INVESTIGATING DEEP NEURAL NETWORK BASED TRANSFORMS OF ROBUST AUDIO FEATURES FOR LVCSR

Enrico Bocchieri and Dimitrios Dimitriadis

AT&T Research, 180 Park Ave, Florham Park, New Jersey 07932
[enrico, ddim] at_research.att.com

ABSTRACT

Micro-modulation components such as the formant frequencies are very important characteristics of spoken speech that have allowed great performance improvements in small-vocabulary ASR tasks. Yet they have limited use in large vocabulary ASR applications. To enable the successful application, in real-life tasks, of these frequency measures, we investigate their combination with traditional features (MFCC's and PLP's) by linear (e.g. HDA), and non-linear (bottleneck MLP) feature transforms. Our experiments show that such integration, using non-linear MLP-based transforms, of micro-modulation and cepstral features greatly improves the ASR with respect to the cepstral features alone. We have applied this novel feature extraction scheme onto two very different tasks, i.e. a clean speech task (DARPA-WSJ) and a real-life, open-vocabulary, mobile search task (Speak4it^{SM}), always reporting improved performance. We report relative error rate reduction of 15% for the Speak4it^{SM} task, and similar improvements, up to 21%, for the WSJ task.

Index Terms— Neural networks, feature extraction, robustness, speech recognition.

1. INTRODUCTION

In recent years, great efforts have been focused on the task of Continuous Speech Recognition (CSR), and significant advances in the state-of-the-art have been achieved. CSR is becoming a preferred user interface for mobile applications, often in "difficult" acoustic environments. Therefore, one of the main challenges is the estimation and modeling of robust to noise speech features that can enhance the ASR performance in noisy environments.

In this context, many methods have been proposed for robust ASR feature extraction. We can distinguish these methods into two large clusters: either noise-robust feature extraction or feature post-processing to suppress some of the noise introduced. Micro-modulation features[†] capture the fine-grain formant frequency variations and are extremely robust to noise [1]. It is also quite common to post-process features by smoothing, e.g. MVA [2] or RASTA [3] filtering, and by feature transformations like HDA and/or MLLT [4, 5, 6, 7, 8]. Especially this last scheme is widely adopted in most of the state-of-the-art LV-CSR systems. However, one of the shortcomings of this method is that it is still based on the non-robust *cepstral* features, like the MFCCs [9]. Due to the fact that this transformation is linear, it fails when combining features that are very different in nature.

[†] The "micro-modulation" term is used to highlight the fine-grained time resolution of these features.

This paper presents a novel method to combine these noise-robust modulation features with cepstral features and also, filter out some of the present noise. The final features are shown to outperform any of the single-stream, cepstrum-based features already been used in real-life CSR applications. The instantaneous formant frequencies are not widely used in real-life CSR systems, although they contain significant acoustic information. This might be due to the fact that the scheme to optimally combine these formant measurements at the feature domain appears to be critical, as the experimental results of this research reveal.

As described in Section 2 we have studied two feature transformation methods for the integration of the formants measures with other features (MFCC, PLP [10]). These methods are: the linear HDA/MLLT transform [4, 6, 7, 8] and the non-linear bottle-neck neural network (hybrid MLP-HMM or tandem) [11, 12, 13, 14, 15, 16] approach. Section 3 describes the algorithm for measuring the micro-modulation formant-related audio features. The experiments in Section 4 show that the MLP-based feature transform successfully uses the combination of formant measures with PLP's to obtain 2.3 % absolute WER reduction (11% relative) over the linear HDA/MLLT transform of PLP's alone, and 3.2% absolute WER reduction (15% relative) over the linear transform of MFCC's. The adoption of the MLP-based transform is essential, because HDA/MLLT proves ineffective with the formant frequencies. These results have been obtained on a real-life, open-vocabulary, mobile search task (Speak4it^{SM}). Similar results, up to 21% relative improvements, are also reported on the standard DARPA-WSJ task. Sections 5 and 6 contain the conclusion, and the "relation to prior work", as requested by the submission guidelines.

2. FEATURE TRANSFORMATION

2.1. Linear Discriminative Front-End.

We have first adopted a discriminative feature extraction technique known as Heteroscedastic Discriminant Analysis (HDA) [4], a particular formulation of [5]. Given a number of recognition classes with arbitrary Gaussian distributions, the HDA transform provides features that maximize a ratio of between-class and within-class distortion measures. To minimize the loss of likelihood with diagonal covariance GMM states, we also apply a maximum likelihood linear transform (MLLT) [4, 6, 7, 8]. For simplicity, we will refer to the joint application of the HDA and MLLT transforms as "HDA".

Figure 1 depicts the application of the HDA transform to "super-vectors" of concatenated 11 consecutive frames (centered on the "current" frame) of MFCC (or PLP or other) "raw" acoustic features, to extract feature vectors of 60 dimensions for acoustic

HMM training and ASR decoding. CMS on the sentence level is applied to the raw acoustic features.

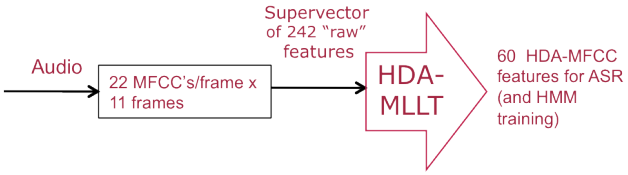

Figure 1. HDA feature extraction from MFCC's.

2.2. Non-Linear Discriminative Front-End

We have also adopted a hybrid, or TANDEM [11], speech recognition approach, based on a multi-layer perceptron (MLP), and on a conventional HMM. The MLP is configured as a non-linear feature extraction and dimensionality reduction mechanism to generate *bottleneck* [12, 13, 14, 15] features from the input raw acoustic features. Speech recognition is based on the conventional HMM, with context-dependent (triphones) Gaussian mixture model states of the bottle-neck feature vectors.

The structure of the adopted MLP bottleneck component is shown in Figure 2. Super-vectors of raw acoustic features (242 MFCC components in the Figure) are built by frame concatenation similarly to the HDA transform (Section 2.1). The global means and variances of the training data super-vector components are normalized to zero and unit values, respectively, before MLP training. The node activation functions are hyperbolic tangents, except for the softmax function at the output layer. During the MLP training the targets of the output nodes are set according to the supervised phoneme state segmentation generated by an HMM recognizer, with the cross-entropy as training criterion.

After training, the MLP outputs give an estimate of the HMM state posteriors given the input raw features. Intuitively the "bottleneck" node-layer (of dimension 60, Figure 2) provides a compact representation of the posterior probabilities: this motivates the adoption of the bottleneck node values (inputs of the node activation functions) as features for the HMM training and recognition.

We have trained the MLP with methods developed in-house, based on the BLAS library and multi-threading for enhanced computation speed. The MLP weights are estimated by iterative stochastic gradient descent with mini-batches of 300 vectors. The mini-batches are randomly created from the training corpus after each training epoch. The training is terminated after 10 epochs. We have spot-checked the ASR accuracies with MLP's trained up to 50 epochs, and we have observed only small changes with respect to the results corresponding to the 10 epochs. The learning rate is adjusted after every mini-batch weight update, according to the formula [17, 18] $\frac{\eta}{1+\theta t}$, where t is the mini-batch index and η,θ are two constants defining the learning rate for the first and subsequent mini-batches, respectively. Setting the value of θ equal to the reciprocal of the number of mini-batches in an epoch, works well for our applications.

We have also experimented with batch MLP training by the resilient back-propagation method (iRPROP) [19, 20], (details in Section 4.1).

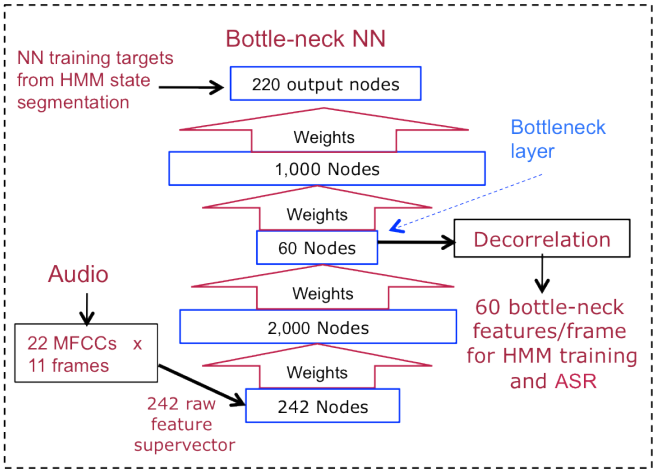

Figure 2. Structure of bottleneck MLP.

3. MICRO-MODULATION FEATURE ESTIMATION

The micro-modulation features capture the speech formant fine structure taking advantage of the excellent time resolution of the Energy Separation Algorithm (ESA). These features provide information on the instantaneous formant frequency variations and on the transient speech phenomena, and are complementary to the cepstral features. Herein, we employ the regularized GaborESA algorithm [21] for the demodulation process.

In more detail, the AM-FM speech model [22] dictates that the formant frequencies are not constant during a single pitch period, but they can vary around a center frequency. These variations are partly captured by the micro-modulation mean frequency and bandwidth coefficients F_i, B_i [21] defined for the i[th] filter as,

$$F_i = \frac{\int_0^T a_i^2(\tau) f_i \ (\tau) d\tau}{\int_0^T a_i^2(\tau) d\tau}$$

$$B_i^2 = \frac{\int_0^T \left[\dot{a}_i^2(\tau) + \left(f_i(\tau) - F_i\right)^2 a_i^2(\tau)\right] d\tau}{\int_0^T a_i^2(\tau) d\tau}$$

where $a_i(t)$ and $f_i(t)$ are the instantaneous amplitude and frequency signals, $i = 1 \ldots 6$ is the filter[‡] index, and T the time window length. The instantaneous signal $a_i^2(t)$ is used as weight for the estimation of the F_i, B_i coefficients, deemphasizing the contribution of $f_i(t)$ when the instantaneous amplitude signal takes smaller values (and thus, its estimates are not accurate enough, [22]). Finally, the coefficients are estimated over rolling *20ms* long windows with overlap of *10ms* (exactly how the cepstral features are estimated). These micro-modulation features model acoustic phenomena in a much different time-scale than the widely used Cepstral features. Consequently, a simple concatenation of these different features with the MFCC's or PLP's is far from optimal, causing some loss of acoustic information and eventually a degradation of the overall ASR performance. Herein, we suggest using the DNN architecture to combine them, taking advantage of

[‡] There is an almost one-to-one correlation between the filters and the speech resonances [22].

both the nonlinear relation between the different acoustic cues and the concatenated input vectors.

We propose using two different features based on the estimated mean/bandwidth quantities [21]. The first feature set is the "Instant. Frequency Means" (IFMean's), where the feature vector consists of the 6 F_i coefficients, i.e. one coefficient per Gabor filter (using a 6-filter Gabor filterbank), as described in detail in [25][§]. The second feature set is called "Frequency Modulation Percentages" (FMP's) and it consists of the normalized bandwidth estimates, i.e. the B_i/F_i coefficients. After some experimentation, we have found that a 12-filter Gabor filterbank performs better than the first 6-filter one for the case of the FMP's.

The first feature set provides an estimate of the mean formant frequencies in a finer time-scale. The second one provides estimates of the normalized formant frequency variances.

4. ASR EXPERIMENTS

To make valid comparisons between different frontends we apply the same processing to different audio feature types: eleven consecutive frame vectors are concatenated into super-vectors. These are transformed (by HDA or MLP) into 60 dimension feature vectors, used for HMM training and recognition. For a given ASR task, we train the triphonic HMM's for different frontends with the same number of parameters and MLE procedure. The lexicon and language model have been kept the same for all of the examined features, to investigate only the impact of the frontend scheme upon the overall system performance. The examined raw feature types, prior to concatenation into super-vectors, are the following:

- MFCC: 21 mel frequency cepstra [9], and frame energy. Super-vectors of 242 dimensions.
- PLP: perceptual linear prediction coefficients [10], and frame energy, optimized to a total of 16 coefficients per frame. Super-vectors of 176 dimensions.
- PLP+IFMean: the PLP coefficients plus formant frequencies (Section 3.1), estimated over 6 bands, or 22 coefficients in total per frame. Super-vectors of dimensions 242.
- PLP+FMP: the PLP coefficients plus the normalized formant bandwidths (Section 3.2), estimated on 12 bands, or 28 coefficients per frame. Super-vectors of 308 dimensions.

The ASR accuracy is measured on two different CSR tasks with different noise conditions and technical challenges, namely the DARPA Wall Street Journal (WSJ) and the Speak4it[SM] tasks.

4.1 DARPA-WSJ

We performed speaker-independent ASR experiments on the DARPA WSJ corpus (downsampled to 8 kHz), using the Nov93-H1 and Nov93-H2 test sets, and the *3*-gram language models built at MIT Lincoln Laboratories. The bottleneck MLP and the HMM are trained on the WSJ 284 speaker set.

For the baseline linear feature extraction, we have adopted two HDA/MLLT matrices, namely, HDA estimated on a large collection of telephone band-width data and HDA_{WSJ} estimated on the WSJ training data.

Table 1 shows the word accuracies for the two WSJ test sets, and the described frontends. The entries in the first column denote the frontend type, characterized by the transform type (i.e. MLP or HDA) and the respective, dash separated, super-vector of the input audio raw features (e.g. "MLP-PLP+IFMean" stands for MLP transform of the super-vector of the PLP and IFMean features).

The results shown in Table 1 are obtained with MLP's trained by *stochastic* gradient descent. *Batch* iRPROP training has produced lower accuracies than stochastic training, even when using a larger number (up to 1,000) of epochs (compare MLP_{iRPROP}-MFCC and MLP-MFCC in Table 1).

The bottleneck MLP feature transform, when applied to the PLP's, (see HDA_{WSJ}-PLP vs. MLP-PLP) outperforms the HDA transform, with absolute word error rate reductions of 1.1% (8.3% relative) and 0.8% (13% relative) on the Nov93-H1 and the NOV93-H2, respectively (with larger gains, up to 18% relative, for the MFCC's). The integration of the formant frequencies with the PLP coefficients (see MLP-PLP+IFMean further decreases the error rate by 1.8% and 1.3% absolute (14% and 21% relative), on the two test sets respectively.

When the HDA transforms (of the MFCC's and PLP's, respectively) are estimated on the same WSJ data (see the 1st and 3rd lines of Table 1), the accuracies of the MFCC's and PLP's are very similar. Since we believe that the PLP's offer better performance than the MFCC's in noisy conditions we have focused on improving the PLP performance by adding the micro-modulation features of Section 3. In fact, on the noisier speech of the Speak4it task (next Section), the PLP's provide much better accuracy than the MFCC's.

Table 1. Word accuracy for the WSJ task.

Frontend	Nov93-H1	Nov93-H2
HDA_{WSJ} – MFCC (MFCC baseline) :	86.6 %	94.0 %
HDA - MFCC :	87.2 %	94.3 %
HDA_{WSJ} - PLP: (PLP baseline)	86.8 %	93.8 %
MLP_{iRPROP} – MFCC:	87.0 %	94.8 %
MLP - MFCC:	88.4 %	95.1 %
MLP - PLP:	87.9 %	94.6 %
MLP - PLP+IFMean	**88.6 %**	**95.1 %**

4.2. Speak4it[SM]

We have performed more extensive tests of different feature types on the Speak4it[SM] application [23, 24], concerning real-life voice search queries using mobile devices. A noise analysis of the database reveals that the corrupting noise is low-pass (on average) with an average SNR of about 20 dB. The training and testing sets contain 337k and 6.5k sentences, respectively, with an average length of 2.5 words per sentence.

MLPs trained on different raw acoustic features have exactly the same structure (Figure 2), except than the input node layer that must accommodate different super-vector dimensions, respectively. Figure 3 shows the cross-entropy distortion (training data), versus the MLP training epoch. It is interesting to note that

[§] The filterbank configuration is motivated by the formant structure of speech signals. The Gabor filters are chosen for their optimal T-F properties.

the PLP features obtain a lower distortion than the MFCC features, even if the PLP vectors have fewer coefficients and the respective MLP has fewer free parameters. The raw features PLP+IFMean, with the added formant frequencies (and same dimensions as the MFCC's), produce even lower distortion during training. Thus MFCC's, PLP's and PLP+IFMean's show increasingly better fits of the training data, respectively. This is mirrored in the ASR accuracy, even if, admittedly, there is no theoretical relationship between cross-entropy and accuracy.

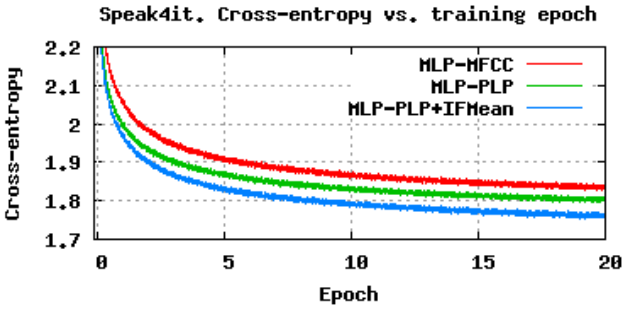

Figure 3. Cross-entropy vs epoch for different features.

After feature extraction, we trained MLE HMM's of 19k triphones, 8k GMM states and 160k Gaussians (60 dimensions), for all features sets, respectively.

Table 2 shows the word accuracy with either the MLP or HDA transforms of different audio features. In the Speak4it task the PLP's are overall more accurate than the MFCC's.

As already observed for the WSJ task, the MLP-MFCC is more accurate (by 1.0% absolute) than the HDA-MFCC, and the MLP-PLP is more accurate than HDA-PLP (by 0.8%). This confirms the improved performance of the non-linear MLP-based feature transformation scheme w.r.t. the HDA-based scheme. Additional 0.8%, and 1.0% improvements over MLP-PLP are obtained, respectively, by adding the IFMean features (MLP-PLP+IFmean) and the FMP features (MLP-PLP+FMP). The MLP transformation of the combined PLP, IFMean and FMP features could not be tested by the paper submission date.

Table 2. Word accuracy for the Speak4it^{SM} task.

Frontend	Word accuracy	Rel. WER reduction
HDA - MFCC : (MFCC baseline)	78.1 %	-
HDA – PLP: (PLP baseline)	79.0 %	4.1 %
HDA – PLP+IFMean	76.4 %	-7.8 %
MLP – MFCC:	79.1 %	4.6 %
MLP - PLP:	79.8 %	7.8 %
MLP - PLP+IFMean:	80.6 %	11.4 %
MLP – PLP+FMP:	**80.8 %**	**12.3 %**

It is noteworthy that the linear transform of the combined PLP and IFMean features (HDA-PLP+IFMean) reduces the accuracy w.r.t. HDA-PLP. Thus, the linear transform seems rather ineffective at the integration of different audio features, which is one of the motivations of this study of MLP-based transforms.

The overall error rate reduction from the baseline HDA-PLP to MLP-PLP+FMP is 1.8% absolute or 8.5% relative. 1.0% absolute (5% relative) is attributable to the use of the FMP features (compare MLP-PLP+FMP and MLP-PLP) as input to the MLP transform.

4.3. Improved MLP structure, Speak4it.

The experiments in Sections 4.1 and 4.2 (Tables 1 and 2) are in part designed to compare the accuracy of the HDA/MLLT and MLP-based feature transforms. Therefore we created input raw feature super-vectors by concatenating the same number, i.e. 11, of consecutive frames, and we adopted the same dimensionality, i.e. 60, for both the HDA output features and the MLP bottleneck features. However, it seems that this parameterization had been optimized for the HDA transform, and it is far from optimal for the MLP system.

We have started experimenting with better MLP structures. For example increasing the input super-vector of MLP-PLP+IFMean to 17 consecutive frames (instead of 11) has improved the accuracy to 81.3% (instead of 80.6% in Table 2).

Thus the best error rate reduction (Speak4it task) with respect to the HDA-PLP baseline is 2.3% absolute (11% relative), and with respect to HDA-MFCC is 3.2% absolute (15% relative).

5. CONCLUSION

The proposed ASR frontend reduces the absolute word error of the Speak4it^{SM} large vocabulary voice-search task by 2.3% absolute (11% relative) w.r.t. to the HDA transform of PLP coefficients. The improvement with respect to the HDA transform of MFCC's is larger (3.2% absolute, or 15% relative). Large improvements, up to 21% relative, are also reported for the standard WSJ task. The proposed frontend is based on:

- micro-modulation formant-related features, and
- non-linear MLP-based feature transform for the integration of micro-modulation and cepstral features (while linear transforms proved ineffective).

We hope that these results may inspire the study of micro-modulation features in conjunction to non-linear feature transforms, for the robustness of ASR applications.

6. RELATION TO PRIOR WORK

This research relates to linear [4, 5, 6, 7] and non-linear MLP-based [11, 12, 13, 14,15] feature transformation methods for ASR. These prior works have not focused on the potential advantages of combining different types of audio features as input to the feature transformation process. On the contrary, this paper extends the prior research on non-linear MLP feature transformations with the successful integration of spectral and formant measurements in the large vocabulary voice-search Speak4it^{SM} task. Herein, it is proved that the MLP architecture can be successfully used to combine multiples of different features, keeping the non-trivial acoustic information. Previous work on the use of formant measures [1, 21] concerns small vocabulary tasks and uses separate information streams by combining their respective likelihoods. Instead our method combines the different features into one stream, by bottleneck MLP transforms. This study also extends our previously published work on the Speak4it task [23,24].

REFERENCES

[1] K. K. Paliwal, "Spectral subband centroid features for speech recognition", in *Proc. ICASSP*, Seattle, WA, May 1998, pp. 617–620.

[2] Chia-Ping Chen and J. A. Bilmes, "MVA Processing of Speech Features" *IEEE Trans. On Audio, Speech and Lang. Proc*. Vol. 15, No. 1, Jan. 2007.

[3] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. On Speech and Audio Processing*, Vol. 2, No. 4, Oct. 1994.

[4] G. Saon, M. Padmanabhan, R. Gopinath, and S.Chen, "Maximum likelihood discriminant feature spaces," in *Proc. ICASSP-2000*, pp. 1129–1132.

[5] N. Kumar and G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[6] R.A. Gopinath, "Maximum likelihood modeling with [Gaussian distributions for classifications," in *Proc. ICASSP*, 1998, pp. 661–664.

[7] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Tech. Rep. CUED/FINFENG/TR291*, Cambridge Univ., 1997.

[8] M. Saraclar, M. Riley, E. Bocchieri, and V. Goffin, Towards automatic closed captioning: low latency real time broadcast news transcriptions", *in Proc. International Conference on Spoken Language Processing (ICSLP*), Sep. 2002, pp. 741-1744.

[9] S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. on Acoustic, Speech and Signal Processing*, 28(4):357–366, 1980.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. of Acoust. Soc. of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[11] H.Hermansky, D. Ellis and S.Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems.", *in Proc. of ICASSP-2000*, Istanbul, June 2000.

[12] F. Gre´zl, M. Karafia´t, S. Konta´r and J. Cˇernocky, "Probabilistic and bottle-neck features for ASR of meetings", in *Proc. of ICASSP-2007*, pp 757-760.

[13] P.Fousek, L.Lamel and J.Gauvain, "Transcribing broadcast data using MLP features", in Proc. Interspeech 2008, pp 1433-1436.

[14] D.Yu and M.L.Seltzer,"Improved Bottleneck Features Using Pretrained Deep Neural Networks", in Proc. Interspeech 2011, pp 237-240.

[15] T.N.Sainath, B.Kingsbury and B.Ramabhadran,"Auto-encoder bottleneck features using deep belief networks", *in Proc. of ICASSP-2012*.

[16] Z.Tüske, R.Schlüter, H. Ney, M.Sundermeyer, "Context-Dependent MLPs for LVCSR: TANDEM, Hybrid or Both?," *Proc. INTERSPEECH 2012*, Portland, Oregon.

[17] W. XU, "Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent", last revised 22Dec 2011, arxiv.org/abs/1107.2490.

[18] L.Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent", *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, 177–187, Edited by Yves Lechevallier and Gilbert Saporta, Paris, France, August 2010, Springer.

[19] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm", in *Proc. IEEE International Conference on Neural Network*s, 1993, pp 586-591.

[20] C. Igel and M. Hüsken, "Improving the Rprop Learning Algorithm", in *Proc. Of The Second International Symposium On Neural Computation* (NC2000).

[21] D. Dimitriadis, P. Maragos, and A. Potamianos "Robust AM-FM Features for Speech Recognition", *IEEE Signal Processing Letters,* Vol. 12, No. 9, Sept. 2005.

[22] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc.Amer.*, vol. 99, no. 6, pp. 3795–3806, Jun. 1996.

[23] E.Bocchieri, D.Caseiro and D.Dimitriadis, "Speech recognition modeling advances for mobile voice search,",in *Proc. ICASSP*, 2011, pp 4888-4891.

[24] D.Dimitriadis, E.Bocchieri and D.Caseiro, "An alternative front-end for the AT&T WATSON LV-CSR system", in *Proc. ICASSP*, 2011.

[25] D. Dimitriadis and P. Maragos, "Continuous Energy Demodulation Methods and Application to Speech Analysis", Speech Communication, vol.48, no.7, pp.819-837, July 2006.