

PORTING CHANNEL ROBUSTNESS ACROSS LANGUAGES

Françoise Beaufays, Daniel Boies, Mitch Weintraub

Nuance Communications
1380 Willow Road
Menlo Park, CA 94025
{francoise,boies,mw}@nuance.com

ABSTRACT

We propose a technique to port channel characteristics from one language to another. This allows us to build acoustic models in a target language that are robust to an environment for which we have no data in that language.

The approach consists in training broad phonetic class maximum likelihood linear regression (MLLR) transformations from a source language, and applying them in the target language. These transforms encapsulate the acoustic specificities of the environment without capturing language-specific characteristics that are difficult to port across languages.

As a case study, we consider the problem of building in-the-car GSM models for UK English, assuming that we have no GSM, and no car data in UK English, but that we have such data in German. We show that this technique can greatly reduce the error rate of the recognition system on English GSM car data.

1. INTRODUCTION

The last few years saw the deployment of a multitude of telephone-based automatic speech recognition (ASR) systems, such as voice-activated financial quote and transaction services, automated directory assistance, and voice portals. As these services become more widespread, it becomes more and more important that they deliver an equally good performance independently of the environment from which the users are calling, be that a quiet office or a moving car, with a land telephone line or a load-stressed cellular network. In addition, multinational companies who deploy such services typically want to offer similar products in several countries, in different languages.

From a technical point of view, although the field of robustness to acoustic distortions has made tremendous progress [1, 2, 3], it remains difficult to develop acoustic models that are intrinsically insensitive to factors such as high levels of noise or non-stationary channel distortion [4]. Instead, developers often recourse to collecting large amounts of data to adapt existing models to the desired acoustic conditions. This approach has a high cost, especially if it is to be repeated for each language in which the service will be deployed.

In this paper, we propose a method to use data collected in one environment and in one language to build acoustic models in another language, for the same environment. In other words, we try to capture the acoustic characteristics of an environment, and port them from one language to another. This of course adds an element of complexity to the robustness problem in the form of a linguistic mismatch between the source and target languages.

Cross-language modeling, *i.e.* the acoustic modeling of a target language from speech collected in other languages, has received a lot of attention recently [5]. The most successful methods proposed to date rely on data-driven matching and clustering techniques for acoustic units such as triphones or triphone-states (*e.g.* [6, 7, 8]). The goal of these approaches is to discover a detailed match between phonetic units in the source and target languages so that data aligned to the source models can be used to train or adapt the corresponding target models (*e.g.* [9]).

In our problem however, a close phonetic match is not required. Rather, we want to *ignore* language specificities and capture instead the channel and noise characteristics of the data. To this effect, we define broad phonetic classes that can readily be translated from one language to another. We train MLLR transforms [10] for these classes from a source language, and apply these transforms to the target language.

2. APPROACH

The approach assumes the availability of 3 ingredients:

1. acoustic models in the source language that are trained with “standard” data, for example this can be landline clean speech
2. acoustic models in the target language trained with similar data
3. adaptation data in the source language that was collected under the acoustic conditions on interest.

Broad phonetic classes (*e.g.* vowels vs. consonants) are defined, and the phonemes of the source and target languages are classified in the appropriate classes. A set of MLLR transformations is then trained to map the standard acoustic models in the source language to the desired environment in this same language. These transforms are then applied to the target-language models trained in the standard environment to make these models better suited to the desired environment. The choice of broad phonetic classes (rather than, for example, one class per phone) is justified by the fact that we want to avoid language-dependent idiosyncrasies to be captured by the transformations as these would be difficult to port to the target language.

In our experiments, we used a form of block-diagonal MLLR transformations [11] with mean adaptation only.

2.1. A Case Study

We illustrate this approach with a case study whose goal is to port acoustic characteristics from German to UK English. The baseline (standard) models are trained with landline telephone speech.

The target environment is recognition in the car, where the input speech is received by a handsfree kit, and sent to the recognition server through the GSM cellular network. We assume we have a small amount of training data collected for this environment in German, but none in English. We also assume we have an adequate amount of English GSM car data for testing. The target environment speech is severely mismatched with the standard training data (clean landline) for the following reasons:

1. Noise: the target data is corrupted by low-frequency car noise. This noise is mostly stationary, with a signal-to-noise ratio (SNR) that depends on the position of the microphone.
2. Reverberation: the target data is collected in a small enclosure.
3. Microphone: low-cost microphones introduce frequency distortions in the signal, and can potentially capture mechanical vibrations in the car.
4. Handsfree kit: handsfree kits typically have noise reduction and other processing that cause non-negligible distortions in the signal. This distortion may be non-stationary.
5. Cellular coding: GSM compression introduces frequency distortions and losses of information to which a standard speech recognizer may be more sensitive than the human ear.

The combination of these factors makes the test data extremely mismatched with the data used to train the baseline acoustic models, and the task of recognition is extremely delicate.

In the experiments below, we will present recognition results with a few MLLR configurations, evaluate the benefit of SNR-dependent models, and study the impact of acoustic differences between the adaptation and test data. Finally, we will examine a scenario in which, a posteriori, a small amount of matched English data becomes available for adaptation.

3. RECOGNITION SYSTEM

The recognition system is based on standard 3-state triphone hidden Markov models (HMM), with a Genone-based state clustering mechanism [12]. Unless otherwise stated, a one-pass architecture is assumed. The front-end is based on 27-dimensional mel-filterbank cepstral coefficient feature vectors, with cepstral mean subtraction and standard noise reduction implemented. The baseline models are trained with a large amount of phonetically rich, clean, landline telephone speech.

4. ADAPTATION AND TEST DATA

The German adaptation and English test data used in the first few experiments below (sections 5.2, 5.3) is perfectly matched from an acoustic point of view: subjects in both countries drove the same make of cars, equipped with the same microphones and the same handsfree kits, and used the same cellular standard, only their language was different. This setup is thus ideal to apply the information extracted from the adaptation data to the test application. In later experiments, we will relax some of these assumptions and use different adaptation sets (see section 5.4). The subjects are all native speakers. All the training and testing material is live speech (no simulations). The German adaptation data consists of 12K sentences with broad phonetic coverage. The test data is divided

in two sets, covered by two different grammars: name dialing (300 first-last name pairs), and telephone number dialing (10- or 11-digit strings with (zero, oh, nought) pronunciations for the digit "0"). The test data was collected under 3 different driving conditions: idle, slow (ca. 50km/h), and fast (ca. 120km/h). Each test speaker contributed to the 3 conditions. Each of the 6 test sets (2 tasks, 3 speeds) contains ca. 3000 sentences.

5. EXPERIMENTS

5.1. Performance of Unadapted English Models

Table 1 shows the error rates of the unadapted landline English models on GSM car data.

PHONE DIALING						NAME DIALING		
String ER			Digit ER			String ER		
Idle	Slow	Fast	Idle	Slow	Fast	Idle	Slow	Fast
47.8	89.9	96.6	9.3	32.7	45.3	7.2	25.0	47.0

Table 1. Error rates of landline models with English car test data.

Because the string error rates are so high on telephone number dialing, we also report per-digit error rates. For comparison, digit strings of equal length but collected from landline phones were recognized with more than 90% accuracy with these same models.

With this level of performance, it can be argued that only name dialing in an idle car would be a deployable application.

5.2. Performance of Cross-Language Adapted Models

In this section, we evaluate the performance of the English models after adapting them with 12K of German GSM car data. We considered two MLLR configurations:

1. 3 transforms: the phones are divided in speech and non-speech. One transform is estimated for non-speech phones, one for consonantal and one for non-consonantal speech phones.
2. 4 transforms: same as above but a specific transform is trained for the pause model. The other non-speech phones (human and other noises) share another transform.

PHONE DIALING						NAME DIALING		
String ER			Digit ER			String ER		
Idle	Slow	Fast	Idle	Slow	Fast	Idle	Slow	Fast
<i>3 TRANSFORMS:</i>								
43.1	80.4	92.3	7.9	20.6	33.9	6.0	15.3	29.1
<i>4 TRANSFORMS:</i>								
42.8	79.5	91.9	7.8	20.9	33.72	5.2	11.5	23.5

Table 2. Error rates of German-car-adapted English-landline models with English-car test data.

The resulting recognition performance is reported in Table 2. We see that cross-language MLLR adaptation with 3 transforms brings from 17 to 38% relative error-rate reduction on the name

dialing task depending on the driving condition, the largest gain being obtained at high driving speeds. A similar gain is observed in the per-digit telephone dialing error rates. String error rates decrease much less because of the relative independence of digit errors and of the length of the strings.

The benefit of using a specific transform for the pause model appears mainly in the name dialing task where it brings an additional 13 to 25% relative error-rate reduction.

In an additional experiment, we generated a series of more detailed phonetic classes with the help of phonetic decision trees. These classes were then used to define MLLR transforms to be trained and used as previously. We found that, within the limits of statistical significance, there was no benefit in further refining the phonetic classes and increasing the number of MLLR transforms. This is in agreement with other studies on block-diagonal or full covariance MLLR transformations which showed that such transforms are usually rich enough that a small number of transforms performs well [11].

5.3. Effect of Explicitly Modeling SNR Dependencies

Since the focus of this study is on acoustic mismatch, we considered modeling explicitly the noise level of the data. For this purpose, we labeled the 12K German adaptation sentences with their SNRs. We then split the adaptation in 3 roughly equal sets based on SNR (low, mid, high), and trained 4 MLLR transforms from each set. We applied these transforms to the baseline English models, thereby creating 3 sets of adapted models, one that is better suited for high-SNR test data, one for low-SNR data, and one for the mid-range. These models were placed in parallel in a 2-pass architecture whose first pass was adapted with SNR-independent transformations as above. Table 3 summarizes the SNR statistics of the 3 adaptation sets as well as those of the test data.

Data Set	Mean SNR	Std SNR
Adapt High SNR	31	4
Adapt Mid SNR	20	3
Adapt Low SNR	12	2
All Adapt Data	22	8
Test Idle	19	5
Test Slow	15	4
Test Fast	12	4

Table 3. SNR means and standard deviations for the adaptation and test sets.

Table 4 shows the performance of this 2-pass system. Comparing with Table 2, we see that explicit modeling of the noise level in the adaptation data brings 9 to 14% relative error-rate reduction on name dialing with respect to the previous best results, and 7 to 11% relative improvement on per-digit phone error rates.

With this setup, the relative error rate improvements with respect to the baseline unadapted models range from 36 to 58% for name dialing, and from 22 to 43% for digits. With this level of accuracy, name dialing becomes deployable up to fast driving speeds included, whereas telephone number dialing remains too difficult a task.

5.4. Adaptation with Acoustically Mismatched Data

In the previous experiments, we used adaptation data that was acoustically very well matched to the test data: same cellular network, same car type, same HF kit, same microphone. In this section, we explore the effect of relaxing some of these assumptions. To this effect, we consider 2 new German adaptation sets:

1. same type of cars as in the test sets, but different handsfree kits and microphones (one from the same manufacturer, one different),
2. different cars (a variety, mostly from different manufacturers), and different handsfree kits and microphones (different manufacturer).

The new adaptation sets were also collected through the GSM network.

We wish to compare baseline English models adapted with these and the previous adaptation sets. To make this a fair experiment, we have to try to control for all the other variables such as the amount of data, the gender of the speakers, the SNR profile of the data, and the phonetic coverage. We achieved this by subsetting the 3 adaptation sets to a common configuration of ca. 8K sentences with 25% female, 75% male, a roughly Gaussian SNR profile with an average of 25dB and a standard deviation of 10 dB, and approximately 20K words in each set with ca. 400 unique words, most of which have roughly the same frequency in each adaptation set.

Table 5 summarizes the error-rates with the 3 adaptation sets. All experiments are with 4 SNR-independent transforms. The baseline numbers are the same as in section 5.1. The “same cars – same kits” numbers are slightly worse than those in section 5.2 because of the downsampling of the adaptation set.

As expected, we see that acoustic mismatches between the adaptation and test data (line 3 and 4) result in lower performance, with a higher degree of mismatch (line 4) resulting in the highest error rate. For example, for name dialing at high speed, acoustically matched adaptation data gives a 46% relative error-rate improvement w.r.t. the baseline, a handsfree kit mismatch brings the improvement down to 40%, and a kit and car mismatch lowers it further to 36%.

5.5. Use of Language-Matched Data to Refine the Cross-Language Models

In this section, we consider the scenario where, after performing cross-language MLLR adaptation, we use the adapted models to run a pilot system and collect field data, this time in the target language (English in this case). The questions are (1) Can we make use of such data? (2) How much data is necessary? (3) Is cross-language MLLR still useful or does the newly collected data make the cross-language data obsolete?

PHONE DIALING						NAME DIALING		
String ER			Digit ER			String ER		
Idle	Slow	Fast	Idle	Slow	Fast	Idle	Slow	Fast
40.5	76.7	90.3	7.3	18.6	31.1	4.6	10.4	20.1

Table 4. Error rates with 2-pass SNR dependent models.

PHONE DIALING						NAME DIALING		
String ER			Digit ER			String ER		
Idle	Slow	Fast	Idle	Slow	Fast	Idle	Slow	Fast
<i>BASELINE UNADAPTED:</i>								
47.8	89.9	96.6	9.3	32.7	45.3	7.2	25.0	47.0
<i>SAME CARS – SAME KITS:</i>								
42.1	79.9	91.8	7.7	21.3	34.2	5.4	12.8	25.1
<i>SAME CARS – DIFFERENT KITS:</i>								
42.4	80.1	91.9	7.9	21.5	34.6	6.2	14.0	28.0
<i>DIFFERENT CARS – DIFFERENT KITS:</i>								
44.4	84.4	94.5	8.4	26.4	38.5	5.5	14.9	30.1

Table 5. Error rates on English GSM car test sets, with baseline landline models and with the same models adapted with different German adaptation sets.

To answer these questions, we did some experiments with increasing amounts of acoustically matched GSM car English data.

Table 6 summarizes these experiments for the name dialing task, at fast driving speed. (The other test sets showed similar results.) As expected, additional MAP adaptation with English data (line 3) helps considerably (34% relative error-rate improvement with 12K sentences) over the cross-language MLLR adapted models (line 2). Comparing lines 2 and 4, it appears that MLLR adaptation with English data instead of German data is slightly better with equal amounts of data (6% relative improvement). The comparison of the MLLR, MAP, and MLLR+MAP UK adaptations (lines 4, 5, and 6 respectively), agrees with [13]: the best results are obtained with the combined MLLR+MAP scenario. Somewhat unexpectedly, comparing lines 3 and 6 shows that MLLR adaptation with German data instead of English data before performing MAP adaptation to English data gives slightly superior performance (up to 6% relative). We have no specific explanation to offer for this observation. However, it does indicate a posteriori that cross-language MLLR seems perfectly adequate for channel robustness adaptation.

System	English Adapt.Set Size			
	<i>n/a</i>	3K	6K	12K
Baseline	47.0			
MLLR GE 12K	23.5			
MLLR GE 12K + MAP UK		20.4	18.4	15.4
MLLR UK		24.3	23.8	22.1
MAP UK		28.6	24.8	21.4
MLLR UK + MAP UK		21.4	19.6	15.8

Table 6. Error rates of adapted models under different adaptation scenarios, for the name dialing, fast driving test set. GE = German, UK = English.

6. CONCLUSIONS

We proposed a simple method based on MLLR to port acoustic characteristics such as channel distortion from one language to another. We showed in a case study that this technique allowed us to

build acoustic models whose performance is sufficient to deploy a server-based in-the-car name dialing application over the GSM network, without collecting any GSM or car data for that language, using instead similar data from another language.

We explored the issue of acoustic mismatch between the adaptation and test data, and observed that car and handsfree-kit mismatches had a non-negligible impact of the performance of the adapted models. Finally, we considered a situation where a pilot system is fielded and language-matched data becomes progressively available. We show that such data can effectively be used for MAP adaptation of the cross-language adapted models.

We did not explore the impact of how similar the adaptation and test languages are, but our experiments seem to indicate that the language mismatch is not a key factor in itself.

Acknowledgments: We wish to thank Ashvin Kannan and Jean-François Crespo for their help.

7. REFERENCES

- [1] Special issue on “Robust Speech Recognition”, Speech Communication, vol.25, 1998.
- [2] Proc. of the Workshop on “Robust Methods for Speech Recognition in Adverse Conditions”, 1999.
- [3] Session on “Noise Robust Recognition: Front-End and Compensation Algorithms”, in Proc. Eurospeech, 2001.
- [4] J.-C. Junqua, “Impact of the Unknown Communication Channel on Automatic Speech Recognition: A Review”, in Proc. Eurospeech, 1997.
- [5] Johns Hopkins Summer Workshop, 1999. At <http://www.clsp.jhu.edu/ws99/projects/asr/index.html>.
- [6] T. Schultz and A. Waibel, “Language Portability in Acoustic Modeling”, in Proc. workshop on Multilingual Speech Communication, 2000.
- [7] W. Byrne, P. Beyerlein, J.M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, J. Picone, D. Vergyri, and W. Wang, “Towards Language Independent Acoustic Models”, in Proc. ICASSP, 2000.
- [8] T. Schultz and A. Waibel, “Experiments on Cross-Language Acoustic Modeling”, in Proc. Eurospeech, 2001.
- [9] P. Fung, Ma C. Y., and Liu W. K., “MAP-Based Cross-Language Adaptation Augmented by Linguistic Knowledge: From English to Chinese”, in Proc. Eurospeech, 1999.
- [10] C. Leggetter and P. Woodland, “Maximum Likelihood Linear Regression For Speaker Adaptation of Continuous Density HMMs”, Computer Speech and Language, May 1995.
- [11] L. Neumeyer, A. Sankar, and V. Digalakis, “Comparison of Supervised and Unsupervised Adaptation Techniques”, in Proc. Eurospeech, 1995.
- [12] V. Digalakis, P. Monaco, and H. Murveit, “Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognition”, in IEEE Trans. on Speech and Audio Processing, pp. 281-289, 1996.
- [13] V. Digalakis and L. Neumeyer, “Speaker Adaptation Using Combined Transformation and Bayesian Methods”, in IEEE Trans. on Speech and Audio Processing, July 1996.