IDIAP RESEARCH REPORT

HMM/ANN Based Spectral Peak Location Estimation for Noise Robust Speech Recognition

Shajith Ikbal [a,b] Hervé Bourlard [a,b]
Mathew Magimai.-Doss [a,b]

IDIAP–RR 04-50

September 20, 2004

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] IDIAP Research Institute, Martigny, Switzerland.
[b] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

IDIAP Research Report 04-50

HMM/ANN Based Spectral Peak Location Estimation for Noise Robust Speech Recognition

Shajith Ikbal Hervé Bourlard Mathew Magimai.-Doss

September 20, 2004

Abstract. In this paper, we present a HMM/ANN based algorithm to estimate the spectral peak locations. This algorithm makes use of distinct time-frequency (TF) patterns in the spectrogram for estimating the peak locations. Such an use of TF patterns is expected to impose temporal constraints during the peak estimation task, thereby yielding a smoother estimate of the peaks over time. Additionally, the algorithm use an ergodic topology for the HMM/ANN, thus allowing an estimation of a varying number of peak locations over time. The usefulness of the proposed algorithm is evaluated in the framework of a recently introduced noise robust feature called spectro-temporal activity pattern (STAP) feature. Interestingly, recently introduced, phase autocorrelation (PAC) spectrum, with enhanced spectral peaks and smoothed spectral valleys, turns out to be more appropriate for this algorithm than the regular spectrum.

Acknowledgements: The authors thank the Swiss National Science Foundation for the support of their work through grant MULTI: FN 2000-068231.02/1 and through National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The authors also thank DARPA for supporting through the EARS (Effective, Affordable, Reusable Speech-to-Text) project.

1 Introduction

Speech signal exhibits spectral and temporal amplitude modulations [1]. Typical speech recognition systems makes use of them by considering spectral representation of the speech signal over the entire span of the frequency axis and over a limited span of the temporal axis. However, this is in quite contrast to the human auditory processing which has been shown to process local time-frequency (TF) patterns. Physiological studies conducted on mammalian auditory cortex show evidences for recognition of local TF patterns by the auditory cortical neurons during the process of recognizing sounds [2]. Another interesting aspects of human perception is a phenomenon called noise masking, as a result of which, unreliable components are either masked or discarded while recognizing the sound in the presence of noise.

These two interesting aspects of the human auditory processing (the local TF pattern processing and the noise masking) have served as motivation for the development of a recently introduced noise robust feature called spectro-temporal activity pattern (STAP) feature [3].STAP features are computed by parameterizing the local TF patterns around the spectral peaks. As the regions around spectral peaks constitute relatively high SNR part of the speech signal, STAP features show improved robustness to high noise conditions [3]. Because of the use of TF patterns around spectral peaks, the effectiveness of the STAP features depend very much upon the estimation of the spectral peak locations. In the previous work [3], a simple frequency-based dynamic programming (DP) algorithm, that utilize the spectral slope values of single time frame, has been used to perform the peak location estimation.

In this paper, we present a peak estimation algorithm that differs from the previous frequency-based DP method in two aspects. First, the method uses an alternative to the regular spectrum called phase autocorrelation (PAC) spectrum. Second, the method uses multi-layer perceptron (MLP) in a simple HMM/ANN [4] framework, for TF pattern modeling in the spectrogram. The use of PAC spectrum is expected to yield more reliable peak location information, as it has been shown to have enhanced spectral peaks and smoothed spectral valleys than the regular spectrum [5]. The use of MLP in a HMM/ANN framework makes it possible to learn distinct TF patterns in the spectrogram, hence locate the peaks by discriminating between the TF patterns.

This paper is organized as follows: Sections 2 and 3, give a short introduction to the STAP features and PAC spectrum, respectively. Section 4 gives a description of the HMM/ANN based algorithm for peak location identification. Section 5 explains the experimental set up used to evaluate the STAP feature computed using the proposed algorithm. Section 6 presents and discusses the experimental results.

2 STAP feature

Inspired by the two interesting aspects of the human auditory processing system namely, the local time-frequency processing [2] and the noise masking, the STAP approach use parameterization of local TF patterns around the spectral peaks as noise robust feature representation for the speech recognition task [3]. The effectiveness of the STAP feature depends upon two crucial factors, namely: 1) spectral peak identification, and 2) parameterization scheme used for describing the activity within local TF patterns around the spectral peaks. In the previous work, peak identification is performed using a simple frequency-based DP algorithm that utilize the single time frame spectral slope values [3]. One distinguishing aspect of this algorithm from the other spectral peak estimation algorithms, reported in the literature, is the fact that there is no constraint imposed on the number of peaks that should be estimated, which in turn avoids possible erroneous estimation of the peak location. An erroneous estimation would lead to inclusion of TF patterns from the non-peak locations in the STAP feature computation.

The parameterization scheme used to describe the activity pattern within TF patterns (i.e., the energy surface), for use in STAP feature are: 1) frequency index of the peak location, 2) energy at the peak location or average energy of the whole TF pattern, 3) delta of energy around peak location along the time axis, 4) acceleration of energy around peak location along the time axis, 5) delta of energy around peak location along the frequency axis, and 6) acceleration of energy around peak location along the frequency axis. An important note to make here is, the STAP features computed according to above explained method will yield feature vectors of varying dimensions over time. This is because the number of peaks identified can vary over time and TF patterns only around the spectral peaks are considered for computing the STAP features. However, traditional speech recognition system requires uniform dimensional feature vectors. To handle this, as explained in [3], the parameters describing activity within TF patterns around non-peak locations are first masked to be zeros and are then used in the feature vector. This results in many components with zero values in the uniform dimensional STAP features.

There is a good scope improving the frequency-based DP algorithm for peak estimation from two different directions. First, recently, an alternative to the regular spectrum called PAC spectrum has been introduced in [5]. PAC spectrum has been shown to possess enhanced spectral peaks and smoothed spectral valleys, which makes it an interesting choice for use in the peak estimation algorithm. Second, the frequency-based DP algorithm considers spectral energy values of only single time frame. This may lead to an unrealistic variation in the estimated peak locations from one frame to the other. In such case, an imposition of temporal constraint during the peak identification can be expected to provide more reliable peak location estimation.

The next section gives a brief introduction of the PAC spectrum and the following section explains a HMM/ANN based method that imposes temporal constraints during the spectral peak identification.

3 PAC spectrum

Time-domain Fourier equivalent of power spectrum is autocorrelation [6]. Traditional autocorrelation computes correlation as a dot product between the time delayed speech vectors. Recently, an alternative measure of autocorrelation called phase autocorrelation (PAC) has been introduced, where the angle between the vectors in the signal vector space is used as a measure of correlation [5]. The motivation for the use of angle is the fact that angle gets less affected in the presence of noise than the dot product [7]. If $R[k]$ represents the traditional autocorrelation coefficients, then the PAC coefficients can be computed as follows:

$$P[k] = \cos^{-1}\left(\frac{R[k]}{\|\mathbf{x}\|^2}\right) \tag{1}$$

where $\|\mathbf{x}\|^2$ represents the signal frame energy. Spectrum computed using PAC coefficients is referred to as the PAC spectrum. The computation of PAC coefficients from the autocorrelation coefficients, using (1), involves two operations namely: 1) energy normalization, and 2) inverse cosine. As explained in [5], the inverse cosine transformation has an effect of enhancing the spectral peaks and smoothing out the spectral valleys. A visual illustration of these are given in Figures 1 and 2, showing respectively the regular and the PAC spectra, for a sample speech frame corresponding to phoneme /ih/.

4 HMM/ANN based peak identification

The use of HMM/ANN for spectral peak identification is basically motivated by a previous work, where HMM employed along the frequency axis (in a general framework called HMM2 [8]) has been shown to be successful in identifying a fixed number of spectral peaks. In the current case, a simple HMM/ANN is used along frequency axis to locate the spectral peaks. HMM/ANN use multi-layer

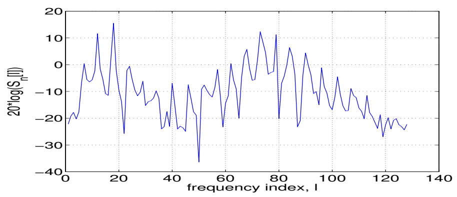

Figure 1: Energy normalized power spectrum for a sample frame of phoneme /ih/.

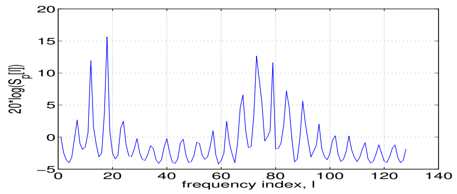

Figure 2: PAC power spectrum for a sample frame of phoneme /ih/.

perceptron (MLP) for emission modeling, as opposed to the previous HMM2 case where Gaussian Mixture Models (GMM) are used. The use of MLP provides an additional flexibility to use, more general, TF patterns in the spectrogram for the peak identification task, as the MLP has been shown to be more effective in handling the temporal contextual information [4]. The inclusion of such temporal contextual information is expected to impose temporal constraints for the peak identification. Another difference between the current and the previous HMM2 based approach is the fact that topology of the HMM/ANN used do not constraint the number of peak locations to be identified. Figure 3 shows topology of a simple HMM/ANN applied to the spectrum, for estimating the peak locations.

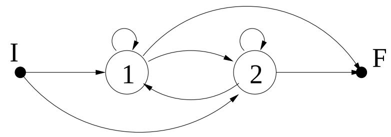

Figure 3: Topology of the HMM/ANN used to locate the spectral peaks. The states have a minimum duration of 2 (frequency bands).

For the successful use of HMM/ANN for the peak identification task, first of all, the constituent states should learn the distinct TF patterns in the spectrogram. For example, suppose along frequency axis, the TF patterns before the spectral peaks are modeled by the first state and the TF patterns after the spectral peaks are modeled by the second state. With such a learning, while Viterbi aligning the HMM/ANN along frequency axis, it is possible to locate the spectral peaks as points of transition from the first state to the second state. Now the training of HMM/ANN and its use for peak identification task raises a few issues explained as follows:

1. The presence of pitch information in the raw spectrum introduces small peaks and dips throughout the frequency range. To avoid this mel-frequency filter bank spectrum is used, where pitch information is reasonably suppressed. Additionally states of HMM/ANN, as shown in Figure 3 are imposed with minimum duration constraints to avoid spurious peaks. For the current case, mel-warped filter bank PAC spectrum of 24 dimension is used, and the minimum state duration value is fixed as 2 (frequency bands).

2. There is no transcription available during the training of the HMM/ANN for discriminating the

spectral regions into hypothesized classes, i.e., TF patterns before the spectral peaks (positive sloped TF patterns) and TF patterns after the spectral peaks (negative sloped TF patterns). In this sense, the training of the HMM/ANN need to be unsupervised. The convergence of such unsupervised training into segmentation of hypothesized regions is not always guaranteed. However, an use of slope spectrum facilitates this to a certain extent. Additionally, the topological constraints of the HMM/ANN along with minimum duration constraints is expected to further facilitate the convergence.

The experimental set up used to evaluate the performance of the STAP feature (computed using the peak locations estimated by the proposed algorithm) is explained in the next section.

5 Experimental setup

The database used for the experiments is OGI Numbers95 connected digits telephone speech database [12], having a lexicon size of 30 words, and 27 different phonemes. For additive noise experiments, factory noise from Noisex92 database has been added with Numbers95 database at various noise levels, such as 12dB, 6 dB, and 0 dB Signal-to-Noise Ratio (SNR). The speech recognition system used to compare the STAP features with the state-of-the-art features is TANDEM system [9]. TANDEM system use pre-nonlinearity outputs taken from a discriminatively trained MLP as feature input to the standard HMM-GMM system. MLP used takes 9 or 19 frames of contextual input and has 27 output units, corresponding to the number of context-independent phones. Hidden layer size is proportional to the feature dimension[1]. The HMM-GMM system consists of 80 triphones, 3 left-to-right states per triphone, and 12 mixture Gaussian Mixture Model (GMM) to estimate emission probability within each state. HMMs are trained using HTK. Mel-Frequency Cepstral Coefficient (MFCC) and CJ-RASTA-PLP [10] are the features used to evaluate the comparative performance of the STAP features. These features are of dimension 39, including 13 static coefficients, 13 delta coefficients, and 13 delta-delta coefficients. STAP features are basically extracted from spectrogram obtained with 24 dimensional mel-warped filter-bank spectrum. For the peak location estimation mel-warped filter-bank PAC spectrum is used. Including all the time-frequency pattern activity describing parameters, as listed in Section 2, STAP feature dimension is 60. However, as explained in Section 2, many of its components (typically 35-45 components) are zeros.

6 Experimental results

Assuming the HMM/ANN has been trained on PAC spectrogram of several utterances, the main factor that affects the peak location estimation performance is the size of the TF patterns, as denoted by $w_f \times w_t$, width along the frequency axis times width along the temporal axis. Figure 4 shows the peak locations identified when $w_f = 1$ and $w_t = 3$, for a sample PAC spectrum belonging to phoneme /ih/.

Figure 5 shows PAC spectrogram of a sample utterance taken from OGI Numbers95 database. Figures 6, 7, and 8 show plots of the peak locations identified by the proposed algorithm for the respective TF pattern sizes as follows:: Figure 6: $w_f = 1$ and $w_t = 1$, Figure 7: $w_f = 1$ and $w_t = 3$, and Figure 8: $w_f = 2$ and $w_t = 5$. From the figures, for the case of $w_f = 1$ and $w_t = 1$, the peak locations estimated looks to be random. The reason for this could be the fact that when single coefficient is used as TF pattern, the MLP is not able to converge to the hypothesized distinct TF patterns, as explained in Section 3. However, with better TF pattern size it is expected to behave well. For the case of $w_f = 1$ and $w_t = 3$, a better match between the peaks identified and the actual

[1]This may raise speculations about the differences in the number of parameters for different features. However, it has been verified that the performances do not change significantly with the parameter increase.

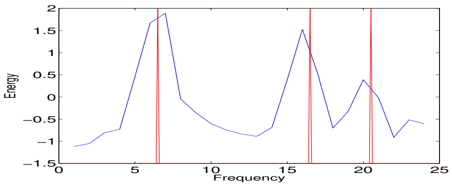

Figure 4: Spikes show the locations of peaks identified in an example filter-bank PAC spectrum corresponding to phoneme /ih/.

peak locations in the Figure 5 can be seen. For the case of $w_f = 2$ and $w_t = 5$, with the increase of TF block size the peak location estimation seems to get more constrained by the temporal context and also possibly by the width along frequency axis. STAP features used for experiments reported in the later part of this section are computed with TF pattern size of $w_f = 1$ and $w_t = 3$.

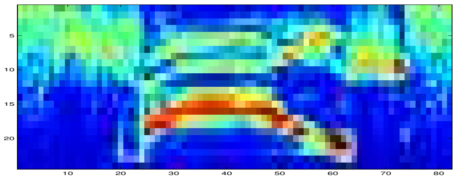

Figure 5: Mel-warped filter-bank spectrogram of a sample speech utterance taken from OGI Numbers95 database.

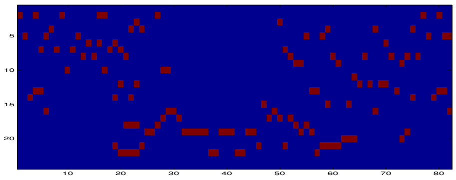

Figure 6: Peak locations identified when TF block size of $w_f = 1$ and $w_t = 1$.

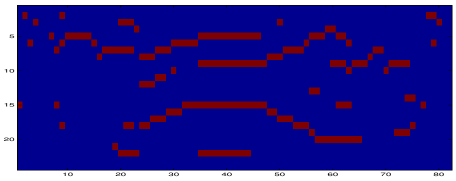

Figure 7: Peak locations identified when TF block size of $w_f = 1$ and $w_t = 3$.

The first three lines of Table 1 show results of the experiments conducted to evaluate and compare the speech recognition performance of the STAP features (computed using peak locations identified by the proposed HMM/ANN based algorithm in the PAC spectrum) with MFCC and CJ-RASTA-PLP features, for clean speech as well as various noise levels of factory noise corrupted speech. As can be seen from the table, the STAP feature is comparatively more robust to high noise conditions. Also in the high noise conditions, its performance is comparable to the CJ-RASTA-PLP. However, it is

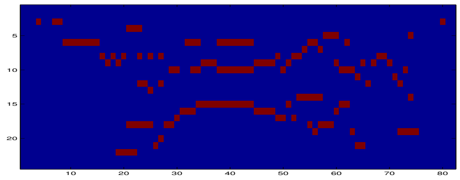

Figure 8: Peak locations identified when TF block size of $w_f = 2$ and $w_t = 5$.

inferior to the other features in clean speech condition. The reason for this (as also explained in [3]) is the masking of non-peak components which also carry significant information for the clean speech recognition. This can be compensated for by combining the STAP features with MFCC features in a TANDEM feature combination framework, as explained in [11]. Such a combination is expected to yield a representation that is reasonably robust in all the conditions. Fourth row of the table show the performances of such combination. The values show that the combination is reasonably robust in all the conditions. The fifth row give results for the case when the regular spectrum is used, instead of the PAC spectrum, in the HMM/ANN based algorithm to estimate the peak locations, keeping all other experimental settings the same. As can be seen from the results, using the PAC spectrum for spectral peak identification is better than using the regular spectrum, for all conditions.

Feature	%Word Error Rate for SNR			
	clean	12 dB	6 dB	0 dB
STAP	**10.6**	**15.3**	**22.1**	**38.3**
MFCC	4.7	12.9	25.8	52.4
CJ-RASTA-PLP	6.3	10.4	20.4	44.7
STAP + MFCC	**6.7**	**11.7**	**19.1**	**37.0**
reg-spec-STAP	14.8	19.9	27.7	45.7

Table 1: Performance comparison of STAP, MFCC, and CJ-RASTA-PLP features in TANDEM system (peak locations required for the computation of STAP feature is estimated from the PAC spectrum). Fourth row gives performance of the combination of STAP and MFCC features in a TANDEM framework. The last row gives performance comparison of the STAP features when peak locations are identified using the regular spectrum.

7 Conclusion

We have presented a HMM/ANN based algorithm for spectral peak location estimation using the PAC spectrum for further use in the computation of STAP features. This algorithm uses distinct time-frequency patterns in the PAC spectrogram for the peak identification task. Such use of time-frequency patterns imposes temporal constraints during the peak identification. Additionally, the use of PAC spectrogram facilitates a more reliable estimation of peak locations as it has enhanced spectral peaks and smoothed spectral valleys than to the regular spectrogram. Experimental results conducted to evaluate the performance of the STAP features, computed with peak information from the HMM/ANN based algorithm, shows robustness to high noise conditions. The combination of the STAP feature with MFCC feature in a TANDEM framework show a reasonable robustness in all the conditions.

References

[1] M. Kleinschmidt, "Robust Speech Recognition Based on Spectro-Temporal Processing," *Ph.D Dissertation*, University of Oldenburg, 2003.

[2] D. Depireux, J. Simon, D. Klein, and S. Shamma, "Spectro-Temporal Response Field Characterization with Dynamic Ripples in Ferret Primary Auditory Cortex." in *Journal of Neurophysiology*, 2001, 85:1220-1234.

[3] S. Ikbal, M. Magimai.-Doss, H. Misra, and H. Bourlard, "Spectro-Temporal Activity Pattern (STAP) Features for Robust ASR," to appear in *Proc. of INTERSPEECH-ICSLP-04*, Jeju Island, Korea, Oct. 2004.

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

[5] S. Ikbal, H. Hermansky, and H. Bourlard, "Nonlinear Spectral Transformations for Robust Speech Recognition," in *Proc. of IEEE ASRU 2003 Workshop*, St. Thomas, Virgin Islands, USA. Nov-Dec, 2003.

[6] A. V. Oppenheim, and R. W. Schafer, "Digital Signal Processing," *PTR Prentice-Hall, Inc.*, Englewood Cliffs, NJ, USA, 1975.

[7] S. Ikbal, H. Misra, and H. Bourlard, "Phase AutoCorrelation (PAC) derived robust speech features," in *Proc. of ICASSP-03*, Hong Kong, Apr. 2003, II-133–II-136.

[8] K. Weber, S. Ikbal, S. Bengio, and H. Bourlard, "Robust Speech Recognition and Feature Extraction Using HMM2," *Computer Speech & Language*, vol. 17, no:2-3, 2003.

[9] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *in Proc. of ICASSP-00*, Istanbul, June 2000.

[10] H. Hermansky, and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, Oct. 1994, Vol.2, No:4, pp. 578-589.

[11] S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, and H. Bourlard, "Entropy Based Combination of Tandem Representations for Noise Robust ASR," to appear in *Proc. of INTERSPEECH-ICSLP-04*, Jeju Island, Korea, Oct. 2004.

[12] R. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," *in Proceedings of European Conference on Speech Communication and Technology*, 1995, vol. 1, pp. 821–824.