

# An Investigation of Acoustic Models for Multilingual Code-Switching

Christopher M. White<sup>1,2</sup>, Sanjeev Khudanpur<sup>2</sup>, James K. Baker<sup>1</sup>

<sup>1</sup>Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA

<sup>2</sup>Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

cmileswhite@jhu.edu, khudanpur@jhu.edu, james.karl.baker@gmail.com

## Abstract

Multilingual speech processing continues to develop as speech technology spreads to heterogeneous clients and applications. We address a distinct problem of *code-switching* — the spontaneous but occasional use, within speech in one language (referred to as  $L_1$ ), of words, phrases, expressions or idioms from a second language ( $L_2$ ). We examine two alternatives for modeling the acoustics of such words: creation of  $L_1$  pronunciations for the *out-of-language* (OOL) words for use with  $L_1$  acoustic models, and retention of their  $L_2$  pronunciations for use with multilingual acoustic models. We test the hypothesis that the latter is a better acoustic model for OOL words. We develop a set of lexica in IPA form, a global phoneme inventory, and handle the problem of  $L_2$  word pronunciation by creating linguistically motivated pairwise mappings. We show that retention of  $L_2$  pronunciations with multilingual acoustic models better explains the observations when restricted to a forced alignment.

**Index Terms:** Cross-lingual and multi-lingual processing, Automatic speech recognition, Accent and language identification, Spoken language resources and annotation

## 1. Introduction

Developments in multilingual automatic speech recognition (ASR) range from building multiple, monolingual ASR systems that merely share design principles but little else, through systems with language-adapted models with a shared multilingual ancestry, to truly language-universal acoustic modeling based on the IPA [1]. A recent survey of relevant work may be found in [2]. In particular, the multilingual acoustic modeling work falls into three broad categories: (i) building acoustic models in one language using speech from other languages because adequate amounts of transcribed speech are unavailable for the language(s) of interest, (ii) building a single ASR system to recognize multiple languages because the language of the intended user is a priori unknown, or (iii) building a recognizer for speakers with a nonnative accent. In most of the work, the ASR systems assume at least one language per utterance, usually one language per speaker.

In this paper, we focus on the related but distinct problem of *code-switching* — the spontaneous but occasional use, within speech in one language (referred to as  $L_1$ ), of words, phrases, expressions or idioms from a second language ( $L_2$ ). E.g., the English word *email* is used in over 12 out of 100 telephone conversations in the Mandarin CallHome corpus. Most ASR systems do not have a mechanism to model code-switching; the resulting errors are ascribed to *out-of-language* (OOL) words and usually ignored because they do not substantially impact *average* WER. Yet, infrequent OOL words are also often information-rich, and recognizing them is particularly important in applications such as spoken term detection

[12], where the OOL word could be present in the query, and spoken dialog systems, where the OOL could be a travel destination. Note that in each case, the ASR system is not required to *transcribe* an OOL word using the  $L_1$ -orthography; it suffices to represent it machine-internally in the  $L_2$ -orthography and the main challenge is to recognize it when it is spoken.

ASR for accented speech, to which this problem is most similar, has been addressed by using techniques such as speaker adaptation, pronunciation adaptation, structural adaptation of the acoustic models and language model adaptation to deal with phenomena such as nonnative articulation, speaking rate, pause distribution, proficiency-related disfluency, word choice, syntax, discourse style etc [2, 3, 4, 5, 6]. Many of these phenomena are not applicable to code-switching in conversational speech: a Mandarin ( $L_1$ ) speaker using an English ( $L_2$ ) word does not try nearly as hard to render its pristine English pronunciation to her Mandarin listener. Yet, OOL words are pronounced clearly enough to be correctly recognized by the listener. We therefore examine two alternatives for modeling the acoustics of such words: creation of  $L_1$  pronunciations for the OOL words for use with  $L_1$  acoustic models, and retention of their  $L_2$  pronunciations for use with multilingual acoustic models. We test the hypothesis that the latter is a better acoustic model for OOL words.

## 2. Multilingual Speech Resources

We chose the CallHome database covering Egyptian Colloquial Arabic, English, German, Japanese, Mandarin Chinese, and Spanish available from the Linguistic Data Consortium [7].

### 2.1. Pronunciation Lexicons - CallHome

The CallHome database contains pronunciation lexica for each of the 6 languages listed; the size of the phoneme set for each language ranges from 30 to 49 phonemes. While these lexica facilitate building 6 monolingual systems, they were *not* created with a common phoneme set with similar (or identical) phonemes that occur in more than one language represented using shared symbols.

### 2.2. Pronunciation Lexicons - CallHomeJHU

We contend along with [3, 9, 10, 6] and others that articulatory representations of phonemes across languages are similar enough to be modeled by the same phoneme in the pronunciation lexica. This creates the need for a global phoneme set such as GlobalPhone [3]. Leveraging the CallHome lexica, part of the Johns Hopkins Summer Workshop 2007 team "Recovery from Model Inconsistency in Multilingual Speech Recognition" mapped the phoneme inventories for all 6 CallHome languages into IPA [1] classes and rewrote the lexicons accordingly.

The lexica facilitated building and development of multilingual acoustic models, which we use here to address code-switching.

For our lexica the global phoneme set can be defined by

$$\Gamma = \cup_{i=1}^{N=6} \Gamma_{L_i} = \Gamma_{LI} + \cup_{i=1}^{N=6} \Gamma_{LDL_i} \quad (1)$$

where  $\Gamma_{L_i}$  is the phoneme inventory (in IPA form) for language  $i$ ,  $\Gamma_{LI}$  is the set of phonemes which are language independent (occur in more than one language), and  $\Gamma_{LD}$  is the set of phonemes that are language dependent (occur in only one language). In Table 1 the  $\Gamma_{LD}$  for each CallHome language are shown in the bottom half, while the  $\Gamma_{LI}$  are described in the upper half based on how many languages share the phoneme.

Table 1: *Global Unit Set for 6 Languages*

Shared by	#	
	58	Phonemes Shared by $\geq 2$ Languages
All	7	nonspch, unk, sil, k, w, m, j
5	11	h, l, p, f, v, z, b, g, n, a, u
4	6	s, ʃ, o, ai, au, i
3	11	d, t, r, x, f r <sup>j</sup> , tʃ, e, i, e
2	23	ts, ə, θ, u, ei, a, ʒ, ou, d <sup>j</sup> æ, o, s, s, n ɹ, y, ɲ, ʒ, ɹ, d, a, ʔ, ʌ
	51	Phonemes in 1 Language
AR	10	ɸ, d <sup>v</sup> , s <sup>v</sup> , s <sup>j</sup> , h d <sup>v</sup> , e, t <sup>v</sup> , ʃ, a:
EN	4	ai, ð, ʊ, ɔ:
GE	10	ū, ø, ä, e: ōe, oi, œ, ɔ, y, pf
JA	11	b <sup>j</sup> , k <sup>j</sup> , p <sup>j</sup> , φ, φ ʃ <sup>j</sup> , h <sup>j</sup> , m <sup>j</sup> , g <sup>j</sup> , t, u
MA	13	t <sup>h</sup> , tʃ <sup>h</sup> , ç, ç̃, p <sup>h</sup> n, r, tʃc <sup>h</sup> , ç̃, i, tʃ <sup>h</sup> , k <sup>h</sup>
SP	3	ɣ, β, r

Another method to analyze multilingual lexica and global phoneme inventory developed by [3] is the notion of a share factor  $sf_N$ . For a set of  $N$  languages the  $sf_N$  is defined in Equation 2 as the ratio between the sum of the number of phonemes in a set of languages and the size of the global set derived from those languages. It describes the number of languages that share a phone, and therefore is dependent on which languages are chosen to be in the numerator.  $sf_N$  ranges between 1 and  $N$ : 1 if no phoneme is shared across languages,  $N$  if each language uses the same phone set. We calculate the average, min, and max  $sf_N$  over all possible  $k$ -tuples ( $k = 1, \dots, 6$ ) of 6 languages for the  $\binom{12}{k}$  combinations.

$$sf_N = \frac{\sum_{i=1}^N |\Gamma_{L_i}|}{|\Gamma|}; |\Gamma| = |\Gamma_{LI}| + \sum_{i=1}^N |\Gamma_{LDL_i}| \quad (2)$$

The results can be seen in the upper most line in Figure 1 where the min, max, and average for the case using all 6 languages is about 2.3, which compares well to the GlobalPhone average share factor for 6 languages of approximately 2.1 [3]. Moreover, the results are consistent in that the share factor increases with the number of languages. However, our share factor plot does not exhibit the large variance reported in [3] indicating less dependence on which languages are involved.

While the share factor describes phoneme coverage on a type basis, not all phonemes have the same distribution of occurrence. For example, although Japanese and Mandarin have about the same number language dependent phonemes given all 5 other languages, those phonemes in Japanese occur infrequently. To illustrate this, we calculate a token weighted coverage coefficient,  $wcc$ , over a given training partition, which shows how much of the training partition is covered by phonemes in other languages:

$$wcc(L_i) = 1 - \sum_{p \in \Gamma_{LDL_i}} count(p) / \sum_{p \in \Gamma_{L_i}} count(p) \quad (3)$$

where  $p$  is a phoneme. This is a function of a language ( $L_i$ ) and the terms in the sum in the numerator depend the size of  $\Gamma_{LD}$ , which is determined by the other languages in comparison. For example, we could consider the coverage of Mandarin given all 5 other languages, or given 2 specific other languages. Each case creates a different set of language dependent phonemes. Therefore, we calculate  $wcc$  for each language given all subsets of other languages (as in calculating share factor), which generates an average, min, and max shown in Figure 1. Returning to Japanese and Mandarin, the Japanese  $wcc$  curve remains above the Mandarin curve illustrating the example above that language dependent phonemes in Japanese are infrequent and do not impact coverage. To interpret a point in the graph, given 3 other languages on average the  $wcc$  for Mandarin is approximately 0.75, which indicates that with 3 other languages about 75% of phoneme tokens in the Mandarin training data come from phonemes that have representation in another language.

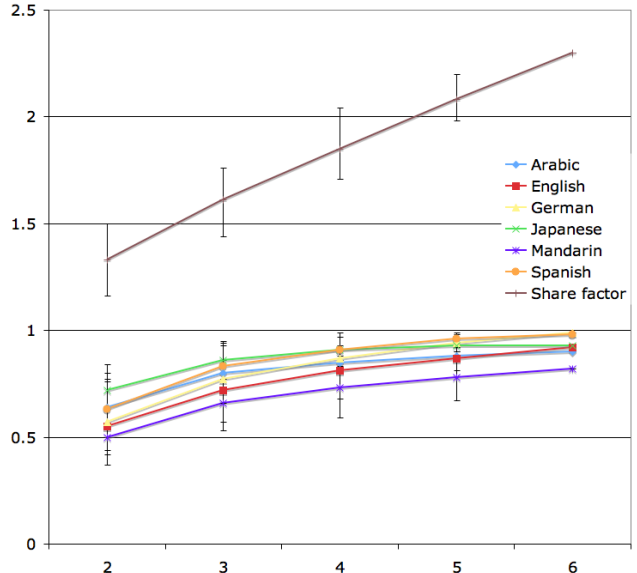


Figure 1: Average Share-factor;  $wcc$  for each Language: average value, min, max as a function of number of languages included.

### 2.3. Code-Switching in CallHome

The CallHome database contains examples of conversational code-switching. In Table 2, the languages in CallHome are listed next to the number of utterances in the train and development partition (about 100 conversations) that contain one or more words of English. A further categorization of OOL words as proper nouns (e.g. “Google” in Mandarin), technical terms

(e.g. “e-mail”) and common expressions (e.g. “hi!”) and back channels will be undertaken to guide future research.

Table 2: *CallHome (tr+dev) with  $L_2 = English$*

$L_1$	# $L_2$	$L_1$	# $L_2$	$L_1$	# $L_2$
Arabic	1084	German	762	Spanish	330
Japanese	111	Mandarin	871		

### 3. Pronunciations of OOL Words

Most of the work in pronunciation modeling (accented, multilingual, or traditional) has assumed a single language for the target utterance [2]. In an ASR system for accented  $L_2$  speech (say, English) by a native speaker of  $L_1$  (say, Spanish), it has been demonstrated [2] that rather than *modifying* the English pronunciation lexicon to capture the Spanish accent, greater ASR accuracy is obtained by preserving the English lexicon and *adapting* the English *acoustic models* to the Spanish-accented speech or speakers. This may initially suggest that if, while speaking  $L_1$ , a speaker code-switches to an  $L_2$  word, the ASR system for  $L_1$  should use the  $L_2$  pronunciation of the OOL word along with accent-adapted  $L_2$  acoustic models. Not only does this solution require a transcribed corpus of  $L_1$ -accented  $L_2$ -speech — a nontrivial resource — it also requires each  $L_1$  ASR system to carry all conceivable  $L_2$  acoustic models.

We investigate two alternate solutions that require neither accented-speech corpora nor carrying more than one additional set of acoustic models. We investigate creating an  $L_1$  pronunciation for each  $L_2$  word and using the  $L_1$  acoustic models available in the ASR system. We contrast this with retaining the  $L_2$  pronunciation of the OOL word, but using language-universal acoustic models instead of the accent-adapted acoustic models. This allows us to compare the effect of using acoustic models with parameters estimated from observing one or many languages. To facilitate this investigation, we create a global phone inventory and mappings between the various  $L_1$  and  $L_2 = English$ .

Consider the English word ‘Rockefeller’ and its  $L_1$  pronuncia-

Table 3: ‘Rockefeller’ in English and Arabic

English	ɹ ɑ k ə f ɛ l ɜ
Arabic	r <sup>j</sup> ɑ k ũ f a l a

tion shown in the upper entry in Table 3. The canonical pronunciation includes phonemes that are not covered in the inventory of other languages. For example, when comparing the inventories of Arabic and English the phonemes ɑ,k,f,ɹ have models with parameters estimated from Arabic speech data, however the phonemes ɹ, ə, ε, ɜ do not have a corresponding representation in the Arabic inventory. Therefore to score ‘Rockefeller’ with acoustic models estimated from Arabic speech it is necessary to find a mapping between the phonemes in English and the ‘closest’ phonemes in Arabic. When considering  $L_2 = English$  there are 24 such phonemes compared to  $L_1 = Mandarin$  and 21 compared to  $L_1 = Arabic$ .

Much work [8, 10] addressed this problem falling into three categories: mapping by hand, linguistic-feature based automatic mapping, and data-driven mapping. Our work falls into the second category.

Efforts in linguistically motivated mapping have included a simple scoring based on a few features and calculating a ham-

ming distance between vectors of features [10, 8]. We continue along this line, but introduce a set of features that take binary values as well as other values. The set of features and their values can be seen in Table 4. For each phone in  $L_2$  not covered in  $L_1$  a feature vector is created and the distance calculated per feature dimension as identity or real-valued where appropriate and then summed.

A resultant example can be seen in Table 3 where the phone ɹ (alveolar approximant) has been mapped to r<sup>j</sup> (palatalized alveolar tap or flap), ə (Schwa) has been mapped to ũ (tied open back unrounded with near-close near front unrounded vowel), and ε,ɜ (open-mid front unrounded vowel, r colored open-mid front unrounded vowel) with a (open front unrounded vowel).

Table 4: *Features for Linguistic Distance*

Values	Feature
binary	vowel, nasal, length, rounding, rhotic, palatalized plosive, fricative, click, approximant, lateral, flap, trill apical, velarized, radical, glottal, alt. air supply
ternary	pulmonic, aspiration
>ternary	height, backness, labial, coronal, dorsal

### 4. Acoustic Models and Code-Switching

Acoustic modeling for multilingual and accented speech continues to show that adaptation [6, 3, 2] and pooling of data [2, 4, 5] improve recognition performance. Although there is recent work using adaptation and context-dependent phoneme modeling for multilingual speech recognition [9], most work until now has shown that context-independent phoneme modeling better explains the data in a language independent setting while avoiding sparsity issues related to context-dependent modeling [2]. Therefore this work builds two types of context-independent acoustic models: one type estimated having observed  $L_1$  using the pronunciation lexicon from  $L_1$ , and one type estimated having observed all of CallHome  $L_{UPR}$  using all the languages’ lexica.

#### 4.1. Experimental Setup

Each system contains 3 State CD-HMM/phoneme with 16 Gaussian/state modeling context-independent phonemes. This entails using the training partition from the CallHome data base with approximately 13 hours of training per language. We consider code-switching in two cases:  $L_1 = Arabic$  or  $L_1 = Mandarin$ . Monolingual systems are built for Mandarin and Arabic as well as a set of universal phoneme models built from all 6 CallHome languages. In both cases  $L_2 = English$ . Table 5 outlines the frequency of  $L_1$  and  $L_2$  for the experiments below. In both cases there are over 1700 English terms embedded in  $L_1$ . Since we treat this as a detection task (detecting English words), the number of words in  $L_1$  becomes relevant as they can be erroneously detected (false alarms).

Table 5: *Code-Switching Word Breakdown with  $L_2 = English$*

$L_1$	# Utts	$L_1$ Words	$L_2$ Words
Arabic	1084	9455	2231
Mandarin	871	8486	1701

## 4.2. Detecting $L_2$ in $L_1$

To investigate which pronunciation- plus acoustic-model combination is better for OOL words, we set up a detection task as follows. From the orthographic transcript of each  $L_1$  utterance containing an  $L_2$  word, we create two alternate phonetic transcripts: one based the IPA-lexicon pronunciation of words, resulting in a mixed transcript with  $L_1$  and  $L_2$  phones, and another based on the  $L_1$  pronunciation of each OOL word (cf Section 3), resulting in a transcript that contains only  $L_1$  phones. The  $L_1$ -only transcript is *Viterbi*-aligned with the acoustics using  $L_1$  acoustic models, while the  $L_1+L_2$  transcript is aligned using a set of *universal phone models* trained by pooling together speech from all six CallHome languages. Following related work in OOV detection [11], the total log-likelihood of the acoustic-frames that align with each word are collected from the two alignments, and the likelihood ratio is used to decide if the word is OOL. I.e., we test whether the universal phone models, along with their native  $L_2$  pronunciation, result in a higher likelihood for the OOL words, while the converse is true for the  $L_1$  words.

Equations 4-5 state these likelihood computations more formally, with the likelihood ratio computation shown in the caption of Figure 2 that forms the sufficient statistic for the OOL detector:

$$p(\bar{X}|P_{1+f(P_2)}, M_1) \equiv p(\bar{X}|L_1) \quad (4)$$

$$p(\bar{X}|P_{1,\dots,N}, M_{1,\dots,N}) \equiv p(\bar{X}|L_{UPR}) \quad (5)$$

$\bar{X}$  denotes the observed acoustics,  $M$  the acoustic model, and  $P$  the pronunciation. The numerical subscript refers to the language. The function  $f(P_2)$  generates pronunciations in  $L_1$  based on the mapping that comes from Section 3. In the case 1 ( $L_1$ ) refers to either Arabic or Mandarin (depending on the curve) and 2 refers to English. The remaining indices refer to the remaining CallHome languages.

## 4.3. Experimental Results

In Figure 2, we can see for both Arabic and Mandarin that the universal phoneme models contribute to the detection of English words. The curves would be much flatter (making many type I and type II errors), or would incorrectly classify everything as either  $L_1$  or English if this were not the case. Furthermore, both languages demonstrate the effectiveness of using multilingual acoustic models indicating something to be leveraged across languages, but also demonstrate a different curve shape indicating potentially less language neutrality than might be desirable.

## 5. Conclusions

In order to test our hypothesis that acoustic models with parameters estimated from observing multilingual speech better explain foreign terms in a code-switching scenario we have developed a set of lexica in IPA form and a global phoneme inventory. We have handled the problem of foreign word pronunciation using language dependent phonemes by creating linguistically motivated pairwise mappings for each language involved in code-switching ( $L_1 = [Arabic, Mandarin]$  and  $L_2 = English$ ). These lexica and pronunciations were used to build context-independent phoneme models for both  $L_1$  and  $L_{UPR}$  (language independent systems). Comparing the likelihood of the data with code-switching under  $L_1$  and  $L_{UPR}$  by

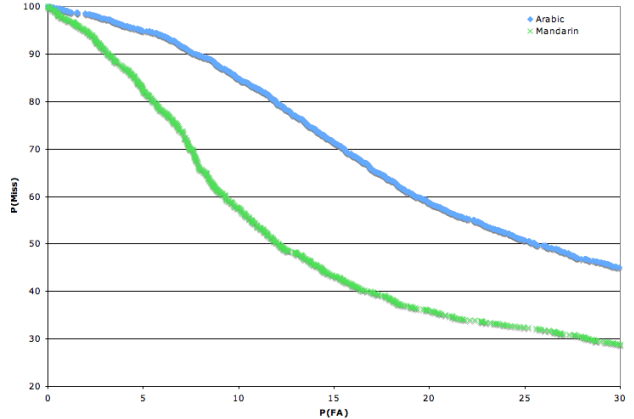


Figure 2: Detection of  $L_2$  using  $\log \left( \frac{p(\bar{X}|L_{UPR})}{p(\bar{X}|L_1)} \right)$ .

detecting English words showed that using canonical pronunciations along with universal acoustic models better explained the observations when restricted to the forced alignment (reference alignment). This result justifies future work to deal with foreign words by incorporating resources in the target language.

## 6. Acknowledgements

We would like to thank Sally Isaacoff for her work on developing the lexica, Jon Nedel for his efforts on linguistic distance, and the rest of the Johns Hopkins University Summer Workshop 2007 Team for their thoughtful contributions to this work.

## 7. References

- [1] IPA, “The International Phonetic Association (revised to 1993) - IPA Chart”, Journal of International Phonetic Association 23, 1993.
- [2] Schultz, T. and Kirchoff, K. (Ed.), “Multilingual Speech Processing”, Elsevier, Academic Press, ISBN 13: 978-0-12-088501-5. April 2006.
- [3] Schultz, T. and Waibel, A., “Language-Independent and Language-Adaptive Acoustic Modeling for Speech Recognition”, Speech Communication, vol. 35, p31-51, 2001.
- [4] Livescu, K., “Analysis and modeling of non-native speech for automatic speech recognition”, Ph.D. Thesis, MIT. 1999.
- [5] Mayfield Tomokiyo, L., “Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in Speech Recognition”, Ph.D. Thesis, Carnegie Mellon University. 2001.
- [6] Goronzy, Silke “Robust Adaptation to Non-Native Accents in Automatic Speech Recognition”, Lecture Notes on Artificial Intelligence. Vol 2560. Springer Verlag. 2002.
- [7] Linguistic data consortium, <http://www ldc.upenn.edu/>.
- [8] Chase, L., “Error-Responsive Feedback Mechanisms for Speech Recognition”, Ph.D. Thesis, Carnegie Mellon University. 1997.
- [9] Bac Viet, L. Besacier, L., and Schultz, T. “Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability”, ICASSP, 2006.
- [10] Withgott and Chen, “Computational Models of American Speech”, 1995.
- [11] Burget et al., “Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs”, ICASSP, 2008.
- [12] NIST, The Spoken Term Detection (STD) 2006 evaluation plan, [http://www.nist.gov/speech/tests/std/docs/std06\\_evalplan-v10.pdf](http://www.nist.gov/speech/tests/std/docs/std06_evalplan-v10.pdf), 2006.