# RECOGNIZING ENGLISH QUERIES IN MANDARIN VOICE SEARCH

*Hung-An Chang*$^{*\dagger}$, *Yun-Hsuan Sung*$^{*}$, *Brian Strope*$^{*}$, *Françoise Beaufays*$^{*}$

$^{*}$ Google, $^{\dagger}$ MIT, CSAIL

## ABSTRACT

Recent improvements in speech recognition technology, along with increased computing power and bigger datasets, have considerably improved the state of the art in the field, making it possible for commercial apps such as Google Voice Search to serve users in their everyday mobile search needs. Deploying such systems in various countries has shown us the extent to which multilingualism is present in some cultures, and the need for better solutions to handle it in our speech recognition systems.

In this paper, we describe a few early data sharing and model combination experiments we did to improve the recognition of English queries made to Mandarin Voice Search, in Taiwan. We obtained a 12% relative sentence accuracy improvement over a baseline system already including some support for English queries.

***Index Terms*—** Multilingual speech recognition, acoustic modeling.

## 1. INTRODUCTION

By some estimates[1], more than half of the world's population is multilingual. In cultures where multilingualism is strong, for example among immigrant groups, or in countries where multiple languages coexist, it is fairly common to hear people borrowing words from one language to use them in another. Even people whose familiarity with a second language is more limited may still choose to systematically use that language for targeted words related to professional activities, consumer goods, or entertainment, depending on the context of their exposure to that language.

The same language patterns are naturally observed in users' interactions with speech-enabled machines. For example, we estimated that 10% of the spoken Voice Search queries we get from Mainland China contain English words, whereas in Taiwan and Hong-Kong these percentages raise to 15% and 30% respectively.

This practice raises a new challenge for speech recognition systems: enabling typically monolingual systems to correctly recognize words from other languages. The problem is not trivial for many reasons: words borrowed from another language may be pronounced with a spectrum of accents, and may be systematically or randomly mispronounced. Moreover, such words would typically need to be expressed with another phone set, and since their occurences in the training data are relatively limited, it is not clear that the underlying models will be trained appropriately. Yet, if learning another language requires a substantial effort of memory and auditory and articulatory discrimination from a human, it seems that machines might have an advantage here.

A lot of research has been devoted to multilingual speech recognition over the last decade. Often the object of these studies is to leverage acoustic data from languages with large training corpora to help train acoustic models for lesser represented languages, see e.g. [2, 3]. This work instead focuses on improving mature, data-rich, systems that happen to contain words from multiple languages.

In this paper, we describe a few acoustic modeling experiments we made towards multilingual recognition. For concreteness, we focus on Voice Search queries made in English by Mandarin-speaking users from Taiwan.

## 2. DATASETS

Several datasets were used in this paper, including a training set of one million Mandarin Voice Search utterances, and a training set of two million American English Voice Search utterances. Both datasets were transcribed by native speakers of these languages. We didn't use any additional untranscribed training data in order to keep a fast experimental turn around.

The primary test set used throughout the paper consists of 16K transcribed utterances (60K words/characters) collected in Taiwan. It consists of roughly 14K utterances containing only Chinese words, and 2K utterances containing only English words. Too few utterances combine words from both languages to make a significant test set. In the rest of the paper, we refer to the entire test set as 'ZH', and to its two subsets as 'ZH.ZH' and 'ZH.US'.

In one experiment, we contrast a British English (GB) recognition system with the American English (US) system. The corresponding test sets consist of 23K Voice Search GB utterances (53K words), and 27K Voice Search US utterances (87K words).

## 3. BASELINE SYSTEM

### 3.1. Mandarin Voice Search System

The speech recognition engine used in Google Voice Search is a standard, large-vocabulary recognizer, with PLP features and LDA, decision trees, GMM-based triphone HMMs with variable numbers of Gaussians per state, STC [4] and an FST-based search [5]. ML training is followed by boosted MMI [6]. The language model is a 3-gram model trained from web search queries. A confidence score between 0 and 1 is estimated for each recognized utterance.

The Mandarin Voice Search models rely on a 75 phoneme/toneme phone set, where different tones are modeled with different units. No special front-end effort (e.g. pitch modeling) was made to handle tones. A detailed description of this system is provided in [7].

All the baseline systems refered to in this paper were thoroughly optimized: the sizes of the models were optimized given the training sets, ML and MMI iterations were repeated until convergence was reached, and the various decoder parameters were tuned for sentence accuracy under real-time recognition constraints.

Recognition performance is reported in terms of sentence accuracy (SACC).

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

### 3.2. Handling of English Words in the Mandarin Baseline System

To train the baseline acoustic model, an expert-made English-to-Mandarin context-independent phone mapping was first created. For the most part, this is a one-to-one mapping, with English vowels most often mapped to a fourth tone Mandarin phone (see example in Table 1). In some instances, English diphtongs were mapped to a sequence of two Mandarin phones, e.g. 'oy → uo4 i4'. This mapping was applied to our baseline US lexicon to produce mapped pronounciations that could be appended to the Mandarin lexicon. The appended lexicon was used for training and testing.

| | | |
|---|---|---|
| US Lexicon: | `google` | `g uw g ax l` |
| ZH Lexicon: | `google` | `g u4 g @1 l` |

**Table 1**. *Mapping of US lexicon entries to ZH in the ZH baseline system.*

Performance with this approach is reported in Table 2. The sentence accuracy on the English subset of the test set is roughly 6 points worse than that on the Mandarin subset.

| System | SACC (%) | | |
|---|---|---|---|
| | ZH | ZH.ZH | ZH.US |
| ZH | 58.1 | 59.7 | 53.9 |

**Table 2**. *Recognition accuracy with the baseline ZH models, on the ZH test set and its subsets, ZH.ZH and ZH.US.*

## 4. SYSTEM COMBINATION

System combination has long been a favorite technique in speech recognition: systems that make different errors can be combined to produce more accurate results. This principle can be applied at various levels of the system, including in the front-end as in e.g. [8], or at the system output as in ROVER [9].

In the context of multilingual recognition, it is natural to consider combining monolingual systems, and choosing what appears to be the best recognition output from the various systems. This approach won't handle queries where users switch languages, but if one of the systems has some support for multilingual queries, which our baseline ZH system does, then combining it with, for example, a US recognizer could help recognize English-spoken queries.

In the experiments below, we combine a 'primary' system, A, with a 'secondary' system, B, by chosing the recognition result of the primary system unless the confidence of the secondary system is greater by some margin:

$$\text{A+B:} \qquad \text{Pick A unless Conf(B)} > \text{Conf(A)} + \delta, \tag{1}$$

where the margin, $\delta$, can be empirically optimized. This method can easily be extended to combining multiple systems. In the next two sections, we evaluate and compare two model combinations: GB+US and ZH+US. We will see that they behave quite differently.

### 4.1. Combination of British and American English Systems

The results of a system combination for two dialects of the same language, GB and US English, are summarized in Table 3.

| System | SACC (%) | |
|---|---|---|
| | US | GB |
| US | 75.6 | 60.5 |
| GB | 63.0 | 66.3 |
| GB+US | 76.1 | 66.9 |

**Table 3**. *Recognition accuracy for the baseline GB and US systems, and for the GB+US combination, measured on the US and GB test sets. The confidence margin, $\delta$, was 0.1.*

Whereas the US system performs worse on the GB test set than the GB system (60.5% SACC vs 66.3%), and vice-versa (63.0% vs 75.6%), the combined system (third line) helped both test sets by roughly half a percent. This simple model combination was thus very effective in reducing the error rates of two well-optimized systems.

### 4.2. Combination of Mandarin and American English Systems

The same technique was then applied to combine the baseline Mandarin system with the US system, in the hope of improving the recognition of English-spoken queries in the Mandarin test set. However, as indicated in Table 4, the combination brought almost no improvement (58.1 to 58.2% SACC). For reference, the oracle performance evaluated by marking each sentence as correct if either system got it right was 59.4%, indicating that the US system could have corrected some recognition errors on English queries.

| System | SACC (%) |
|---|---|
| | ZH |
| ZH | 58.1 |
| US | 5.2 |
| ZH+US | 58.2 |
| Oracle | 59.4 |

**Table 4**. *Recognition accuracy combining the baseline ZH and US systems, evaluated on the ZH test set. The confidence margin, $\delta$, was 0.7.*

It is interesting to note that the combination algorithm chose to set the confidence margin to 0.7: a lower $\delta$ would have produced a worse overall accuracy. This relatively large value reflects the fact that the combined system needs to prevent the US model from false-triggering on ZH sentences, which it achieves by increasing the confidence value required from the US model to surpass the ZH model. This in turn prevents the US system from contributing much to the combination, even on utterances where it theoretically could. The GB+US combination didn't suffer from that problem because they both 'understand each other's language', like two bilingual systems, so they don't need to compete so hard with each other.

## 5. TAG ACOUSTIC MODEL

The model combination approach described in Section 4 combines entire systems, chosing the recognition result from one or the other. Another path for multilingual modeling consists in training several languages simultaneously within a single acoustic model. Such a model can be used as is, or as a basis for further parameter tying as will be described in Section 6.

The Tag model defined here is trained by concatenating the training corpora and lexicons of multiple languages, after tagging all words and phones with the language of the corpus or lexicon they come from. Thus the word 'google' appearing in the US training corpus will be tagged with a 'US' tag, and the word 'google' appearing (in English) in the ZH training corpus will be tagged with 'ZH', along with Mandarin words. Pronunciations for these words are showed in Fig. 1. The lists of decision-tree questions for the two languages are also tagged and concatenated.

| | |
|---|---|
| google_US | g_US uw_US g_US ax_US l_US |
| google_ZH | g_ZH u4_ZH g_ZH @1_ZH l_ZH |
| 谷歌_ZH | g_ZH u3_ZH g_ZH @1_ZH |

**Fig. 1**. *Language-tagged lexical entries with words from the US training corpus (first line) and from the ZH corpus (next two lines).*

The Tag model is different from both the "ML-Sep" and "ML-Tag" approaches described in [2] in that we use data from each corpus to train triphone-state models specific to that corpus tag. This acoustic model is similar but not equivalent to a 'merge' of corpus-specific acoustic models; sharing occurs at various levels. First, the silence model is shared across languages. Second, LDA is performed on the merged phone set, resulting in a shared feature space for further ML and MMI iterations. Third, since the word lattices used for MMI training are estimated from the (merged) training corpus, discriminative training will consider inter-language errors as well as intra-language errors. This may be beneficial for language discrimination, but it may also impose an unwanted differentiation between instances of the same word in different languages, such as google_ZH and google_US.

| System | SACC (%) | | |
|---|---|---|---|
| | ZH | ZH.ZH | ZH.US |
| ZH | 58.1 | 59.7 | 53.9 |
| Tag | 57.8 | 59.2 | 55.7 |
| ZH+Tag | 59.6 | 61.1 | 57.4 |

**Table 5**. *Recognition accuracy of the Tag model and ZH+Tag combination on the ZH testset and its subsets. The confidence margin, $\delta$, was 0.2.*

Table 5 compares the Tag model to the baseline ZH model, along with the combination of both. Overall, the Tag model does worse on ZH utterances (-0.5% SACC) and better on US utterances (+1.8%). However, the system combination improves on both subsets, with a 1.4% SACC gain on the ZH utterances, and 3.5% on the US utterances. This model combination assembles two bilingual systems, and its confidence margin is small, 0.2, very much like the GB+US combination.

## 6. DATA-DRIVEN PHONE MAPPING

The Tag model described in Section 5 contains triphone units from the various languages we wish to combine, and constitutes a good support to explore parameter tying strategies. In contrast with previous work on this topic, e.g. [10], the long-term aim of our research is to share individual Gaussians across languages. As an early experiment in this direction, we chose to use the Tag model we trained for Mandarin and English to derive a phone-level mapping and build a system similar to the baseline ZH system, but where the phone mapping is data-driven rather than expert-made.

### 6.1. Gaussian Clustering

A simple measure of phone similarity can be derived by clustering all the Gaussians in the Tag model in a single, phone-independent, language-independent, KL-based, Vector Quantization (VQ) codebook. If any two phones in the model have the majority of their Gaussians lying in common VQ clusters, then we can assume that these phones are similar.

In this experiment, we trained a 256K cluster VQ partitioning of the Gaussians in the Tag model. Since the Tag model has 588K Gaussians, the clusters contain a little over two Gaussians each on average. Gaussians clustered together may belong to triphone states tagged with the same language, or with different languages. Interestingly given how different the two languages are, we found that as much as 20% of the Gaussians were clustered in bilingual clusters.

For each US/ZH phone pair, we then computed a phone similarity as the fraction of shared Gaussians across all triphone states of these two monophones. For example, if k_US has a total of 5K Gaussians across all its triphone states and t_ZH has 10K Gaussians across all its states, if 3K out of the 15K Gaussians lie in common VQ clusters, we define the similarity betwen k_US and t_ZH as 3K/15K = 0.2.

### 6.2. Similarity Map for US/ZH

Monophone similarities were computed for all US/ZH phone pairs, and the resulting similarity matrix was plotted as a color-coded map, as shown in Fig. 2, with English phones on the Y axis, and Mandarin phones on the X axis. The knowledge-based mapping used in the ZH baseline system is shown as black dots in the colored grid boxes.

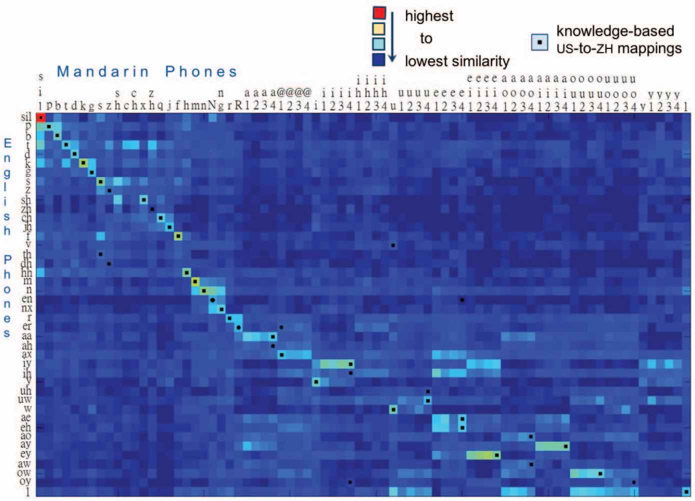

**Fig. 2**. *Similarity matrix for English/Mandarin phone pairs.*

The highest similarity box, in red in the upper-left most corner, corresponds to the sil_US/sil_ZH phone pair, which has a similarity of 1.0 since the silence model is shared across languages in the Tag model. Other phone pairs, such as f_US/f_ZH or m_US/m_ZH have a high similarity (yellow boxes on the map). Low similarity phone pairs are colored in increasingly dark shades of blue. Lighter color horizontal stripes, especially in the lower-right quarter of the map, show that English vowels are often similar to all 4 (or 5) tones of the corresponding Mandarin vowels. Also, short English vowels tend to be similar to several Mandarin vowels, tone set aside. For example, iy_US maps with a relatively large similarity to i_ZH, ey_ZH, and ih_ZH, indicating how complex the similarity patterns are, and how suboptimal a simple phone-level mapping is likely to be. Another expected phenomenon illustrated on the map is the high similarity betwen plosives (e.g. p, t, k, b) and silence.

The knowledge-based mapping (black dots) appears to correlate fairly well with brighter spots on the map, but not necessarily with the brightest spot, as we will explore in Section 6.4.

### 6.3. Similarity Map for ZH/ZH

Out of curiosity, we also derived a similarity map for Mandarin only, see Fig. 3. It shows, as might be expected, a certain amount of cofusability amongst consonants, and also a strong similarity amongst tones of the same vowels, reminding us that our recognizer does not display a very strong tone discrimination. Nonetheless, a nice red line runs diagonaly through the map, indicating a high self-similarity for all the phones.

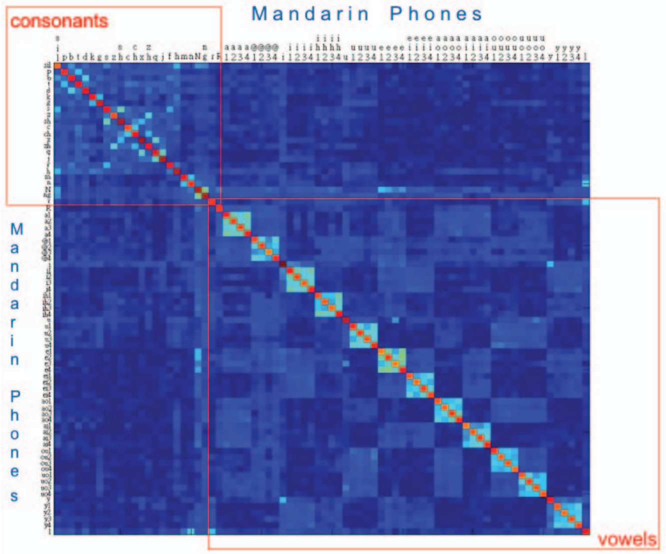

**Fig. 3**. *Similarity matrix for Mandarin/Mandarin phone pairs.*

### 6.4. Data-Driven Phone Mapping

An English to Mandarin phone mapping can be derived from Fig. 2 by chosing for each English phone the highest-similarity (brightest) Mandarin phone. This mapping can be used instead of the knowledge-based mapping to train a model similar to the baseline ZH model. We call it the 'Remap' model.

Table 6 shows the performance of the Remap model: the data-driven mapping appears to be as good as the knowledge-based mapping, but no better. As it turns out, the data-driven mapping was actually 0.5% better than the knowledge-base mapping in the context-independent iterations of model training, but that gain vanished as more Gaussians were grown during the context-dependent iterations. It is possible that a triphone-state mapping as proposed in [11] would provide a better final system than the monophone mapping.

The Remap model however significantly helps the ZH+Tag combination, especially on English-spoken queries. With the three-way combination, the recognition accuracy of Mandarin utterances improved by 1.7% over the ZH baseline, while the accuracy of the English utterances improved by 5.5% absolute, showing how much the three models differ in their ways of handling queries spoken in English.

| System | SACC (%) | | |
|---|---|---|---|
| | ZH | ZH.ZH | ZH.US |
| ZH | 58.1 | 59.7 | 53.9 |
| Remap | 58.1 | 59.7 | 53.6 |
| ZH+Tag | 59.6 | 61.2 | 57.4 |
| ZH+Tag+Remap | 60.3 | 61.4 | 59.4 |

**Table 6**. *Recognition accuracy with the Remap model, and with the combination of the three bilingual models, ZH, Tag, and Remap.*

## 7. CONCLUSION

This paper described some early experiments to improve the recognition performance of queries spoken in English to a Mandarin Voice Search system. With a system combination approach, we improved the sentence accuracy of English-spoken queries by 5.5% (12% relative) while increasing that of Mandarin-spoken queries by 1.7% (4% relative), over a baseline already containing some support for bilingual queries. We learned from these experiments that the Tag model we defined provides an interesting framework to apply data sharing techniques such as the Gaussian clustering. It is likely however that sharing at the monophone level is too coarse, and that we should consider lower-level modeling units, such as triphone states. We also verified, again, that system combination is a powerful method to achieve accuracy wins.

## 8. REFERENCES

[1] G. R. Tucker, "A Global Perspective on Bilingualism and Bilingual Education", CMU, 1999.
http://www.cal.org/resources/Digest/digestglobal.html.

[2] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition", Speech Communication (35), 2001.

[3] T. Niesler, "Language-dependent state clustering for multilingual acoustic modeling", Speech Communication (49), 2007.

[4] M. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models", Proc. IEEE Trans. SAP, May 2000.

[5] OpenFst Library, http://www.openfst.org

[6] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, K. Visweswariah, "Boosted MMI for model and feature-space discriminative training", Proc. ICASSP, 2008.

[7] J. Shan, G. Wu, Z. Hu, X. Tang, M. Jansche, P. Moreno, "Search by Voice in Mandarin Chinese", Proc. Interspeech, 2010.

[8] A. Janin, D. Ellis, N. Morgan, "Multi-Stream Speech Recognition: Ready for Prime Time?", Proc. Eurospeech, 1999.

[9] J.G. Fiscus, "A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", Proc. ASRU workshop, 1997.

[10] Q.-Q. Zhang, J.-L. Pan, Y.-H. Yan, "Development of a Mandarin-English Bilingual Speech Recognition System with Unified Acoustic Models.", Journal of Information Science and Engineering 26, 2010.

[11] H. Cao, T. Lee, P.C. Ching, "Cross-Lingual Speaker Adaptation via Gaussian Component Modeling", Proc. ICSLP, 2010.