IDIAP RESEARCH REPORT

Hierarchical Neural Networks Feature Extraction for LVCSR system

Fabio Valente [a], Jithendra Vepa [a], Christian Plahl [b]
Christian Gollan [b], Hynek Hermansky [a], Ralf Schlüter [b]

IDIAP–RR 07-08

March 2007

published in
Interspeech 2007

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Lehrstuhl für Informatik 6, Computer Science Department RWTH Aachen University, Aachen, Germany

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-08

Hierarchical Neural Networks Feature Extraction for LVCSR system

Fabio Valente, Jithendra Vepa, Christian Plahl
Christian Gollan, Hynek Hermansky,Ralf Schlüter

March 2007

published in
Interspeech 2007

Abstract. This paper investigates the use of a hierarchy of Neural Networks for performing data driven feature extraction. Two different hierarchical structures based on long and short temporal context are considered. Features are tested on two different LVCSR systems for Meetings data (RT05 evaluation data) and for Arabic Broadcast News (BNAT05 evaluation data). The hierarchical NNs outperforms the single NN features consistently on different type of data and tasks and provides significant improvements w.r.t. respective baselines systems. Best result is obtained when different time resolutions are used at different level of the hierarchy.

1 Introduction

Data driven feature extraction aims at producing features for ASR using a statistical front-end. An efficient method of generating discriminative features is based on the use of Neural Networks (NN) trained to classify phonetic targets [1]. Input to the NN is generally a section of the time-frequency plane. Different kinds of input have been investigated in the past: short temporal context input (9 frames PLP, see [1]), long temporal context input (HATS see [2]) and multiple time resolution input (MRASTA see [3]).

Data driven features provide complementary information to classical spectral features (PLP, MFCC) obtained from a temporal window of 30ms and considerable improvements in LVCSR tasks ([4]). Neural Network structure has been an active research topic as well: three-layers ([1]) and four-layers NN ([2]) has been studied and several hierarchical structures have been considered [5], [6].

In this paper we study two kinds of hierarchical neural network structures: the first structure based only on short time-frequency input and the second structure that combines long and short time-frequency input. We show that in the hierarchy, the second NN performs error correction on the most probable errors of the first one. Features are tested in two LVCSR systems trained on Meetings data and Arabic Broadcast News transcription, showing consistent improvement over single NN features and classical spectral features (PLP, MFCC).

The paper is organized as follows: section 2 describes the general idea of NN based feature extraction, section 3 reports some preliminary experiments with a single NN, sections 4 and 5 describe experiments with hierarchal neural networks with short and long temporal context, sections 6.1 and 6.2 report results of hierarchical NNs feature extraction in two different LVCSR tasks: transcription of meetings data (RT05 evaluation data) and transcription of Arabic Broadcast News (BNAT05 evaluation data).

2 NN based feature extraction

In a multi-class problem, NN can be trained so that the output approximates class posterior probabilities (see [8]). Generally, a three-layer NN structure is used but other topologies have been investigated as well. Output from third layer are then normalized using a softmax function so that the sum of outputs is one.

In speech recognition, targets are represented by phonetic units, thus NN estimates posterior distribution of a given phoneme. A speech segment can be turned into a *posteriogram* i.e. a representation of the posterior probability of phonemes for each time frame (e.g. figure 4). Ideally a well trained NN will activate an output unit when a given spectro-temporal pattern is presented as input. For instance, in [13] input consists of 9 consecutive PLP frames while in [3] it consists of one second long segment obtained from the time-frequency plane. In order to use NN features in the classical HMM system, they are first gaussianized using log transform and then transformed using a KLT transform: this technique is referred as TANDEM [1]. In LVCSR systems they are typically appended to spectral features [4].

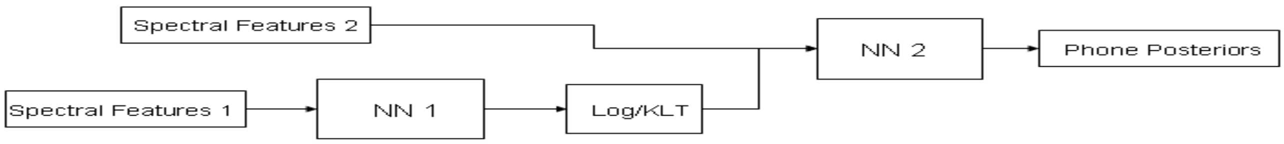

Figure 1: Hierarchical Neural Net processing

3 Experiments with single net

In this section we briefly describe experiments with a single NN on meetings data. Database consists of about 100 hours of meetings recorded at different sites. The independent headset microphone channel is used. Phoneme set consists of 46 targets including silence. Training data are phonetically labeled using forced alignment through the AMI LVCSR system (see section 6.1 and [10]).

We trained a Neural Net on those data with a temporal context of 9 frames. Almost 25% of total frames are labeled as silence thus it is interesting to report frame error rate without silence. In Table 1, Frame Error Rate (FER) with silence and without silence are reported together with FER obtained considering the two and three-best output list.

	w silence	w/o sil	2 best w/o sil	3 best w/o sil
FER	34.6 %	43.0%	32.3 %	26.3%

Table 1: FER computed with silence, without silence and using the two highest output and the three highest output.

We notice that around 40% of the errors are generated by confusion in between two or three phonemes. For instance figure 2 plots confusion patterns for phonemes */g/*. Several instances of phoneme */g/* are classified as phoneme */k/*. Very similar error patterns are seen for all other phonemes in which the largest part of confusion is found in between two or three best competitors.

In literature, different approaches have been proposed for correcting confusion pattern in NN based phoneme classifiers. For instance, in [7] the use of error correction code is investigated. We show here that the use of a hierarchy of NNs provides an effective tool for correcting those kinds of errors.

4 Hierarchical Neural Network

We consider a hierarchical processing based on a cascade of NNs in which the second net has as input, posterior features (i.e. posteriors after Log/KLT transform) from the first net together with spectral features. Intuitively, the first net will activate a given output when a certain spectro-temporal pattern is seen as input. As described before, this will lead to a certain amount of errors, most of them in between two of three competing phonemes. The second net ideally will correct those kind of errors, using the spectro-temporal patterns that disambiguates in between different phonemes.

Figure 1 plots the general scheme for hierarchical NNs used for experiments. An initial set of spectral features is used for training a first net. Output is processed using a Log/KLT transform. New features are used for training a second neural net together with a second set of spectral features.

We investigate two different schemes; in the first one, already investigated in [6], we build the hierarchy using only short-context time-frequency input i.e. a block of 9 frames PLP features augmented with dynamic features. We refer to this as Hierarchical TANDEM. In the second scheme, we consider both long temporal context features (one second) and short-context features. We refer to this as Hierarchical MRASTA.

	w silence	w/o sil	2 best w/o sil	3 best w/o sil
NN	34.6 %	43.0%	32.3 %	26.3%
HNN 1	29.2 %	35.9 %	27.5 %	22.7 %
HNN 2	27.0 %	33.0 %	25.5 %	21.1 %

Table 2: FER computed with silence, without silence and using the two highest output and the three best outputs for single NN, and hierarchy of two and three NN with 9-frame PLP input.

In Table 2, we reports FER for the single NN, and for hierarchical TANDEM with two and three NNs. Input is 9 frames PLP features. 5.5% absolute FER reduction is obtained by the second NN and further 2% is obtained by the third. It is also interesting to notice that the difference between the FER of the best output and of three highest outputs is progressively reduced. Figure 3 plots confusion patterns for phonemes /g/ across different level of the hierarchy: the largest gain in performance is obtained against most confusing phonemes from the previous NN. This behavior is observed for *all* the phonemes i.e. there is a reduction in frame error rate for every phoneme in the set.

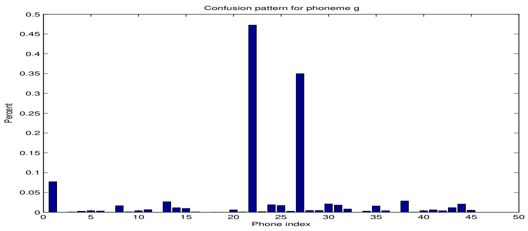

Figure 2: Confusion pattern for phoneme /g/; several instances of /g/ are classified as /k/

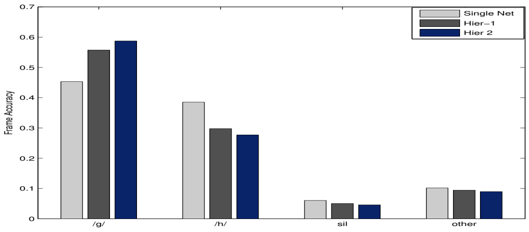

Figure 3: Confusion pattern for phonemes /g/ for the three different level of hierarchy.

An interesting effect of such a hierarchy (already described in [6]) is that at each layer the acoustic context is progressively increased: if the first NN has a temporal context of 9 frames, the second NN will use a temporal context of 9+8 frames. In next section, we investigate the use of a hierarchy of NN using directly different temporal context inputs at different layers.

5 Hierarchical NN with different temporal context input

MRASTA features are obtained giving as input to a NN a spectro-temporal cut of one second processed according to a zero mean multi-resolution filter ([3]). Thus MRASTA features are very robust to channel distortions.

Hierarchy of Neural Network can be used to incorporate different time resolution input at different level. In this section we investigate this framework. Considering schema of figure 1 the first set of spectral features has a temporal context of one second while the second set of spectral features uses only 9-frames context. We refer to this method as Hierarchical MRASTA. Table 3 shows FER for single MRASTA and Hierarchical MRASTA.

	w silence	w/o sil	2 best w/o sil	3 best w/o sil
NN	36.3 %	45.9%	34.9 %	28.4 %
HNN 1	28.9 %	35.6 %	27.7 %	23 %

Table 3: FER computed with silence, without silence and using the two highest output and the three best outputs for MRASTA and hierarchical MRASTA.

Results in terms of FER are close to those obtained in Table 2. Hierarchy with three levels doesn't further improve FER. The same behavior previously described can be noticed i.e. confusion between

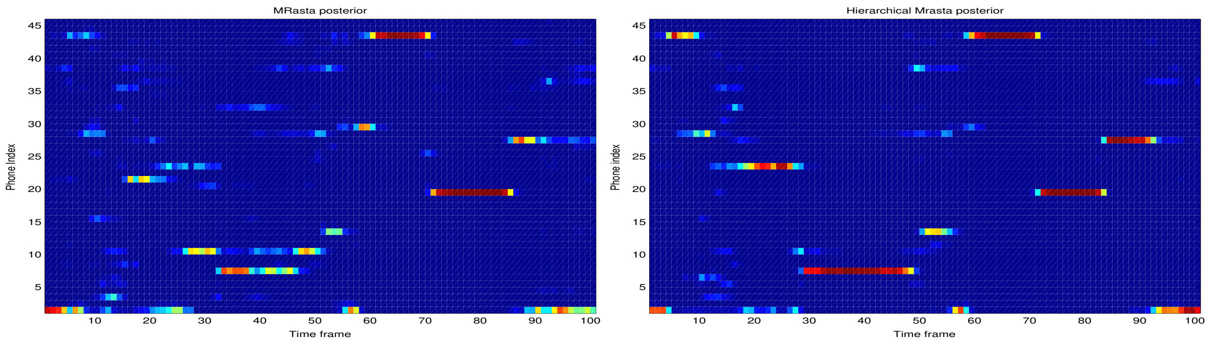

Figure 4: Posteriograms for MRASTA (left) and Hierarchical MRASTA (right)

phonemes is drastically reduced. It is possible to visualize the difference between the two methods using a *posteriogram*. Figure 4 plot posteriograms for MRASTA (left) and hierarchical MRASTA (right). Posteriogram on the right is much smoother than the one on the left, achieving at the same time a lower FER.

6 LVCSR experiments

6.1 Meeting system

For investigation purposes, we run experiments on RT05 [9] data without concatenation with other features and compare with PLP features. The training data for this system comprises of individual headset microphone (IHM) data of four meeting corpora; the NIST (13 hours), ISL (10 hours), ICSI (73 hours) and a preliminary part of the AMI corpus (16 hours). Acoustic models are phonetically state tied triphone models trained using standard HTK maximum likelihood training procedures. The recognition experiments are conducted on the NIST RT05s [9] evaluation data. We use the reference speech segments provided by NIST for decoding. The pronunciation dictionary is same as the one used in AMI NIST RT05s system [10]. Juicer large vocabulary decoder [11] is used for recognition with a pruned trigram language model.

Table 4 reports results obtained using PLP (baseline system), TANDEM, MRASTA, hierarchical TANDEM and hierarchical MRASTA.

Out of the proposed feature set, hierarchical MRASTA is providing the best performance and reduces WER of 3% absolute compared to PLP features. As general remark the use of the hierarchy always improves w.r.t. the single net. Let us consider results in more details.

TANDEM features slightly outperforms PLP features on AMI and ICSI data. On the other hand, there is a consistent drop in performance in VT data which are particularly noisy . Let us consider now the effect of the hierarchy. When a second NN is used an average improvement of 2.2% absolute is obtained; this improvement is verified on all type of data. Thus second NN is improving recognition at both phoneme and word level when used for generating features. On the other hand, when a third NN is added, overall performance deteriorates of 0.6% probably because of over fitting to the data.

MRASTA features are designed to remove mean value in the modulation spectrum trough the use of a multi-resolution filter thus more robust to noise and distortions. Furthermore they use an acoustic context of one second. Overall performance of MRASTA is better than TANDEM features. Contrarily to TANDEM, on VT data, they hold performance comparable to PLP.

Most interesting results we obtained is based on using hierarchical MRASTA with different temporal context as described in section 5: an average improvement of 6% absolute is verified w.r.t single net MRasta. Improvements are seen on all data sets in the RT05; 2% improvement is also observed on VT data where results for TANDEM are very poor. Furthermore this set of features outperforms by 3% classical PLP front-end.

Features	TOT	AMI	CMU	ICSI	NIST	VT
PLP+D+A	42.4	42.8	40.5	31.9	51.1	46.8
TANDEM	46.6	41.4	43.7	31.3	54.5	64.9
Hier TANDEM 1	44.4	39.6	42.3	28.9	51.5	62.0
Hier TANDEM 2	45.0	40.5	44.4	29.4	51.1	61.9
MRASTA	45.9	48.0	41.9	37.1	54.4	48.8
Hier MRASTA 1	39.4	38.1	36.9	28.2	48.0	46.9

Table 4: WER for Meeting data.

6.2 Arabic BN system

In this section we investigate the use of hierarchical features in concatenation with spectral features in a LVCSR system for transcription of Arabic Broadcast News. As described in [14], the acoustic front end uses MFCC features with cepstral mean normalization. The MFCC features are augmented with a *voicedness feature* [12] and includes Vocal Tract Length Normalisation (VTLN). The MFCCs and voicedness features from nine consecutive frames are concatenated and a linear discriminative analysis (LDA) is used to reduce the feature dimensions. The Neural Network features were concatenated with the LDA-transformed MFCC baseline system. Acoustic models were triphone based Viterbi trained Gaussian mixture models (GMMs) with a global pooled covariance matrix. The triphones are top down clustered using CART, rendering 4501 generalized triphone states with cross-word context.

The training corpus consists 120 hours of speech, derived from the FBIS (30h) and the Arabic TDT4 (60h) corpus. 30h of additional data is taken from the first two quarter releases of the first year (Y1Q1, Y1Q2) of the GALE project. All data consists of Arabic Broadcast News. Most available training material for the Arabic speech recognition don't inlcude diacritics. Ignoring theses diacritics increase the error rate ([16]). For this purpose the Buckwalter Arabic Morphological Analyser [1] is used to vowelise the transcriped data. Because not all words are mapped to a diacritic form we used a data-driven approach known as Grapheme-to-Phoneme conversion [15]. In the training process for the Grapheme-to-Phoneme conversion a mapping beetween the orthographic form of a word an its phonetic transcription is build. These models are used to create the transcription of unknown words.

The language model used for the experiments is derived from the Gigaword Arabic, the Arabic TDT4 and from the FBIS corpus. Additional data is taken from the Y1Q1 and Y1Q2 corpora of the GALE project. The language model is a bigram with a vocabulary of 256k words.

The BNAT05 evaluation corpus has been used for evaluation purpose Table 5 summarizes the improvement produced by using the concatenated Neural Network features. As shown in Table 5, the single net features improves the recognition rate by 0.3% absolute for TANDEM and by 0.6% absolute for MRASTA. Using the hierarchical features there is an additional improvement by 0.6% absolute towards the single MRasta and TANDEM approach. Overall, the best results are archieved using the hierarchical MRasta feature approach and results are consistent with what is observed on meeting data.

7 Conclusion

In this paper we investigate the use of a hierarchy of Neural Networks for performing data driven feature extraction. Two different framework are proposed: one with the same temporal context at both level of the hierarchy (Hierarchical TANDEM) and one with changing context (Hierarchical MRASTA). Consistent reduction in Frame Error Rate are observed for both framework. The second

[1] http://www.qamus.org/morphology.htm

Features	WER
MFCC	23.6
MFCC + TANDEM	23.3
MFCC + MRASTA	23.0
MFCC + Hier TANDEM	22.7
MFCC + Hier MRASTA	22.4

Table 5: WER for Arabic BN task.

net has the property of correcting the most confusable patterns of the first one. NN based features are investigated in two LVCSR systems for transcription of meetings and Arabic BN. Hierarchical MRASTA method is showing the best performance in both systems providing consistent improvements w.r.t. respective baseline systems. Changing time resolution across different level of the hierarchy seems to be very effective and must be further addressed in future works.

8 Acknowledgments

This material is based upon work supported by the EU under the grant DIRAC IST 027787 and by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 . Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). We thank SRI International for the help to build up the Arabic recognition system, especially Dimitra Vergyri and Andreas Stolke for the closed collaboration. We thank Thomas Hain and the AMI ASR team for their help with the meeting system.

References

[1] Hermansky H., Ellis D.P.W. and Sharma S. , “Tandem connectionist feature extraction for conventional HMM systems”, Proceedings of ICASSP 2000.

[2] Chen B., Zhu Q., and Morgan N., “ Learning Long-Term Temporal Features in LVCSR Using Neural Networks”, Proceedings of ICSLP 2004.

[3] Hermansky H. and Fousek P., ”Multi-resolution RASTA filtering for TANDEM-based ASR.”, Proceedings of Interspeech 2005.

[4] Zhu Q., Chen B., Morgan N., and Stolcke A., “On using MLP features in LVCSR”, Proceedings of ICSLP 2004.

[5] Sivadas S. and Hermansky H., ”Hierarchical Tandem Feature Extraction”, Proceedings of ICASSP-2002.

[6] Schwarz P., Matejka P., Cernock J.,”Hierarchical structures of neural networks for phoneme recognition”, Proceedings of ICASSP 2006.

[7] Hagen H. and Bourlard H., “Error Correcting Posterior Combination for Robust Multi-Band Speech Recognition”, in Proceedings of EUROSPEECH, 2001

[8] Bourlard, H. and Wellekens, C.J. (1989), “Speech Pattern Discrimination and Multilayer Perceptrons” Computer, Speech and Language (Academic Press), vol. 3, pp. 1-19.

[9] http://www.nist.gov/speech/tests/rt/rt2005/spring/

[10] Hain, T. et al, “The 2005 AMI System for the Transcription of Speech in Meetings” NIST RT05 Workshop, 2005, Edinburgh, UK.

[11] Moore, D et al. “Juicer: A weighted finite state transducer speech coder” Proc. MLMI 2006 Washington DC.

[12] Zolnay A., Schlter R., Ney H. “Robust Speech Recognition using a Voiced-Unvoiced Feature”, Proc. of ICSLP, Denver, CO, Vol. 2, pp. 1065–1068, Sept. 2002.

[13] Bourlard H. and Morgan N.,”Connectionist Speech Recognition - A Hybrid Approach”,Kluwer Academic Publishers, 1994.

[14] Lööf J. et al. “The 2006 RWTH Parliamentary Speeches Transcription System”, In Proceedings of ICSLP 2006.

[15] Bisani M., Ney H. “Multigram-based grapheme-to-phoneme conversion for LVCSR”, Proc. Eurospeech, 2003.

[16] Vergyri D., Kirchhoff K. “Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition”, COLING Workshop on Arabic-script Based Languages , 2004.