An approach to multilingual acoustic modeling for portable devices

Yan Ming Cheng, Chen Liu, Yuan-Jun Wei, Lynette Melnar, Changxue Ma

Human Interface Lab, Motorola Labs
Motorola, Schaumburg, IL, USA
{ywei,cliu1,ycheng,melnar,cxma}@labs.mot.com

Abstract

There is an increasing need to deploy speech recognition systems supporting multiple languages/dialects on portable devices worldwide. A common approach uses a collection of individual monolingual speech recognition systems as a solution. However, such an approach is not practical for handheld devices such as cell phones due to stringent restrictions on memory and computational resources. In this paper, we present a simple and effective method to develop multilingual acoustic models that achieve comparable performance relative to monolingual acoustic models but with only a fraction of the storage space of the combined monolingual acoustic model set.

1. Introduction

As communication and voice-enabled devices are more widely used, multilingual ASR (ML-ASR) applications have become increasingly important. To develop such an ASR system, complex acoustic modeling methods (often realized by Hidden Markov Models (HMMs)) are needed. In practice, for a server–based ASR system, which has sufficient CPU and memory resources, individual sets of monolingual acoustic models can be trained separately and stored in the system. When a user calls, the system will switch to a particular set of acoustic models according to the language information it obtains. However, such an approach requires a large amount of memory for storing all of the HMMs, which is not practical for an embedded ML-ASR system (e.g., implemented on a mobile phone or PDA), where the available space is generally insufficient for storing the HMMs of multiple languages. Hence, the number of HMMs has to be greatly reduced in order to accommodate such implementations.

Strategically, the reduction in number of HMMs is achieved by taking advantage of acoustic similarities across languages ([8], [2]). One approach, termed *knowledge-driven*, uses phonological knowledge to group together phonetic units with similar features across languages. These aggregate phonetic units are used to represent the acoustic units for training the acoustic models. Accordingly, speech data from multiple languages can be commonly represented by a single symbol provided such data are determined to be phonologically similar. An advantage of this approach is that acoustic models created for a particular language can apply to other languages that include the same or similar phonetic units in their local inventories. The ability to share models across languages crucially allows for the creation of acoustic models for those languages for which little or no acoustic data is available. However, this approach does have some disadvantages. First, the determination of acoustic units is made independently of acoustic data. Second, the quality of the acoustic units (or models) will likely be degraded as unit sharing across languages increases.

A second approach, termed *data-driven*, determines cross-language acoustic units based on sufficiently large databases. For instance, previous work [3] uses HMM distance to determine whether two HMMs from two monolingual systems can be merged. The resulting unified acoustic models support multilingual phonetic units. Contrary to the *knowledge-driven* method, the cross-language acoustic model can only be applied to those phonetic units that have sufficient instances in the database. However, it is incapable of producing acoustic models based on knowledge in the absence of sufficient data.

In this paper, we present a novel approach to modeling that combines the two approaches discussed above. This method maximizes the advantages of using expert knowledge of cross-language phonetic similarity to produce acoustic models with little or no data. A data-driven method is used to determine the acoustic units based on acoustic similarities to precisely model a given database. This combined approach consequently creates a set of multilingual acoustic models that is only a fraction of the size of the collective monolingual acoustic model set.

In the next section, we are going to describe in detail the two steps of creating the multilingual acoustic models. Then, in Section 3, we present our database preparation and experiment conditions. In Section 4, we provide some of our experiment results and comparisons. Finally, we will make some concluding remarks based on this study.

2. The two steps of building multilingual acoustic models

2.1. Step One: The Motorola Polyphone Network: a feature-based merge specification network of phonetic symbols

The IPA (*International Phonetic Alphabet*) is a well-known and useful tool in exploring phonetic similarity across languages. However, while the IPA differentiates phonetic units with remarkable detail and consistency, it is not in its purview to explicitly organize distinct symbols into units larger than their base signs. The IPA, therefore, does not expressly define possible phonetic symbol mergers across languages. Nevertheless, the insights inherent in the formulation of the IPA can be used to create a network of phonetic similarities that *can* be used to organize phonetic symbols into larger ‘super-phonetic’ categories. These categories, in turn, can be shared across languages.

The Motorola Polyphone Network is such a system of phonetic symbols. Its construction, described in detail in [1], begins with a large and representative subinventory of symbols that derive from the IPA. These symbols are leaves or terminal points in the network and are loosely organized into large sets based on broad phonetic similarities. They merge to create increasingly more abstract ‘super-phonetic’ categories (nodes in the network). Actual merge at each node is governed by progressively more abstract phonetic, phonological, and frequency constraints and may continue *ad absurdum*.

From the Motorola Polyphone Network, a multilingual symbol inventory, termed *MotPoly*, may be defined. The MotPoly inventory is a subset of the network’s phonetic symbols (either nodes or leaves) that are selected based on a number of criteria. For instance, any given MotPoly symbol should correspond to phonemes in a maximum number of languages. This ensures that the inventory will have the greatest possible language coverage. In addition to this frequency criterion, the MotPoly inventory should also observe those contrasts in languages that are typically phonemic. Thus, any given MotPoly symbol should not encompass phonetic distinctions that are known to contrast phonologically. Finally, though not exhaustively, the MotPoly inventory typically must conform to a target size - which governs the relative extent to which the previously mentioned criteria will be violated. Of course, in practice, the constituents of this inventory can and should be determined experimentally.

2.2. Step Two: Building a decision tree using MotPoly symbols and language and phonetic information

In Step One, we select the MotPoly symbol inventory, where each symbol corresponds to one or more distinct IPA symbols (as defined by the Motorola Polyphone Network). While this inventory is designed to maximally represent the sounds of most languages, it does create some problems from an acoustic point of view. For example, any particular MotPoly symbol can represent an over-grouping (or over-merging) of IPA symbols for some languages or dialects and for particular phonetic contexts. To accommodate this possibility and achieve high–performance acoustic models, a decision tree is employed that associates MotPoly symbols with potentially several acoustic models. In traditional monolingual acoustic modeling, this type of division is solely based on phonetic context, such as triphones and contextual allophones. In our approach, we base the division on an optimal mixture of language and phonetic information.

Much like the generic decision tree construction algorithm (CART) described in [9], our decision tree construction algorithm uses a Euclidean distance as the purity measure. Each instance of a MotPoly symbol in our database is represented by a vector of 117 dimensions. This vector concatenates the mean vectors of a 3-state HMM, which is trained with the single instance of the MotPoly symbol. The question set of the decision tree algorithm contains two types of questions, one pertaining to phonetic information and the other to language information. The latter question type includes questions regarding language identity, language class, dialect identity, dialect class, sub-language, etc. The sub-language questions refer to particular tasks that are well represented in the database. The database used for the decision tree algorithm is a sequence of the vectors mentioned above. Each vector is tagged with both phonetic and language information. A simple CART algorithm is used to construct the decision tree with the purity measure, the question set, and the tagged database.

2.3. The search engine for multilingual acoustic modeling

In this study, we focus on using multilingual acoustic models in a single language speech recognition task. In other words, we assume that the language information is known prior to recognition. We use MLite++ [10], a computation and memory efficient search engine to perform speech recognition on embedded devices. Given the language information, this engine is able to select a corresponding subset of models (many of which are shared across languages) from the multilingual model set to perform speech recognition.

3. The database for building the decision tree and for training/testing the HMMs

The database selected for the experiments is a subset of the database produced by the LDC multilingual CallHome and Polyphone projects (see [6]). This subset comprises six diverse languages, including: American English (EN), Spanish (both European and Latin American) (ES), German (DE), Mandarin Chinese (ZH), Japanese (JA), and Arabic (AR). The speech in this database has limited speaker coverage, consisting of uncontrolled telephone conversations between two speakers, the caller and callee. For training and testing purposes, those files that contain foreign speech and poor quality transcriptions are filtered out of the database. All speech files are down sampled to 8k Hz from their original high sampling rate. Table 1 provides data distribution and partitioning and approximate vocabulary size of the test sets for each of the targeted languages.

Table 1. Database, test vocabulary size and number of phonemes

	EN	AR	DE	JA	ZH	ES
Train	166k	10k	8k	13k	52k	66k
Test	4234	2363	1041	2864	16201	4400
Vocab size	~2k	~3k	~1k	~1k	~5k	~3k
Phone	36	35	39	29	36	32

The front-end is a standard mel-cepstral coefficient (MFCC) without pre-emphasis. The feature vector contains 12 MFCC, log energy and their first and second derivatives. Cepstral mean removal is also used. The acoustic models used in the experiments are conventional 3-state HMMs with left-to-right topology, with each state containing 10 Gaussian mixtures.

For each language in consideration, the test portion of CallHome database is used as the main test set. We use a finite state grammar that generates testing utterances.

4. Experimental results of multilingual acoustic models in comparison with the monolingual models

4.1. Baseline system 1 (B1) of monolingual acoustic triphone models

For the baseline system, the traditional triphone-based acoustic modeling technique is used. The triphone inventory for each language is determined based on the available training data of that language so that each triphone unit includes at least 15 samples in the training data. Each triphone set is language dependent. Table 2 shows the number of triphones per language. In total, there are more than 12K triphones, implying a huge demand for storage space with the traditional monolingual approach.

Table 2. Number of triphones per language and speech recognition performance (word accuracy) with B1

%	EN	AR	DE	JA	ZH	ES	Total / Average
Triphones	4210	1831	1019	775	1866	2309	12010
Accuracy	68.2	63.7	53.8	36.2	67.1	66.7	65.0

The average is weighted proportional to the size of the test set. As expected, the languages with sufficient data for training, EN, ZH and ES, have better performance. With the exception of AR, the languages with insufficient data, DE and JA, are associated with low performance. It is possible that AR's slightly higher performance is due to its relatively large triphone set.

4.2. Baseline system 2 (B2) of monolingual acoustic allophone models made with decision trees

To create a baseline system that is especially suitable for comparison with our multilingual system, we generate monolingual allophones by utilizing the same decision tree technology as is used in our multilingual acoustic modeling. Unlike B1, however, for this test we employ a much smaller number of decision tree allophones per language as acoustic model units. The same decision tree construction algorithm as described in Section 2.2 is used except that only the phonetic contextual information is included in the decision tree question set. Table 3 shows the number of allophones and word accuracy performance for each language.

Table 3. Number of contextual allophones per language and speech recognition performance (word accuracy) with B2

%	EN	AR	DE	JA	ZH	ES	Total / Average
Allophones	625	520	577	434	590	555	3301
Accuracy	56.2	67.2	56.	44.4	68.9	67.1	64.1

Note that the performance of B2 is comparable to that of B1. However, B2's storage requirements are only one quarter of that of B1.

4.3. Multilingual triphone-based acoustic modeling (MLT)

As a reference, we also produce a simple multilingual acoustic modeling system by slightly modifying B1. In this system, we model the identical triphones across languages with a single acoustic model. In this way, we are able to compress the number of triphones from about 12K to 9469. Table 4 shows the number of triphones and performance in word accuracy.

Table 4. Number of total triphones and speech recognition performance (word accuracy) with MLT

%	EN	AR	DE	JA	ZH	ES	Total / Average
Triphones	9469						9469
Accuracy	67.9	44.7	45.7	21.5	67.4	65.1	62.0

4.4. Multilingual decision-tree based acoustic modeling (MLD)

For this final experiment, we use the multilingual acoustic modeling strategy described in Section 2. The 782 total allophones of this system are generated via our decision tree and the optimal mixture of language and phonetic information, derived from the construction algorithm. Table 5 shows the number of allophones and recognition word accuracy for each language. Note that all the languages have comparable performance except JA and ES. As with the previous experiments, recognition accuracy for JA is very low. A possible explanation for JA's poor performance is both lack of training data and phonetic incongruity with the other languages in this study. However, for ES, the multilingual decision-tree based acoustic models significantly outperform the other languages. Furthermore, the ES result from this experiment surpasses the ES results from the previous experiments. Perhaps because the sound system of ES is very common from a universal perspective, it is highly compatible with other languages, including those used in this study. Thus, additional, non-ES data can be effectively used by ES via the decision tree to boost both its modeling robustness and ability to deal with the unseen data.

Table 5. Number of total allophones and speech recognition performance (word accuracy) with MLD

%	EN	AR	DE	JA	ZH	ES	Total / Average
Allophones	782						782
Accuracy	64.8	61.4	59.1	26.2	67.3	78.0	66.4

For comparison purposes, we summarize the four acoustic modeling systems of our experiments in Table 6. Here, we provide weighted average performances (where the weights are proportional to the test data size), and the number of acoustic units (either triphones or allophones), which directly determine the storage size.

The triphone-based monolingual acoustic modeling system (B1) and the multilingual decision-tree acoustic modeling system (MLD) offer the best performances. However, the storage size of the MLD is only a small fraction (6.5%) of the size of B1. The traditional decision-tree monolingual acoustic modeling system (B2) and MLD employ similar data-driven algorithms. However, B2 uses language information (language identity) as an independent attribute to generate separate acoustic units, whereas MLD uses an optimal mixture of language and phonetic information to derive acoustic models via the decision tree. MLD outperforms B2 while MLD's storage is only 23% that of B2. The multilingual triphone-based acoustic modeling system (MLT) yields the lowest performance and it also suffers from a very large storage requirement. It's likely that MLT's poor performance is related to the MotPoly inventory. As mentioned, MotPoly symbols may very well have over-grouped (or over-merged) the IPA symbols for the given database. This is the cost of both maximizing knowledge-based prediction across languages and restricting the MotPoly inventory to a target size. The result is that the triphones based on the MotPoly symbols may be acoustically quite distinct across languages and the strategy of using only one acoustic model per triphone may have led to the observed low performance.

Table 6. Summary of the four systems' performances and numbers of acoustic units

%	B1	B2	MLT	MLD
Number of units	12010	3301	9469	782
Word accuracy	65.0	64.1	62.0	66.4

5. Concluding remarks

In this paper, we have shown a two-step methodology for multilingual acoustic modeling, which combines the advantages of knowledge-driven and data-driven multilingual acoustic modeling approaches. This methodology can also be extended in a straightforward fashion to multi-dialect and sub-language acoustic modeling. In Step One of the procedure, we use a multilingual MotPoly inventory, which is experimentally derived from the Motorola Polyphone Network, to maximize knowledge-based, cross-language phonological predictability. In the second step, we derive a set of acoustic models supporting a MotPoly symbol, via a multilingual decision-tree construction algorithm for the given database. Contrary to traditional monolingual decision-tree algorithms, the multilingual decision tree utilizes an optimal mixture of language information (language, dialect, sub-language, etc.) and phonetic information, obtained from its construction algorithm, to derive acoustic models for a MotPoly symbol. For comparative purposes, we also provide several reference approaches, the B1 and B2 systems, and an alternative approach, MLT.

Our experiment results demonstrate that the performance of the presented two-step methodology for multilingual acoustic modeling outperforms all of the other approaches, while only occupying a very small fraction of the storage space. The advantage of small storage is crucial for deploying speech recognition systems in resource sensitive computing environments, such as portable devices and wireless communicators.

As expected, our methodology boosts overall performance of some languages compared to traditional monolingual modeling. This is an important factor in developing multilingual speech recognition systems. Not atypically, databases of sufficient size are not readily available for every ML-ASR targeted language, and at times, no database may be available. In these cases, our methodology is able to maximally explore the language resources available from other languages to produce a starter system.

Acknowledgment: The authors would like to express their appreciation to Dusan Macho for his thoughtful comments during this work.

6. References

[1] Melnar, Lynette, Talley Jim, "Phone Merger Specification for Multilingual ASR: The Motorola Polyphone Network", to appear in the Proc. of ICPhS, 2003.

[2] International Phonetic Association, "Handbook of the International Phonetic Association", Cambridge University Press, 1999

[3] B. Mak, E. Barnard, "Phone clustering using the Bhattacharya distance", Proc. ICSLP 96, Vol. 4, pp. 2005-2008, Philadelphia.

[4] T. Schultz, A. Waibel, "Experiments on Cross-language Acoustic Modeling", EuroSpeech 2001.

[5] T. Schultz, A. Waibel, "Polyphone decision tree specialization for language adaptation", Istanbul, ICASSP 2000.

[6] CallHome database, http://www.ldc.upenn.edu.

[7] Waibel, A., et al., "Multilinguality in speech and spoken language systems," Proc. IEEE, v. 88, pp. 1297-1313, 2000.

[8] Cohen, P., et al., "Towards a universal speech recognizer for multiple languages," Proc. ASRU, pp.591-598, 1997.

[9] Brieman, Olshen, Friedman, and Stone, "Classification and Regression Trees", CRC Press, 1984.

[10] Yan Ming Cheng, "MLite++ book", Motorola Labs technical report, 1999.