FEATURE EXTRACTION AND ACOUSTIC MODELING: AN APPROACH FOR IMPROVED GENERALIZATION ACROSS LANGUAGES AND ACCENTS

Stéphane Dupont[†], *Christophe Ris*[†], *Olivier Deroo*[‡], *Sébastien Poitoux*[‡]

[†]Multitel, Avenue Copernic 1, B-7000 Mons, Belgium
[‡]Acapela Group, Boulevard Dolez, 37, B-7000 Mons, Belgium
{dupont,ris}@multitel.be - {deroo,poitoux}@acapela-group.com

ABSTRACT

The paper proposes a solution that brings some advances to the genericity of the ASR technology towards tasks and languages. A non-linear discriminant model is built from multi-lingual, multi-task speech material in order to classify the acoustic signal into language independent phonetic units. Instead of considering this model for direct HMM state likelihood estimation, it rather operates as a first stage to produce discriminant features that can be further used in cascade with a traditional task/language specific ASR system.

This first stage structure is expected to achieve a strong modeling of the cross-language variability of speech that can better handle pronunciation variations due for instance to regional and non-native accents. Moreover, the flexibility of this architecture still allow the development of small task/language dedicated ASR systems as a second stage structure, possibly with small amount of data.

The benefit of this architecture is demonstrated through a fine analysis of modeling performance at the phoneme level and on two different isolated word recognition tasks featuring accent variabilities.

1. INTRODUCTION

Many published researches to date have highlighted the needs to increase the independence of the technology towards factors such as noise and speaker voice, and hence increase the genericity of the technology. Task independence is nowadays achieved through the use of context dependent modeling, more specifically triphone models. Robustness against several noise environments is achieved by tackling the acoustic analysis part of the ASR chain. Speaker independence is handled with speaker compensation or adaptation mechanisms. Speaking styles are handled with some form of multi-style training. All these techniques are nowadays considered as state-of-the-art.

Recently, language independence has also been sought, with the goal of reducing the development effort for additional languages. Several attempts towards language genericity have been made, mostly in view of facilitating the fast porting to minority languages and domains for which few speech corpora are available. In [1], model adaptation and model transfer to a new language are proposed and investigated. Different methods for building the cross-language mapping of the phonemes are explored. In [2], the use of a multi-lingual phoneme set with some common phones is proposed. No degradation in comparison to uni-lingual models has been observed. In [3], it is shown that recognition rates can be improved when the phonetic similarities across dialects are obtained using decision tree clustering, instead of a priori SAMPA inventories. In [4], straight transfer of uni-lingual acoustic models from several source languages towards a target language is investigated, where phoneme mapping is defined based on the SAMPA symbols, or else on a data-driven method. These papers generally show that if similarities between languages exist, efficient transfer across languages is possible, and may be desirable. The more convincing results build on multi-lingual acoustic models based on multi-lingual phoneme inventories.

These conclusions seem natural as, indeed, speech sounds from similar languages exhibit similar structure as they are all produced by human physiological systems. Such similarity between sounds from several languages has been acknowledged in studies about spoken language, where the SAMPA [5] alphabets share common symbols across languages and, more recently, where generic articulatory gestures have been proposed [6, 7].

This paper further advances along this path. Here, however, multi-linguality is not built in the acoustic models used to estimate the HMM state likelihoods but rather in a novel method that attempts to extract discriminant features that present some form of genericity across several factors,

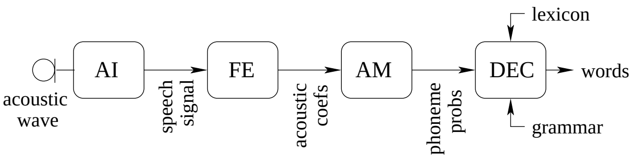

Fig. 1. *A typical ASR system: microphone, audio interface (AI), front-end (FE), acoustic model (AM) and word decoder (DEC).*

including the target language. Language and domain specific acoustic models are then built using these features, allowing language and task specificity if required, while also bringing the benefit of detailed modeling and robustness to any task and language. The possible advantage with respect to straightforward transfer of models lies in the fact that same phoneme symbols can be pronounced differently in different languages, and hence straightforward transfer may not be optimal. More particularly, pronunciation variations due to co-articulation may strongly differ from a language to another.

In this paper, the developed models are evaluated on tasks characterized by strong regional or foreign accents, highlighting that besides model transfer (not investigated here), stronger robustness is achieved to such phenomena.

The remaining of the paper is organized as follows. The next section describes the approach; the discriminant feature extraction, the modeling structures and the corpora that have been used. Section 3 presents an analysis at the phonetic level of the performance of the different models for native and foreign speakers in German. The recognition experiments presented in section 4 demonstrate the benefits of the proposed approach in cases of non-native and regional accents of English language.

2. APPROACH

2.1. Architecture

The general architecture of an automatic speech recognition system is depicted in Figure 1.

The feature extraction (FE) block is designed to extract an intermediate representation of the speech signal that ultimately may be independent of different factors, such as task, background noise, speaker characteristics, speaking style and, as proposed in this paper, languages.

The implementation of the FE module that we propose in this study lies on the non-linear discriminant analysis (NLDA) described in [8]. In our approach, a multi-layer perceptron (MLP), trained on a large amount of data from several languages, is used to transform the acoustic signal into a “language independent” representation (Figure 2). In that framework, a two hidden layer MLP is trained. During the recognition phase, we consider the outputs of the last hidden layer as our new feature parameters. The size of this layer can be optimized or adjusted to get the desired number of features (in our experiments, it has been set to 75). These parameters can then be used as usual acoustic features to further build task/language dependent acoustic models used by HMMs.

consonants				vowels			
p	b	t		i	i:	I	
d	k	g		e	e:	E	
ts	dz	tS	dZ	a	a:	A	A:
f	s	S		o	o:	O	
v	z	Z		u	u:	U	
m	n	N		y	y:	Y	
r	R	l	L	2	2:	9	
j	J	w	H	3	6	{	
h	x	?		V	Q	@	
C	T	D	4	e∼	a∼	o∼	
				aI	eI	OI	OY
				aU	@U		
				I@	e@	U@	

Table 1. *Multilingual phoneme inventory used for training the multi-lingual MLP (refer to SAMPA for detailed significance).*

We used MLPs as they are known to provide powerful locally discriminant acoustic modeling, and simplify the integration of contextual information from the original feature space.

A crucial part in the development of this system concerns the selection of the *phonetic* units used as targets for the training of the multi-lingual MLP. On one side multi-linguality is desired, which probably requires an appropriate coverage of speech sounds specific to several languages. On the other side, some cross-language generalization is also sought, which implies the use of common targets in the case of similar sounds from different languages. As a first attempt, we found it natural to start from the SAMPA phonetic alphabet, which allows defining a library of speech sounds, some of which are common to several languages. The phoneme superset that we defined covers the four languages that we used (see section 2.4), and contains 74 different symbols (75 with silence), listed in Table 1.

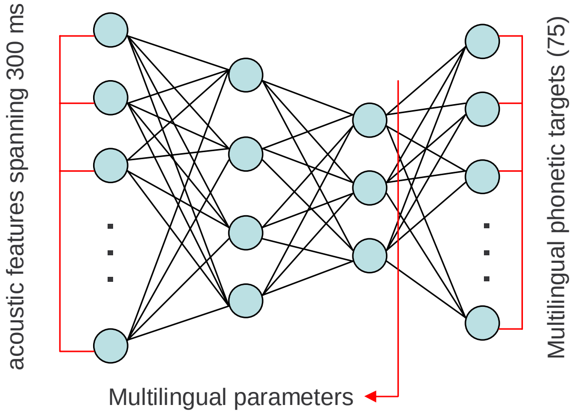
acoustic features spanning 300 ms
Multilingual phonetic targets (75)
Multilingual parameters

Fig. 2. *Estimation of multilingual acoustic parameters through a non-linear discriminant transformation (NLDA) of spectral based features.*

We acknowledge that this rather simple approach presents some weaknesses. Indeed, in the case of strong foreign accents, the speaker would rather fall back on sounds that are taken from his mother language. Based on the SAMPA symbols, the pronounced and target phonemes may have different symbols while cross-lingual modeling would probably be improved by using common symbols (think about the English, French and Italian /r/ for instance).

Given recent studies about speech sounds and cross-linguality, an alternative approach would be to use articulatory targets that are more closely related to the production mechanism [6, 7]. This has not been studied yet by our team, but we believe that this sound methodology should be investigated.

2.2. Advantages

In effect, the simple architecture and training procedure described here attempts to make use of more diverse training data, providing a wider coverage of speech sounds. Hence, besides the increased genericity across languages and accents that is expected when corpora in several languages are used, the approach consisting of pooling together large amounts of training data naturally offers a higher potential for investigating more detailed modeling techniques [9] as well as methods that can relax some of the typical underlying assumptions of HMM systems. More and more diverse data can make it possible to investigate models exhibiting less constraining statistical structures. In this paper, we draw on previous results that show that models can benefit from feature spaces covering larger time segments (up to 310 ms) [10], hence relaxing the HMM output independence assumption. In order to perform fair comparisons with the uni-lingual systems, this large time-span configuration has been implemented for all our acoustic models besides the typical feature space configuration (static feature and temporal derivatives). For each of the following experiments, we only present the best feature space configuration.

From the technological perspective, the proposed architecture may also yield some benefits in terms of footprint of multilingual systems, as modeling parameters are partially shared across languages. Naturally, reduced application development time can also be expected as wider language portfolio can be achieved easier.

It is important to emphasize that the multilingual architecture described here can be cascaded or inserted in any ASR architecture. The use of ANNs to get discrimination across several languages allows a lot of flexibility in the choice of input features, even with strong correlation between the feature vector elements. Also, the discriminant features obtained can be used as inputs to different kinds of acoustic models; the hybrid HMM/ANN system in our case but could also be mixtures of Gaussians, through a structure know as “Tandem” processing [11].

2.3. Feature extraction & acoustic modeling

The reference front-end is based on PLP processing [12] but MFCC processing would yield similar recognition performance. Additive and channel noises are handled using a combination of Wiener filtering and temporal trajectory filtering.

This feature set defines the first-level features from which language/domain independent discriminant second-level features are computed by a non-linear transformation implemented using a MLP (see previous section). As stated above, this MLP owns two hidden layers; the first one has 4000 hidden nodes, the second one, 75, totaling 1.2 M parameters.

On the acoustic modeling side (cf. Figure 1: AM module), our speech recognition systems are based on the hybrid HMM/ANN architecture [13] where ANNs - typically multilayer perceptrons (MLP) - are used to estimate the HMM state likelihoods. This collaboration between ANNs and HMMs has proven its efficiency on many different speech recognition tasks. In its “classical” form, used in this paper, a single ANN is trained in order to classify the frames of acoustic features into language dependent acoustic units (phonemes, diphones, ...). However, in the architecture proposed in this paper, classification does not operate on the usual first-level feature set but rather on the multi-lingual second-level discriminant features.

Language	Corpus	Source	Speakers	Hours
French	EUROM-1	ELDA	60	10
	Bref 120	ELDA	120	115
	BDSONS	ELDA	32	10
Italian	EUROM-1	ELDA	60	16
	SPK	ELDA	100	6
German	EUROM-1	ELDA	60	20
	PhonDat-1	ELDA	201	30
US English	WSJ0	LDC	123	25
UK English	WSJCAM0	LDC	140	19
Total	9 corpora	-	896	251

Table 2. *List of databases used for training the multi-lingual feature extraction engine. The amount of data (hours of speech) that has actually been used from each corpus is also indicated in the table.*

2.4. Databases

Nine corpora have been used for developing the system that will be evaluated in the next sections. We have chosen two Germanic languages (English and German) and two Latin languages (French and Italian). This selection of corpora provides a large amount of training data, speaker and style coverage. This amounts to 250 hours of speech and about 900 speakers (Table 2).

In the following experiments, the new architecture is compared to uni-lingual systems trained on single corpora, for specific recognition tasks, namely the ELDA EUROM German corpus, the LDC Wall Street Journal corpus (WSJ0) and WSJCAM0, the ‘British version’ of WSJ0. Note that the second-stage MLPs of the multi-lingual architecture are trained on the same data. Experiments have shown that those second-stage MLPs can be quite small (about 100,000 parameters) without performance loss.

3. PHONETIC ANALYSIS

In this section, we study the impact of the multi-lingual modeling on a free phoneme classification task, comparing to the uni-lingual ASR system. Test data consist in a subset of the ELDA STRANGE II SC-BAS corpus. This database contains German sentences spoken by native and foreign speakers. Besides the native speakers, we retained the French, Italian, US and UK British speakers, all languages used to train the multilingual system, and also the Dutch and Spanish speakers for exploring the ability of the approach to extend to other languages. The test sets contains about 2300 phonemes per mother language and all the speakers are pronouncing the same set of 100 sentences. Note that the foreign speakers are reported as being fluent in German.

Table 3 shows the phoneme classification rates. The uni-lingual system is a hybrid HMM/MLP (4000 hidden nodes, 1.1 M parameter) trained on the EUROM German database.

The reference performance on this task for the German native speakers using the uni-lingual system is 58.3% classification rate. The average performance for non-native speakers using the same model is only 49.0%. It is interesting to note that this degradation is almost completely compensated with the multi-lingual feature extraction technique. Note also, that the improvement is still in the same order for Dutch and Spanish, two European mother languages not seen during the training of the multi-lingual system, but probably phonetically well covered by the four training languages.

Mother language	Uni-lingual rec. rate (%)	Multi-lingual rec. rate (%)	Relative error reduction (%)
US English	49.0	57.4	16.5
UK English	40.2	48.5	13.9
French	51.7	59.2	15.5
Italian	51.5	59.0	15.5
Dutch	61.2	69.3	20.1
Spanish	40.7	49.0	14.0
Average	49.0	57.1	15.9

Table 3. *Phoneme classification rates with multi-lingual modeling compared to uni-lingual system for different foreign accents.*

In order to have a finer analysis of the performance of different acoustic models, we not only evaluate phoneme recognition rates but also the average posterior probabilities of problematic phonemes. At the local level (frame or phoneme, compared to word or sentence), this measure gives a more reliable idea of the intrinsic performance of the acoustic models than phoneme classification rates. The higher this value is, the better the acoustic model globally matches the acoustic data.

For each mother language, we have selected the phonemes which classification rates were significantly worse than the average recognition rate, assuming those are symptomatic of mispronunciations. Average posteriors of these phoneme subsets computed with the unilingual and the multilingual systems are shown in Figure 3. The systematic increase of the average posteriors is a good indicator of the benefit of the new architecture to handle local pronunciation variations due to foreign accents.

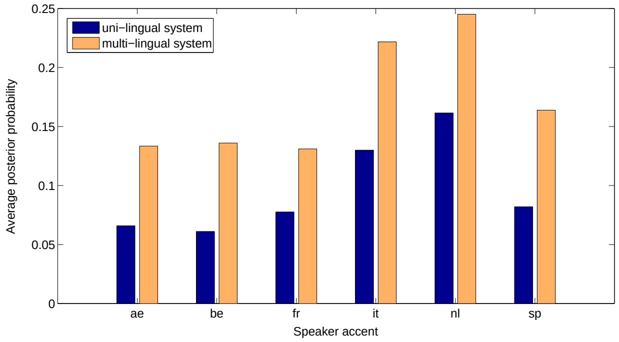

Fig. 3. *Comparison of average posteriors of problematic phonemes for the uni and multi-lingual systems for non-native accents. 'ae' stands for US English, 'be' for UK English, 'fr' for French, 'it' for Italian, 'nl' for Dutch and 'sp' for Spanish.*

4. EXPERIMENTS

Finally, word recognition experiments have been performed in order to understand and assess the possible benefit of this technique on user groups that have a strong regional or non-native accent. These accents are not or few represented in the available training corpora. Besides, depending on the level of practice of a language, acoustic specificities of the mother language exist (even for a SAMPA symbol that is present in both languages) and are more or less present, which may make generalization across languages interesting. The current understanding is however that proper handling of foreign accent will require improved acoustic modeling coupled with improved cross-lingual pronunciation modeling [14]. This is not addressed in the current paper.

4.1. Results & discussion

The first experiment addresses the problem of regional accents. The test data consist in isolated words in English taken from a vocabulary of 292 command & control words pronounced by eight speakers from different UK regions. The language specific acoustic models have been trained on WSJCAM0.

The overall robustness of the new architecture to regional accent is improved by about 28% relative (Table 4), the multi-lingual MLP offering a much better intrinsic modeling of pronunciation variabilities through the coverage of different languages and by the contribution of much more speech training material.

The second experiment consists in isolated word recognition of English words pronounced by 11 French speakers. The vocabulary is composed of the radio alphabet and the digits (36 words). The language specific acoustic models have been trained on WSJ0. The word phonetic transcriptions are taken from US dictionaries. No specific lexical adaptation has been performed.

System	Reco. rate	Rel. error red.
Uni-lingual	87.3%	-
Multi-lingual	90.8%	27.6%

Table 4. *Word recognition rate and relative error rate reduction for different UK accents. 8 speakers uttering command & control word (292 word lexicon).*

The recognition accuracy is improved by a factor of two on this task (see Table 5). The benefit of the cross-language modeling is here clearly demonstrated. Indeed, strong pronunciation digressions have been observed in this test set. For instance, mapping of English phonemes to French phonemes have been clearly identified. Probably the performance could still be improved by combining with adequate lexical modeling.

System	Reco. rate	Rel. error red.
Uni-lingual	85.9%	-
Multi-lingual	92.8%	48.9%

Table 5. *Word recognition rate and relative error rate reduction for a non-native accent. French speakers uttering English words (36 word lexicon)*

4.2. Ongoing and future work

This work provides one of the building blocks towards improved genericity and performance of the feature extraction and acoustic modeling components. A similar structure (based on discriminant feature extraction followed by acoustic modeling) known as multi-band, has been proposed in the past to provide increased robustness to background noise. A previous paper [15] summarized work on that approach and confirmed its potential. In that paper, the sub-band discriminant ANNs were developed on task/language specific training corpora. Joining these two approaches makes sense as task/language independent noise robust sub-band discriminant features could then be produced.

On another side, we feel that the definition of the multi-lingual target phoneme inventory for training the NLDA module is probably a key aspect of the proposed technique. The use of the SAMPA set of symbols is probably not optimal and we should rather focus on a really language independent set of symbols, describing some sort of “universal” language. Studies on articulatory features may be good inspirations.

5. CONCLUSIONS

We have proposed a two-stage architecture for improving the genericity of ASR systems towards task, speaking styles and speaker accents. In the first stage, a generic MLP is used to extract task/language “independent” acoustic features. This MLP was beforehand trained on a large amount of multi-lingual/multi-task speech material labeled with a common inter-language phoneme set. In a second-stage, those discriminant acoustic features feed a standard ASR system, typically focused on a particular task and a particular language.

Building a system on more and more diverse training data not only brings some advance towards the genericity of ASR technology but also opens the way to more detailed modeling, possibly relaxing some constraints of traditional ASR systems.

The performance of this architecture has been brought out on the particular problem of the regional and non-native accents. Phoneme and isolated word recognition experiments have been conducted and led to 27.6% and 48.9% relative error rate reduction for regional accents and non-native accents respectively.

6. ACKNOWLEDGEMENTS

This work has been partly supported by the EU 6th Framework Programme, under contract number IST-2002-002034 (DIVINES project). The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

7. REFERENCES

[1] T. Schultz and A. Waibel, *”Experiments on Cross-Language Acoustic Modeling”*, in Proc. of Eurospeech’01, Alborg, Denmark, Sept. 2001.

[2] J. Marino, A. Moreno and A. Nogueiras, *”A first experience on multilingual acoustic modeling of the languages spoken in Morocco”*, in Proc. of ICSLP’04, Jeju Island, Korea, Sept. 2004.

[3] M. Caballero, A. Moreno and A. Nogueiras, *”Data Driven Multidialectal Phone Set for Spanish Dialects”*, in Proc. of ICSLP’04, Jeju Island, Korea, Sept. 2004.

[4] A. Zgank et al., *”The COST 278 MASPER initiative - crosslingual speech recognition with large telephone databases”*, in Proc. of LREC’04, Lisboa, Portugal, May 2004.

[5] *http://www.phon.ucl.ac.uk/home/sampa/home.htm*

[6] A. Geumann, *”Towards a new level of annotation detail of multilingual speech corpora”*, in Proc. of ICSLP’04, Jeju Island, Korea, Sept. 2004.

[7] L. Deng, G. Ramsay and D. Sun, *“Production models as a structural basis for automatic speech recognition”*, Speech Communication, no. 22, pp. 93-111, 1997.

[8] V. Fontaine, J.M. Boite and C. Ris, *”Nonlinear Discriminant Analysis for Improved Speech Recognition”*, in Proc. of Eurospeech’97, Rhodes, Greece.

[9] L. Lamel, J.L. Gauvain, and G. Adda, *“Lightly Supervised and Unsupervised Acoustic Model Training”*, Computer Speech and Language, vol. 16, no. 1, pp. 115–129, Jan. 2002.

[10] S. Dupont, C. Ris, L. Couvreur and J.M. Boite, *”A study of implicit and explicit modeling of coarticulation and pronunciation variation”*, in Proc. of Interspeech’05, Lisboa, Portugal, Sept. 2005.

[11] H. Hermansky, D. Ellis and S. Sharma, *”Tandem connectionist feature extraction for conventional HMM systems*”, in Proc. of ICASSP’00, Istanbul, Turkey.

[12] H. Hermansky, *”Perceptual linear predictive (PLP) analysis of speech”*, The Journal of the Acoustical Society of America, vol.87, nr.4, april 1990, pp. 1738-1752

[13] H. Bourlard and N. Morgan, *“Connectionist Speech Recognition: A Hybrid Approach”*, Kluwer, 1994.

[14] C. Teixeira, I. Trancoso, and A. Serralheiro, *“Recognition of non-native accents”*, in Proc. of Eurospeech’97, pages 2375–2378, Rhodes, Greece, Sept. 1997.

[15] S. Dupont and C. Ris, *”Robust Feature Extraction and Acoustic Modeling at Multitel: Experiments on the Aurora Databases”*, in Proc. of Interspeech’03, Geneva, Switzerland, Sept. 2003.