On the Use of a Multilingual Neural Network Front-End

On the Use of a Multilingual Neural Network Front-End

Stefano Scanzio [1], *Pietro Laface* [1], *Luciano Fissore* [2], *Roberto Gemello* [2], *Franco Mana* [2],

[1] Politecnico di Torino, Italy, [2] Loquendo, Torino, Italy

{Stefano.Scanzio, Pietro.Laface}@polito.it

{Luciano.Fissore, Roberto.Gemello, Franco.Mana}@loquendo.com

Abstract

This paper presents a front-end consisting of an Artificial Neural Network (ANN) architecture trained with multilingual corpora. The idea is to train an ANN front-end able to integrate the acoustic variations included in databases collected for different languages, through different channels, or even for specific tasks. This ANN front-end produces discriminant features that can be used as observation vectors for language or task dependent recognizers.

The approach has been evaluated on three difficult tasks: recognition of non-native speaker sentences, training of a new language with a limited amount of speech data, and training of a model for car environment using a clean microphone corpus of the target language and data collected in car environment in another language.

Index Terms: ANN, ANN adaptation, Multilingual training

1. Introduction

One of the goals of the ongoing researches in speech recognition is robust modeling. The goal is to train models able to deal with variations due to speaker voice, to noise, and channel distortions.

Speaker independent models are customized to the acoustic-phonetic characteristics of a speaker by means of speaker adaptation techniques [1]. Adaptation is particularly important for non-native speakers.

Robustness against noise and channel distortions is obtained by appropriate filtering the acoustic features during the acoustic analysis [2], or by performing model compensation [3][4].

Moreover, recently some attention has been devoted to reduce the development efforts for training new languages, in particular for languages having low resource in terms of available speech corpora. Model adaptation and model sharing with a new language have been proposed in [5]. The use of a multi-lingual phoneme set with some common phones has been proposed in [6] and [7].

In this work we propose a multilingual ANN architecture, similar to the one proposed in [7], that attempts to face the three afore mentioned problems. It allows the acoustic particular we train ANN front-ends to produce discriminant task dependent recognizers.

Stefano Scanzio is supported by a Lagrange Project scholarship of the CTR Foundation and Institute for Scientific Interchange Foundation, Torino, Italy.

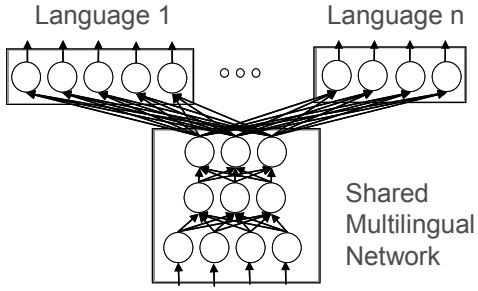

Figure 1. *Multilingual network architecture*

In this work we evaluated the discriminative front-end architecture on three difficult tasks.

First we trained a multilingual front-end using 10 language corpora, and we used it to train very simple language dependent models. The quality of these models was tested on a set of benchmark databases. Two of these tests focused on the recognition of non-native speaker sentences.

In the previous tests, the test languages were among the languages seen in training. Thus, we trained also a simple model for a previously unseen language using the same multilingual front-end and a limited amount of data of the new language.

Finally, about the robustness against noise and channel distortion, an original approach is proposed, where we use the multilingual framework to train a model for an automotive environment using clean microphone corpora of a language and data collected in car environment in another language.

The paper is organized as follows: Section 2 introduces the architecture of the multilingual network. Section 3 presents a large set of performed experiment, and the conclusions are given in Section 4.

2. Multilingual network architecture

The most popular Artificial Neural Networks used for modeling the phonetic units are Multi-Layer Perceptrons (MLP) with two or more hidden layers and a softmax output layer. The use of the softmax output layer allows discriminative models to be trained, and posterior probabilities of the phonetic units to be obtained [8].

The Loquendo-ASR [9] uses a hybrid HMM-ANN model, where each phonetic unit is described in terms of a gender

Table 1. *Features of the train corpora for the 10-language multilingual network*

Language	Dutch	French	German	Greek	Italian	Portuguese	Spanish	Swedish	UK English	US English
Phonetic units	775	924	950	636	683	719	451	905	948	906
Sentences	37897	55104	40683	51787	48176	41420	34067	55334	47530	50754
Time (hours)	28.6	45.0	36.2	48.7	35.6	42.6	25.0	40.3	24.8	39.6

independent single or double state left-to-right automaton with self-loops [10]. The phonetic units are stationary context-independent phones that consist of a single state, and diphone transition units, modeled by a double state. The ANN is a three layer Multilayer Perceptron that estimates the posterior probability of each unit state, given a context window of 7 frames. A frame vector includes 13 RASTA PLP parameters [11] and their first and second derivatives.

Figure 1 shows the architecture of our multilingual network. It consists of a set of independent softmax output layers, one for each language contributing to the training set, and two shared hidden layers. Each softmax block includes a language dependent set of phonetic unit models. The input layer has 273 nodes corresponding to the sliding window of 7 frames of 39 features. The first and second hidden layers have 600 and 500 fully connected nodes respectively. The output of the last hidden layer of the network can be considered a transformation of the input frames into language independent discriminative features, which are supplied as input to a final linear classifier.

It differs from the one proposed in [7], which defines a set of common phones in the case of similar sounds from different languages.

2.1. Multilingual network training

Each training sentence has associated its phonetic transcription and its language. During training, back-propagation is performed from the subset of the output states corresponding to the sentence language only. The weights of the output layer are, thus, language dependent, whereas the weights of the shared layers are updated using the sentences of any available language in the training set.

The multilingual network that has been used in all the experiments described in Section 3 has been trained using subsets of 10 language corpora available for producing the models of the Loquendo ASR speech recognizer. These subsets, detailed in Table 1, include a total of 462752 sentences and about 366 hours of telephone speech utterances. The distribution of the number of sentences and of frames per language is not uniform because of the lack of data for some languages. However, the relatively small difference among the corpora size should not affect the results as it would be possible if the common network was trained using much more data for given language than for any other language.

It is worth noting that the use of a multilingual network does not increase the processing time in testing because the resulting model has the same architecture of a standard network with the lower layers identical for all languages, and language dependent weights in the output layer.

2.2. Multilingual front-end

Since the lower part of the multilingual ANN acts as a front-end that integrates the acoustic variations included in databases of different languages, the shared layers of the network can be embedded in the front-end module of the recognizer. The multilingual features produced by this language independent front-end are more discriminative and more robust to inter-language variations. These features can be used to train classifiers that are not necessarily based on Neural Networks. They can be used, for example, as observation features in a GMM-HMM framework in the "Tandem" approach [12], after performing a logarithmic transformation and dimension reduction through principal component or linear discriminant analysis.

3. Experiments

Three sets of experiments were performed to test the capability of the multilingual model in different application environments.

The first set of tests aims at comparing the results of the standard and the multilingual network tested on different languages and tasks.

3.1. Tests with native and non-native speakers

The multilingual model has than been tested using a set of Italian, German and English databases just appending the last layer of the corresponding language to the multilingual network front-end.

The results of the baseline and of the multilingual networks are compared in Table 2, where:

- *SpeechDat2* is a continuous speech subset of the Italian and the German SpeechDat 2 databases. The Italian and German *SpeechDat2* databases include 4293 and 3395 sentences respectively. The Italian and German systems have a vocabulary of 9.4K and 7.5K words respectively. No language models has been used for these tasks.
- The *Application Words* database consists of 9006 Italian and 11693 German utterances of application words. Since the databases include pronunciations of a small number of different application words, the vocabulary is augmented with the most frequent words appearing in the training database of each language, for a total of about 2000 words.
- *Connected Digits* refers to a database of 84911 Italian and 118462 German connected digits.
- *Spelling* is a letter recognition task for a total of 8443 Italian and 5585 German spelled words.
- *Italian Non-native* is a database of 5106 English command words spoken by Italian speakers; the test vocabulary has 361 words.
- *Wsj0* is the 5K Wall Street Journal 0 corpus (recognition is performed using the trigram Language Model provided by Lincoln Labs).
- *Wsj1* is the SPOKE 3 test part of the Wall Street Journal 1 database including 10 American English non-native speakers.

The results are given in terms of Word Accuracy (%).

Excluding the tests with non-native speaker databases, *Italian Non-native* and *Wsj1*, the results obtained using the multilingual models are slightly worse than the baseline ones. This is not surprising because the baseline models are trained using only the language specific corpus and can specialize on it. However, also in this condition, in all the tests the loss is less than the 10% relative. On the other hand, the tests with the two non-native databases show a small performance increase, as shown in the last two lines of Table 2.

The multilingual network, trained with the combination of different languages, better accounts for foreign speaker pronunciations.

Further experiments were performed for English, using the multilingual network as a front-end. In particular, a two layered ANN was trained using the features of the multilingual front-end. Table 3 shows, in column 4, that the use of two layers on top of the multilingual front-end is successful compared to the baseline results. As already noticed, training from scratch an English model with the same number of layers and weights of the multilingual network, but trained with the English corpus only, produces a model giving (slightly) better performance.

3.2. Training a network for a new language

The multilingual network, trained as illustrated in Section 2.1, has been used as a front-end for the training a network for Polish. The experiment was performed to check if the features extracted by the multilingual front-end were able to cover the acoustic-phonetic space of a language never seen before, and to verify if the Polish specific phones could be classified by linearly separating language independent features. More precisely, we compared the performance of two Polish networks trained using an increasing amount of training data. The first network is the standard one, trained from scratch. The second one is trained by keeping fixed the weights of the multilingual front-end, thus, it is a single layer network with softmax output nodes. The weights of the output layer are initialized by small random values. The two networks have 642 output nodes corresponding to the states of the phonetic unit models defined for Polish.

Figure 3 shows the results on a test set of 1174 phonetically balanced Polish sentences with a vocabulary of 4432 words, no language model was used in these tests.
Using less than 10 hours of training data, the multilingual network performs better than the three layer MLP network trained with the same amount of data. Increasing the dimension of the training corpus, the complete set of weights of the network can be reliably trained, thus the standard network achieves better performance than a single layer network using the features of the multilingual front-end. However, even using all the available Polish training data (20.8 hours), the Word Accuracy of the multilingual network compared with the standard one shows a small 5% relative decrease, from 79.5% to 78.5%.This experiment shows that whenever a small amount of training data of a new language is available, the use of a multilingual network can take advantage of the prior knowledge induced by the variety of the other languages that contributed to train the front-end model.

Table 2. *Comparison of (%) Word Accuracy results using the baseline and the multilingual network tested on different languages and tasks.*

Language	Database	Baseline	Multilingual
Italian	*SpeechDat2*	71.9	68.8
Italian	*Application Words*	96.9	96.8
Italian	*Connected Digits*	97.9	97.6
Italian	*Spelling*	61.0	60.6
German	*SpeechDat2*	67.4	64.5
German	*Application Words*	95.4	94.7
Gernam	*Connected Digits*	95.3	93.0
German	*Spelling*	78.3	78.2
English	*Wsj0*	83.5	81.8
English	*Italian Non-native*	49.0	**50.3**
English	*Wsj1*	41.8	**43.0**

Table 3. *Use of the multilingual network as a front-end for training a new English ANN.*

Database	Baseline 3layers	Multil. 3layers	Multil. 4layers	Baseline 4layers
Wsj0	83.5	81.8	**84.0**	85.6
Italian Non-native	49.0	50.3	**53.4**	51.4
Wsj1	41.8	43.0	**44.2**	45.4

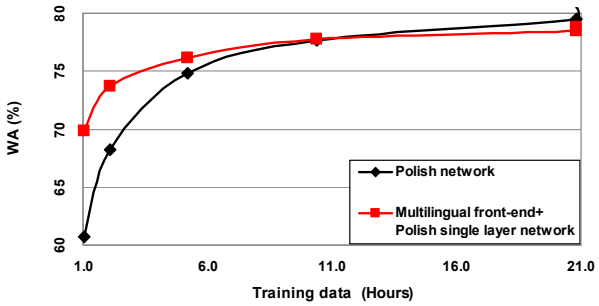

Figure 3. *Test of two Polish networks trained with increasing amount of data. The first network is the classical three layer MLP, the other one is a single layer network, trained using the discriminative multilingual features*

3.3. Car environment adaptation

Recently speech recognition in the automotive environment is attracting relevant resources due to its commercial importance. Since the mismatch between the training and test environment has a dramatic impact on the recognition performance, it is important to use training data consistent with the test environment. The use of specific database collected on cars allows robust recognizers to be trained compared with the ones trained with clean microphone data. Car databases, however, require considerable efforts to be collected. Thus, they are expensive when available, and often not existing at all for many languages.

A possible procedure to reduce the mismatch between training and test conditions in car environments is to add

Table 4. *Test of the US English - Italian multilingual network on US English car environment data.*

Database	Channel	Base	Multilingual	Relative error %
Application Words	ch0	78.5	73.3	-24.2%
	ch1	52.9	54.5	3.4%
	ch2	48.1	48.6	1.0%
	ch3	49.7	52.4	5.4%
Connected Digits	ch0	93.9	94.7	13.1%
	ch1	74.5	82.5	31.4%
	ch2	71.0	78.2	24.8%
	ch3	71.3	77.8	22.7%
Spelling	ch0	74.8	79.4	18.3%
	ch1	50.5	57.0	13.1%
	ch2	47.3	53.6	12.0%
	ch3	47.3	53.9	12.5%

automotive noise to the available telephone or microphone databases. We propose, instead, to embed in the shared layers of a multilingual network the information related to car noise. In these experiments we assume that only clean microphone data of a target language are available (US English), but that we have at our disposal data acquired in car environment for a different language (Italian).

In particular, in the framework of the multilingual approach, a common ANN front-end and a special US English network was trained using the TIMIT and Wall Street Journal 0 microphone data, and the Italian SpeechDat-Car corpus [13][14]. Both databases were down-sampled to 8 KHz, because the front-end of the multilingual ANN was implemented for telephone speech.

The tests have been performed on three US SpeechDat-Car databases collected with 4 different microphones [15]. These tests consist of 3728 *Application Words* utterances with a vocabulary of 2166 command and phonetically balanced words, 12013 *Connected Digits* and 1336 spelled words (*Spelling*).

In Table 4, ch0 refers to a close-talking microphone, while ch1, ch2 and ch3 refer to three far talk microphones placed on the car dashboard. Ch1 is the microphone closest to the speaker, ch3 is the farthest, and ch2 is in the middle between ch1 and ch3. The results of Table 4 show that this technique offers a good relative error reduction in far talking conditions. Surprisingly, excluding the *Application Words* task, there is no performance reduction for the clean – close-talking - ch0 channel condition.

4. Conclusions

A multilingual neural network architecture has been presented that allows merging the acoustic-phonetic knowledge of different languages in one model. The performance obtained with this model are slightly worse than the ones achieved using a network trained for a specific language, due to the fact that the mono-language network focuses on its target language only, whereas the multilingual has to learn many languages. On the other hand, the multilingual network elegantly faces some problems like non-native speaker pronunciations, insufficient training data for a new language and absence of training databases for a specific noisy environment.

Further experiments have been planned to evaluate the benefits of using a common silence unit across languages. The use of a different learning rate for each softmax output block, depending on the amount of training data available for the corresponding language need to be investigated, to reduce the possibility that the discriminative features are polarized toward the most represented languages.

Finally, the multilingual front-end ANN can substitute the random setup of the weights to speed up the training procedure from scratch of a network for a new language or application.

5. References

[1] G. Zavaliagkos, R. Schwartz and J. Makhoul, "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition," in Proc. ICASSP-95, pp. 676–679, 1995.

[2] A. de la Torre, A.M. Peinado, J.C. Segura, J.L Perez-Cordoba, M.C: Benitez and A.J. Rubio, **"**Histogram equalization of speech representation for robust speech recognition**,"** IEEE Transactions on Speech and Audio Processing, Vol. 13, n.3, pp. 355–366, 2005.

[3] M. Gales and S. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," IEEE Transactions on Speech and Audio Processing, Vol. 5, pp. 352–359, 1996.

[4] M. Gales, "Predictive model-based compensation schemes for robust speech recognition," Speech Communications, 25(1-3), pp. 49–74, 1998.

[5] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling" Speech Communication, Vol. 35, no. 1-2, pp. 31–51, 2001.

[6] J. Marino, A. Moreno, and A. Nogueiras, "A First Experience on Multilingual Acoustic Modeling of the Languages Spoken in Morocco," in Proc. INTERSPEECH-2004, pp. 833–836, 2004.

[7] S. Dupont, C. Ris, O. Deroo, and S. Poitoux, "Feature Extraction and Acoustic Modeling: An Approach for Improved Generalization Across Languages and Accents," in Proc. of ASRU 2005, San Juan, Puerto Rico, 2005.

[8] H. Bourlard and N. Morgan, "Connectionist Speech Recognition: A Hybrid Approach", Kluwer, 1994.

[9] http://www.loquendo.com/en/technology/asr.htm

[10] D. Albesano, R. Gemello, and F. Mana, "Hybrid HMM-NN for Speech Recognition and Prior Class Probabilities," in Proc. of the 9th International Conference on Neural Information Processing, Vol. 5, pp. 2391–2394, 2002.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", The Journal of the Acoustical Society of America, Vol. 87, no. 4, Apr. 1990, pp. 1738–1752, 1990.

[12] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional hmm Systems," in Proc. ICASSP-00, Vol. 3, pp. 1635–1638, 2000.

[13] H. van den Heuvel, J. Boudy, R. Comeyne, S. Euler, A. Moreno, and G. Richard, "The SpeechDat-Car Multilingual Speech Databases for In-Car Applications: Some First Validation Results," in Proc. Eurospeech-1999, pp. 2279–2282, 1999.

[14] Available at http://catalog.elra.info

[15] P. Heeman, D. Cole, A. Cronk, "The U.S. Speechdat-Car Data Collection", in Proc. Eurospeech-2001, pp. 2031–2034, 2001.