Application of EαNets to Feature Recognition of Articulation Manner in Knowledge-Based Automatic Speech Recognition

Sabato M. Siniscalchi[1,3], Jinyu Li[1], Giovanni Pilato[2], Giorgio Vassallo[3], Mark A. Clements[1], Antonio Gentile[3], and Filippo Sorbello[3]

[1] Center for Signal and Image Processing,
School of Electrical and Computer Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332, United States of America
{jinyuli, clements}@ece.gatech.edu
[2] Istituto di CAlcolo e Reti ad alte prestazioni,
Italian National Research Council,
Viale delle Scienze (Edif. 11), 90128 Palermo, Italy
g.pilato@icar.cnr.it
[3] Dipartimento di Ingegneria Informatica,
Universita' degli studi di Palermo,
V.le delle Scienze (Edif. 6), 90128 Palermo, Italy
siniscalchi@csai.unipa.it,
{gvassallo, gentile, sorbello}@unipa.it

Abstract. Speech recognition has become common in many application domains. Incorporating acoustic-phonetic knowledge into Automatic Speech Recognition (ASR) systems design has been proven a viable approach to rise ASR accuracy. Manner of articulation attributes such as vowel, stop, fricative, approximant, nasal, and silence are examples of such knowledge. Neural networks have already been used successfully as detectors for manner of articulation attributes starting from representations of speech signal frames. In this paper, a set of six detectors for the above mentioned attributes is designed based on the E-αNet model of neural networks. This model was chosen for its capability to learn hidden activation functions that results in better generalization properties. Experimental set-up and results are presented that show an average 3.5% improvement over a baseline neural network implementation.

1 Introduction

State-of-the-art speech recognition technology utilizes frame-based feature vectors, corresponding to about 10-20 milliseconds (ms) of speech (frame length). Within this framework, Mel-Frequency Cepstrum Coefficients (MFCCs)[4] are the most commonly employed features because of their properties to capture the main characteristics of the vocal cords and tract. Moreover, these features are usually computed by means of short-term spectral techniques, such as linear prediction (LP) analysis, or band-pass filter benches (BPFs). In addition,

B. Apolloni et al. (Eds.): WIRN/NAIS 2005, LNCS 3931, pp. 140–146, 2006.

as the Continuous Density Hidden Markov Model (CDHMM)[1] models the sound classes, data-driven machine learning techniques allow the training of the CDHMM parameters directly from the speech data by way of dynamic programming algorithms, e.g. Baum-Welch procedure [1]. Nonetheless, even if speech researchers have learned a lot on how to build speech recognition systems, the performance of Automatic Speech Recognition (ASR) systems are comparable to Human Speech Recognition (HSR) only when working conditions match training conditions [7]. In this context, it is interesting to note that human beings integrate multiple knowledge sources in bottom-up fashion. The HSR system gathers acoustic and auditory information from the speech signal, combines them into cognitive hypotheses, and then recursively validates these hypotheses until a final decision is reached. Conversely, data-driven automatic systems, such as the Hidden Markov Model (HMM) [2] or Artificial Neural Networks (ANN) [1], address the speech recognition problem as a top-down paradigm, directly trying to convert the speech signal into words, and thus neglecting all the rich set of information that a speech signal conveys, such as gender, accent, speaking style, etc.

To overcome these limits one could attempt to incorporate the above information by collecting more data for the training data set. Nevertheless, C.-H. Lee has recently pointed out that the performance of these knowledge-ignorant modelling approaches can be improved integrating the knowledge sources available in the large body of speech science literature [11]. In the same work, he proposed a detection-based automatic speech recognition (ASR) paradigm through automatic speech attribute transcription (ASAT). Furthermore, in [6] it is showed that the idea of a direct incorporation of acoustic-phonetic knowledge, as knowledge-based features (also referred to as speech attributes in the same work) into ASR design rises the accuracy. This goal was achieved augmenting the front-end module of a conventional ASR system by means of a set of feature detectors able to capture the above-mentioned speech attributes.

The problem addressed in this paper is to build a set of detectors to recognize six attributes, namely *vowel, stop, fricative, approximant, nasal, and silence.* These attributes represent the manner of articulation, and in this paper are referred to as *manner of articulation* attributes. The choice of these attributes was dictated by not only their strong relation to human speech production [3], but also by their robustness to speech variations [6]. These six manner events are extracted directly by short time MFCCs, and represent the direct input to the six detectors.

It is well known that neural networks are widely used since they can learn a mapping from an input space to an output space realizing a compromise between recognition speed, recognition rate and hardware resources[8]. The generalization capability of neural networks is acquired during the training phase and the generalization degree achieved is strictly related to the training set characteristics. In [6], feed-forward neural networks are used to implement the six speech attribute detectors, the output of which may be interpreted as a posteriori probabilities of an attribute given the speech signal [3]. This set of detectors is used as baseline for comparisons against the model herein proposed.

Recently, a feed-forward neural architecture (EαNet) capable of learning its hidden neuron activation function has been introduced [8], and proven to perform better than traditional feed-forward neural architectures [9][10]. In this architecture, the activation functions of the hidden units are not chosen a priori, but rather they are approximated with a regression formula based on orthonormal Hermite polynomial functions. Each activation function belonging to the hidden layer, along with the neuron connection weights is then learnt during the training phase with the use of the Conjugate Gradient Descent technique with the Powells restart conditions[8].

In this paper, a set of six attribute detectors is designed based on the EαNet neural architecture. Each one of the EαNet detector classifies input speech frames into a single attribute category. The performance is evaluated on continuous phone recognition using the TIMIT database [12]. Experimental results demonstrate the effectiveness of this design for speech attribute classification, with an average 3.5% improvement with respect to the traditional ANN (maximum 8.5% improvement for plosives). The rest of the paper is organized as follows. Section 2 describes the architecture of the EαNet neural network. Section 3 describes the general framework of the *knowledge extraction* module. Section 4 presents the experimental set-up and results, with comparison to the baseline architecture. Some concluding remarks close the paper.

2 E-αNet Architecture

EαNet [8][9][10] is a feed-forward neural architecture, or multi-layer perceptron, that is able to learn the activation function of its hidden units during the training phase. Compared to a traditional feed-forward network that uses sigmoidal or sinusoidal activation functions for its hidden unit, this model is characterized by lower values of the gradient of the network output function in the surroundings of the training points. Furthermore, to avoid introduction of additional information beyond what already available in the training set (see [9][10]) in EαNet architectures the activation function is modelled through a Hermite regression formula and the optimization algorithm is based on the conjugate gradient descent with Powell restart conditions [8]. The choice of the Hermite regression algorithm is motivated by i) the smoothness of the resulting interpolation, and ii) its easy to compute first derivative [8][9].

3 Knowledge Extraction Module

The Knowledge Extraction (KE) module uses a frame-based approach to provide K manner of articulation attributes ($A_i, i \in [1, \ldots, K]$) from an input speech signal $s(t)$. In this paper the manner classes were chosen as in [6], and are listed in Table 1.

The KE module, depicted in Figure 1, is composed of two fundamentals blocks: the feature extraction module (FE), and the attribute scoring module (SC).

Table 1. Manner articulation attribute

Articulation manner	Class Elements
VOWEL	IY, IH, EH, EY, AE, AA, AW, AY, AH, AO, OY, OW, UH, UW, ER, AX, IX
FRICATIVE	JH, CH, S, SH, Z, ZH, F, TH, V, DH
STOP	B, D, G, P, T, K, DX
NASAL	M, N, NG, EN
APPROXIMANT	L, R, W, Y, HH, EL
SILENCE	SIL

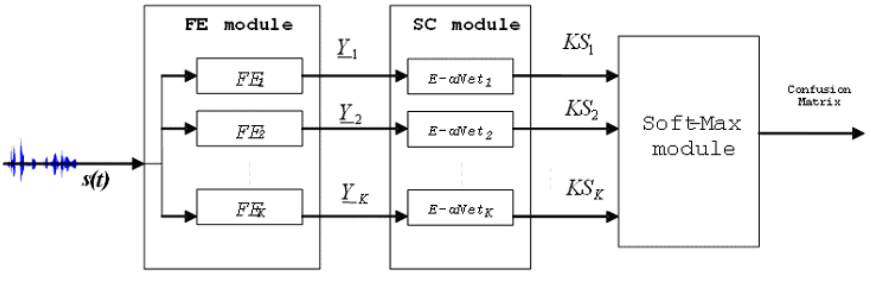

Fig. 1. Procedure of encoding the k-th word of the i-th lexical set

The FE module consists of a bank of K feature extraction blocks FE_i, where $i \in [i \dots K]$ and it maps a speech waveform into a sequence of speech parameter vectors $Y_i, i \in [i \dots K]$. Actually, each of the FE_i is fed the same speech waveform $s(t)$ and for each 10 ms-frame it computes a thirteen-MFCC feature vector X_i (12MFCCs + Energy). The frame length of 30 msec, overlapped by 20 msec.

Finally, FE_i produces, as output, a 117-feature vector Y_i combining the actual frame with the eight surrounding frames, 4 frames before and after, so that each speech parameter vector represents nine frames.

The SC module is composed of six E-αNets feed-forward neural networks, and its goal is to attach a score, referred to as *knowledge score* (KS_i), to each vector Y_i. The input of each network is a 9 frames of 12MFCCs + energy, so that the input layer is of 117 nodes. The output layer has two nodes, one for the desired class, and one for the anti-class (which are the the elements belonging to the other classes). Actually, the value obtained for the desired class for case i is defined to be the knowledge score (KS_i).

4 Experiments and Results

The evaluation of the proposed Manner of Articulation Manner Extraction module was performed on the TIMIT Acoustic-Phonetic Continuous Speech Corpus database [12], which is a well-known speech corpus in the speech recognition field.

This database is composed of a total of 6300 sentences; it has a one-channel, 16-bit linear sampling format, and it was sampled at 16000 samples/sec. The EαNet detectors were trained on 3504 randomly selected utterances, and to be consistent with [6] and [5] the four phones "cl", "vcl", "epi", and "sil" were treated as a single class, thus reducing the TIMIT phone set to a set of 45 context-independent (CI) phones.

Each of the six EαNet detectors is a three-layer network the input of which is a window of nine frames, that is, 117 parameters. The nodes of hidden layers are 100. The output layer contains two units, and a simple linear activation function is used. Finally, the soft-max module applies a soft-max function to the outputs in order to compute the overall confusion matrix.

Furthermore, an algorithm based on a mapping table was used to generate the training labels of each detectors from the phone transcription. In addition, each generated training set was normalized using the following formula:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where z is the new normalized value, x is the original value, μ is the training set mean value and σ is the training set standard deviation. These work-condition constrains were adopted in order to compare fairly the results presented in this paper with those shown in [6]. As previously stated, the detectors work in a frame-based paradigm, so that their performance was evaluated in term of frame error rate. Each frame was classified according to the neural network with the largest value. The global confusion matrix for the manner of articulation manner

Table 2. Phoneme accuracies (as percentages) for the manner of articulation attributes using EαNet architectures. Confusion Matrix of the manner attributes.

%	Vowel	Fricative	Stop	Nasal	Appr	Silence
Vowel	**91,00**	1,38	1,53	1,26	4,64	0,19
Fricative	3,16	**88,06**	5,53	1,02	0,89	1,24
Stop	6,32	7,41	**81,03**	1,71	1,57	1,96
Nasal	9,65	2,44	3,25	**81,45**	2,20	0,90
Approximant	30,82	2,88	3,26	2,74	**59,11**	1,19
Silence	1,10	1,09	1,88	0,61	0,58	**94,74**

Table 3. Relative improvement to the Li's Improvement of the articulation manner classification over the baseline ANN [6]

%	Vowel	Fricative	Stop	Nasal	Appr	Silence
Vowel	**2,00**	-0,12	0,03	-0,54	-0,36	-0,01
Fricative	-0,54	**2,86**	-1,27	-0,18	-0,41	-0,46
Stop	-1,28	-3,59	**8,53**	-1,19	-0,53	-1,94
Nasal	-1,55	-0,06	-1,45	**3,95**	-1,00	0,10
Approximant	-1,48	-0,02	-0,44	-0,46	**2,61**	-0,21
Silence	0,00	-0,11	-1,32	-0,09	-0,32	**1,84**

attributes is given in Table 2. The $(p, q) - th$ element of the confusion matrix measures the rate of the $p - th$ attribute being classified into the $q - th$ class.

5 Conclusions

Incorporating acoustic-phonetic knowledge into Automatic Speech Recognition designs has been proven a viable approach to rise their accuracy. Manner of articulation attributes such as vowel, stop, fricative, approximant, nasal, and silence are examples of such knowledge, and they represent speech attributes. A set of six attribute detectors was designed based on the EαNet neural architecture and their performance has been studied. The evaluation demonstrates the effectiveness of this design for speech attribute classification, with an average 3.5% improvement with respect to the use of a traditional ANN approach, showing a maximum 8.5% improvement in the case of plosives.

Acknowledgements

Authors are indebted with Prof. Chin.-H Lee, for the insightful discussions on the topics and for his help defining the general experimental framework. Part of this effort was supported under the NSF SGER grant, IIS-03-96848 and NSF ITR grant, IIS-04-27413.

References

1. L. R. Rabiner, : A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, Vol. 77, No.2, pp. 257-286, 1989.
2. S. Haykin, : Neural Networks: a Comprehensive Foundation (2nd edition). Prentice Hall, 1998.
3. K. Kirchhoff: Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments Proc. ICSLP98, Sydney, Australia, 1998.
4. S. Davis, and P. Mermelstein: Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences. IEEE Trans. on Acoust., Speech and Signal Process., Vol. 28, No. 4, pp. 357-366, 1980.
5. K. F. Lee, H. W. Hon: Speaker-independent phone recognition using hidden Markov models. IEEE Trans. On Acoust., Speech and Signal Process., Vol. 37, No. 11, pp. 1641-1648, 1989.
6. J. Li, Y. Tsao and C.-H. Lee: A study on knowledge source integration for candidate rescoring in automatic speech recognition. Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, December, 1998, 891-894.
7. R. P. Lippmann: Speech recognition by machines and humans Speech Communication, Volume: 22 (1), July, 1997, pp. 1-15.
8. S.Gaglio, G. Pilato, F.Sorbello and G.Vassallo: Using the Hermite Regression Formula to Design a Neural Architecture with Automatic Learning of the 'Hidden' Activation Functions AI*IA99:Advances in Artificial Intelligence - Lecture Notes in Artificial Intelligence 1792 - Springer Verlag - pp. 226-237, 2000.

9. G.Pilato, F.Sorbello and G.Vassallo: An Innovative Way to Measure the Quality of a Neural Network without the Use of the Test Set IJACI International Journal of Advanced Computational Intelligence - Vol. 5 No 1, 2001, pp:31-36.
10. A.Cirasa, G.Pilato, F.Sorbello and G.Vassallo: An Enhanced Version of the aNet Architecture: Automatic Pruning of the Hermite Orthonormal Functions Atti del Workshop "Apprendimento e Percezione nei Sistemi Robotici" - Parma, Italy - 29-30 November 1999.
11. Lee, C.-H.: From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition, Proc. ICSLP, 2004.
12. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. U.S. Dept. of Commerce, NIST, Gaithersburg, MD, February 1993.