In a lot of realizations the spectrum is defined by a set of cepstral parameters because the cepstral coefficients proved to be less correlated in comparison to spectral coefficients. Under this assumption a single state of a HMM models a segment by a set of individual Gaussian distributions for each acoustic parameter. We will focus on the means of these parameters in our approach.
The average duration of a speech segment can be estimated from the transition probability for remaining in a certain state. The duration is usually in the range of 20 to 100 ms. This makes it obvious that the detailed description by a room impulse response is not really needed when thinking about adaptation approaches on the basis of HMM modeling.

The idea of the new approach is based on the occurrence of the acoustic excitation, as defined within a single HMM state, as attenuated versions at later HMM states. These attenuated versions of the acoustic excitations from previous states will superpose the acoustic excitation of an observed HMM state. The acoustic excitation is described by the means of the spectral parameters and the frame

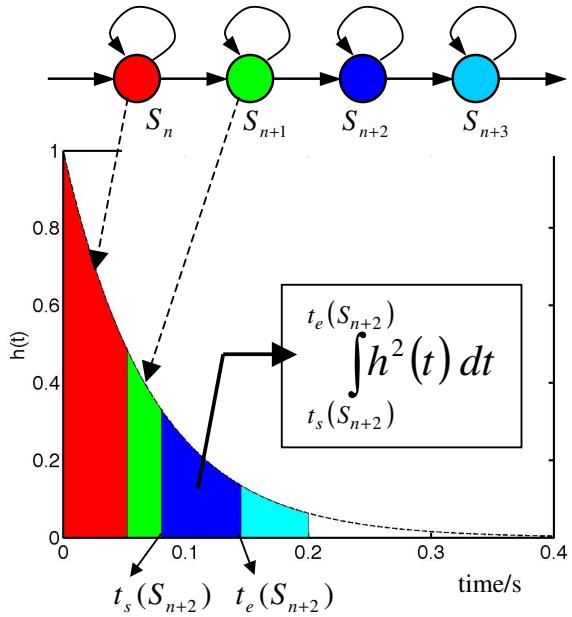

Figure 1: Determination of weighting coefficients

energy. The weighting coefficients defining the individual attenuations are derived from the description of the room impulse response as an exponentially decaying characteristic. This is visualized in figure 1.

The reverberation time T60 is needed as the only parameter for defining the exponential characteristic where a value of 500 ms is chosen for the exemplary curve in figure 1. Four HMM states are considered in this figure. Their average durations can be estimated from the transition probabilities to remain in the corresponding state. The durations are used for defining the length of the corresponding segments in the exponential characteristic. In general the energy weighting coefficients can be calculated as

$$\alpha(n+i,n) = \int_{t_s(S_{n+i})}^{t_e(S_{n+i})} h^2(t)\,dt \qquad (1)$$

where n is the index of an HMM state and $n+i$ is the index of the later state for which the energy contribution of the acoustic excitation at state n is calculated. $t_s(S_{n+i})$ and $t_e(S_{n+i})$ are the corresponding start and end time of state S_{n+i} assuming a time measuring starting at the beginning of state S_n.

$h^2(t)$ is normalized so that $\int_0^\infty h^2(t)\,dt = 1$

The weighting coefficients for the example in figure 1 describe in terms of spectral energy how much of the acoustic excitation at state S_n will be seen in the later states S_{n+1} to S_{n+3}. The coefficients have to be individually calculated for each HMM state due to the different length of the segments that are modeled by the HMM states.

The weighting coefficients can be immediately used to adapt the frame energy of each HMM state by adding the corresponding contributions of the previous frames. In the same way the power density spectrum X can be adapted after transforming back the cepstral coefficients to the linear Mel spectral domain. The adaptation approach can be described as weighted sum of the spectrum $X(S_n, mix_j)$ at HMM state S_n and for the individual mixture component with index mix_j and the average spectra $\overline{X}(S_{n-i})$ of previous HMM states:

$$\tilde{X}(S_n, mix_j) = \alpha(n,n)\cdot X(S_n, mix_j) + \alpha(n,n-1)\cdot \overline{X}(S_{n-1}) + \alpha(n,n-2)\cdot \overline{X}(S_{n-2}) + \cdots$$

This way the power density spectra in the linear Mel domain are individually adapted at each state and for each mixture component by taking into account the attenuated average spectra of previous HMM states. These average spectra are derived from a set of average cepstral coefficients that are calculated as weighted sum over all mixture components with their individual mixture weights. The adapted spectra $\tilde{X}(S_n, mix_j)$ have to be transformed to the cepstral domain again.

The effect of adapting the spectral means is shown in figure 2 by looking at spectrograms that are derived from the cepstral means and the average segment durations of all HMM states. 3 different versions of the spectrogram are shown for the word "six" that have been derived from 3 different HMMs.

The HMMs consist of 16 states. A spline interpolation is applied to recover the spectrogram from the 16 states at a frame rate of 10 ms. The set of average cepstral means is taken for the transformation back to the spectral domain as they can be calculated from the cepstral means of several mixture components under consideration of the individual mixture weights.

The upper graph shows the spectrogram as derived from the HMM of the word "six" (s-ih-k-s) where the HMM has been created as output of a training with the clean TIDigits. We are looking at the end of the word where the spectral characteristics of the vowel and the high frequency contribution of the fricative at the end can be clearly seen.

The graph in the middle shows the spectrogram as derived from a HMM that has been trained on a modified version of the TIDigits training data. All training data has been processed with a tool for simulating the hands-free speech input in a room [5]. The exponential reverberation tails can be seen in this graph, most obvious for the vowel "ih". The "valley" between the vowel and the fricative representing the pause before the "k" does no longer exist.

Finally the spectrogram is shown in the lower graph after applying the adaptation technique to the clean HMM. The adaptation has been done with a fixed value for the reverberation time T60 as only parameter. The reverberation tails are also visible in this graph. In general the spectrogram of the adapted HMM shows a lot of similarities with the HMM trained on reverberant speech data. This indicates that the new adaptation approach seems to produce useful results.

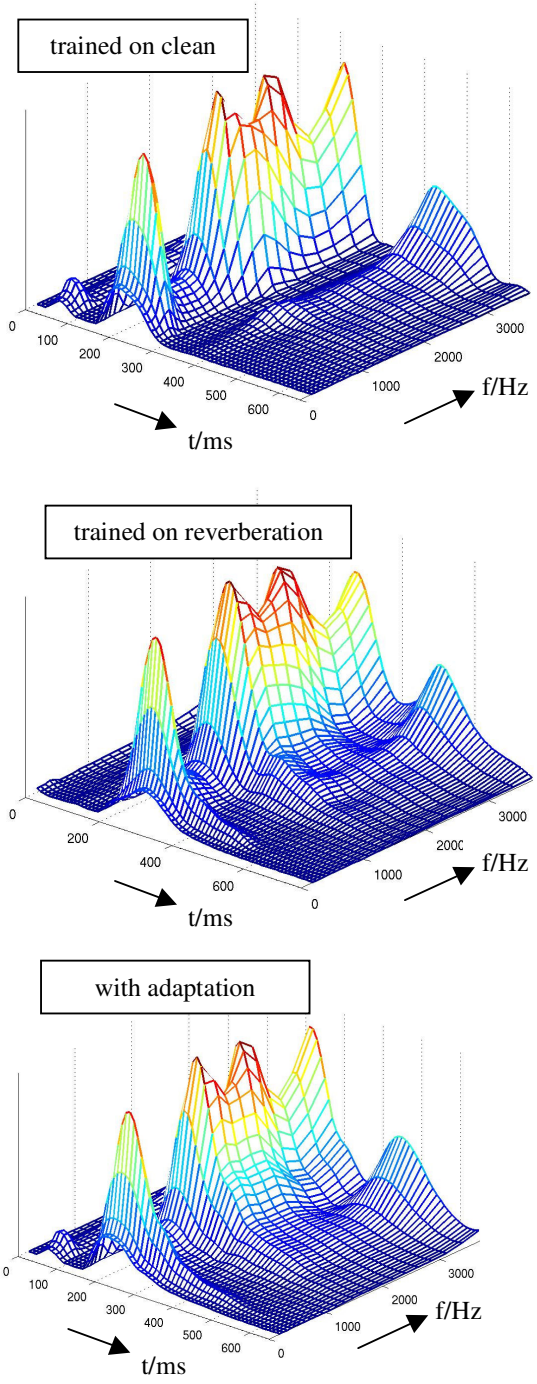

Figure 2: Spectrograms of the HMMs for the digit "six"

Comparing the spectrograms of the clean and reverberated versions it gets obvious that the trajectories at individual frequency bins look quite different due to the reverberation tails. The influence of reverberation can be described as a low-pass filtering in the modulation frequency domain [6]. Thus it should be also worthwhile to adapt the Delta and Delta-Delta coefficients.

Average Delta coefficients can be derived from the interpolated average spectrograms of the clean and the adapted HMMs as e.g. shown in figure 2. This can be realized by transforming the sequences of spectra to the cepstral domain. Due to the Spline interpolation a cepstrum is available every 10 ms. Thus the standard procedure for calculating the Delta coefficients can be applied on these sequences of cepstra as it is done in the feature extraction. Processing the average spectra of the clean and the adapted HMM this way, the average Delta cepstral coefficients are available for both conditions. The difference between these sets of average Delta coefficients is taken to adapt the Deltas of the clean HMM where the adaptation is individually done for each mixture component.

It turned out that best recognition results are not achieved when adding the difference completely but weighting the difference with a factor of about 0,7 before adding it. The Delta-Delta coefficients are adapted in the same way.

The only parameter needed for the adaptation is the reverberation time T60. In order to estimate T60 the following steps are performed. First the recognition of an utterance is performed with the current set of adapted HMMs. Then the set of clean models is adapted again several times by slightly varying the previously estimated value of T60. With each of this newly adapted model sets a forced recognition is carried out based on the result of the first recognition. This value of T60 and thus the new model set is selected, which leads to a maximum likelihood of the recognized sequence. Assuming no extreme change of the room acoustics the estimated T60 is lowered or increased by a maximum of 40 ms during this search for the maximum likelihood.

3. RECOGNITION EXPERIMENTS

A first series of recognition experiments has been run to verify the applicability of the new adaptation approaches. The well known TIDigits data base is taken for the recognition of single digits and sequences of digits based on the use of whole word HMMs. 13 Mel cepstral parameters including the zeroth cepstral coefficient C0 are extracted as acoustic features every 10 ms. The cepstral coefficient C0 is only needed to transform the cepstrum back to the Mel frequency domain for the adaptation processing. But C0 is not used for the recognition.

The feature vector for the Viterbi decoding consists of 39 parameters in total, including the Delta and Delta-Delta coefficients of the 12 cepstral parameters and the logarithmic frame energy as it can be calculated as sum of squared values from the speech samples. Gender

dependent HMMs with 16 states and 2 mixture components per state are calculated from the clean training data with the training tools of HTK [7]. Viterbi decoding and adaptation has been realized with own software modules.

The recognition of single digits only is considered as first recognition task. This avoids the superposition of the acoustic information at the beginning of a word by the acoustic information at the ending of the preceding word as it is the case when looking at fluently spoken sequences of digits.

A simulation tool [5] is applied to create versions of the TIDigits that have been recorded in a room with a reverberation time of about 600 ms. Word error rates are shown in figure 3 for the about 2500 single digits that are part of the TIDigits test data.

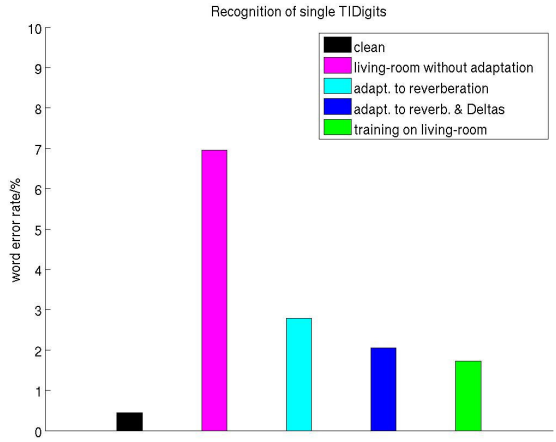

Figure 3: Word error rates for single TIDigits

The error rate for the clean data is 0,44 %. It increases to a value of about 7 % for the recognition of the reverberated digits. Adapting the static frame energy and the static spectral parameters, the error rate decreases to about 2,7 %. Adapting additionally also the Delta coefficients, the error rate is reduced further to a value of about 2 %. This is a first proof that both new approaches can help to improve the recognition performance in case of a hands-free speech input. The recognition performance with adaptation is in the same range as for the case of training the HMMs on reverberant data. For this experiment all TIDigits training data has been processed with the same simulation of a hands-free speech input as for the test data.

The word error rates for the recognition of all TIDigits test data are presented in figure 4 including also sequences of digits. Results are shown for three different conditions. These are the recognition of clean data, of data recorded in an office room with a reverberation time of about 0,4 s and of data recorded in a living room with a reverberation time of about 0,6 s. All data have been created with the already mentioned simulation tool.

It can be seen for both reverberant conditions that the error rate can be reduced by adapting the static parameters. A further improvement can be achieved by additionally adapting the Delta parameters.

The improvement is not as impressive as in the case of single digits because the effects of reverberation are more complex for fluently spoken sequences of words. Some speakers utter sequences of digits fast with coarticulation effects between the digits. Thinking about the further smearing by the reverberation it will be hard to model these effects in the adaptation approach.

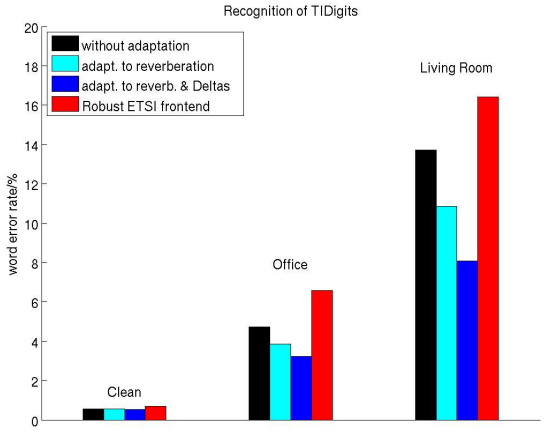

Figure 4: Word error rates for TIDigits

The recognition performance is compared to the case of applying the robust feature extraction scheme as standardized by ETSI [8] to training and test data. This front-end allows a robust recognition for the conditions of additive noise and unknown frequency characteristics. It turns out that the performance is even slightly worse for the recognition of the reverberated data in comparison to applying an usual cepstral analysis scheme. Training has also been done on the clean TIDigits.

4. REFERENCES

[1] B. Kingsbury, "Perceptually inspired signal processing strategies for robust speech recognition in reverberant environments", *dissertation at UC Berkeley*, USA, 1998.

[2] C.K. Raut, T. Nishimoto, S. Sagayama, "Model adaptation by state splitting of HMM for long reverberation", *Interspeech conference 2005*, pp. 277-280, Lisbon, Portugal, 2005.

[3] K. Kinshita, T. Nakatani, M. Miyoshi, "Efficient blind dereverberation framework for automatic speech recognition", *Interspeech conference*, pp. 3145-3148, Lisbon, Portugal, 2005.

[4] H.G. Hirsch, "HMM adaptation for applications in telecommunication", *Speech Communication 34*, pp. 127-139, 2001

[5] H.G. Hirsch, H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems", *Interspeech conference 2005*, pp. 2697-2700, Lisbon, Portugal, 2005.

[6] T. Houtgast, H.J.M. Steeneken, R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics", *Acustica*, Vol.46, pp 60-72, 1980

[7] http://htk.eng.cam.ac.uk

[8] ETSI standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm", *ETSI ES 202 050 v1.1.1 (2002-10)*, Oct. 2002.