RECOGNITION OF MULTILINGUAL SPEECH IN MOBILE APPLICATIONS

Hui Lin[*†], *Jui-ting Huang*[*‡], *Françoise Beaufays*[*], *Brian Strope*[*], *Yun-hsuan Sung*[*]

[*] Google Inc, [†] Univ. of Washington, [‡] Univ. of Illinois

ABSTRACT

We evaluate different architectures to recognize multilingual speech for real-time mobile applications. In particular, we show that combining the results of several recognizers greatly outperforms other solutions such as training a single large multilingual system or using an explicit language identification system to select the appropriate recognizer. Experiments are conducted on a trilingual English-French-Mandarin mobile speech task. The data set includes Google searches, Maps queries, as well as more general inputs such as email and short message dictation. Without pre-specifying the input language, the combined system achieves comparable accuracy to that of the monolingual systems when the input language is known. The combined system is also roughly 5% absolute better than an explicit language identification approach, and 10% better than a single large multilingual system.

Index Terms— Multilingual speech recognition, acoustic modeling.

1. INTRODUCTION

By some estimates[1], more than half of the world's population is multilingual, however most commercial recognition systems remain monolingual. At the same time, speech recognition is now being used both to get information from machines (e.g. speak a Google query) but increasingly to communicate with people by dictating short messages. Together there is increased pressure to recognize whatever language might be most appropriate for whatever setting, without requiring the user to navigate language-selection interfaces, which themselves are complicated by keyboard requirements and other geographic considerations. An omnilingual or at least multilingual recognizer would make many of these interactions more natural, but few users would choose that trade-off if accuracy or latency were degraded.

With those constraints in mind, we evaluated several multilingual techniques on datasets representative of our current mobile traffic, which is a mix of Voice Input (usually short dictation), Voice Search (Google queries), and Voice Actions (commands). These are mostly short utterances recognized in real-time. We started with a trilingual English-French-Mandarin scenario, and evaluated each technique on a union of three test sests representative of those languages.

Recognizing mixed or "code-switched" utterances (where the language changes mid-sentence) was not an explicit focus in this work: providing accurate multilingual recognition was considered a more fundamental first step. Some limited support for code-switching is already built into our systems because the underlying data-driven learning techniques respond to foreign terms of reasonably high frequency. Also, as explained below, the Mandarin baseline system has special provisions for recognizing English words since English is so common in our Chinese mobile traffic.

Finally, this paper is not addressing the often-studied problem of under-resourced languages (see e.g. [2, 3]). Rather we're focussing on mature languages where we have a steady stream of recognition requests, and training and test sets collected from real traffic. All the systems described in the paper were trained to be production quality, i.e. discriminatively trained, and optimized for our production accuracy and latency requirements.

2. DATASETS

The English (En), French (Fr), and Mandarin (Zh) training sets each contain roughly 1.7 million utterances (over 1500 hours of speech) that span the different voice-enabled mobile applications used in those languages. These were all recently collected, are mostly untranscribed (a small portion of the Mandarin training set is still transcribed), and were confidence-filtered for unsupervised training. Three test sets and three development sets were drawn from the same traffic at different time intervals and hand-transcribed. We report accuracies on the test sets. The dev sets were used to optimize the combination classifiers described below. The test sets for English, French, and Mandarin contain, 24K, 37K, and 80K utterances respectively, and the dev sets have similar sizes.

Recognition accuracy is expressed in terms of normalized sentence accuracy ('SACC'), where hyphens, apostrophes, etc. are stripped from the word strings before comparisons with the human reference. For each experiment, we report the sentence accuracy of each individual test set (language), and also the overall accuracy ('Avg') where, given no starting estimate for typical multilingual traffic, we assumed all three languages were equiprobable. The Mandarin test set was further split into three subsets: Mandarin only (Zh - 72K utterances), English only (En - 6K utterances), mixed Mandarin-English (Mix - 2K utterances). Accuracies for the subsets are also reported.

3. BASELINE SYSTEM

The speech recognition engine used for mobile Google applications 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 models are N-gram models (N=4 for English and Mandarin, 3 for French) trained from a variety of Google typed and spoken sources relevant to the overall traffic. A confidence score between 0 and 1 is estimated from lattice-posterior metrics for each recognized utterance.

The Mandarin baseline model relies on a 75 phoneme/toneme phone set, where different tones are modeled as different units. A detailed description of this system is provided in [7]. Because more than 10% of our Chinese speech data contains English words, the

Mandarin system also has explicit provisions for English recognition: words from the English lexicon are phonetically mapped to the Mandarin phone set, and added to the Chinese lexicon. While our recognition of English words in Chinese remains significantly inferior to that of Chinese words (roughly 5% absolute), the baseline Mandarin system is by nature bilingual. The French system does not contain any similar explicit support for English words, though the phone sets are closer to start with and some lexical overlap results from the similarity of the two languages. In addition, data-driven training brought extra foreign words in the French models as well.

System	SACC (%)			
	En	Fr	Zh (Zh/En/Mix)	Avg
En	60.5	3.4	2.8 (0.0/35.8/0.0)	22.3
Fr	0.8	46.2	0.9 (0.0/11.2/0.1)	16.0
Zh	3.8	2.1	41.9 (42.6/37.3/32.5)	16.0

Table 1. *Monolingual system accuracies.*

Table 1 shows the baseline accuracies of the three monolingual systems, on the three test sets. Accuracies of less relevant pairs, such as recognizing the French test set with the Mandarin recognizer, are greyed-out, even though they are not strictly zero due to language overlap (the most outstanding being 35.8% accuracy on the English portion of the Mandarin test set, when recognized with the American English recognizer. The overall recognition accuracy is 49.5% (average of the three bold accuracies in the table). This assumes the input language is pre-specified.

4. SINGLE MULTILINGUAL SYSTEM

One direct technique for multilingual recognition is to train a 'universal' acoustic model, capable of recognizing all (relevant) languages. This approach holds the promise of helpful data sharing between languages, and has been explored in various ways by many researchers [10], including us [11]. It is also attractive for its ease of maintainability (one model for all languages), but might require fundamental decoder changes to accommodate large models while maintaining low-latency characteristics.

The mixed model we report on here ('Mix' below) was trained by merging the training sets and phone sets of all three languages (119 phones total). Pronunciations for each training word were extracted from the corresponding lexicon (or pronunciation engine). Words appearing in several languages have pronunciations in several languages. This results in an average of 1.3 pronunciations per word, which is in line with the monolingual lexicons. The model contains roughly 900K Gaussians, which is a little under the sum of the number of Gaussians of the individual monolingual systems. It runs roughly 2 times slower than the monolingual systems. Its accuracy is summarized in Table 2.

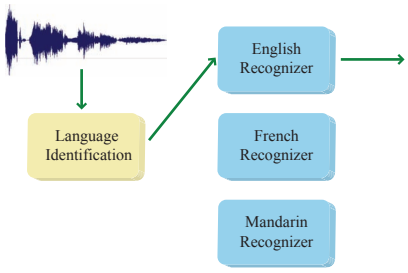
System	SACC (%)			
	En	Fr	Zh (Zh/En/Mix)	Avg
Mix	41.6	39.8	32.6 (31.8/42.6/26.4)	38.0

Table 2. *Mix system accuracy.*

Despite our best attempts, we found it difficult to obtain high accuracies with a single model approach [12]: the overall accuracy is more than 10% worse than the monolingual baseline (38 vs 49.5%). Again, this is a data-saturated environment, so pooling resources does not compensate for data sparsity, and does not readily provide accuracy benefits. The exception is the English subset of the Mandarin test set, which largely benefited from the added American English training data (from 37.3 to 42.6% SACC).

We also trained a 'tagged' system, where all phonemes and lexicon entries were tagged by their language ID (so there is little data sharing), and measured an even lower overall accuracy: 36.9%.

5. LANGUAGE IDENTIFICATION



The next strategy we evaluated for multilingual recognition used a language identification module to direct the utterance to the appropriate recognizer (called the 'LangId' system below). The main advantage of this architecture is its simplicity and low computational cost: it only inserts a relatively simple decision-making module in the recognition flow. This module, however, would need to be very low-latency, and most of the literature on this topic offers off-line, or batch, solutions. Setting this issue aside, we evaluated the recognition accuracy that such an architecture could achieve.

For this experiment we used an implementation of the language identification algorithm described in [8], which is a discriminative extension to the popular MAP-SVM architecture widely used for such tasks [9]. In MAP-SVM, a universal background Gaussian-mixture model (UBM) is used to model each utterance as its maximum a posteriori departure from the UBM. The parameters of this model are then stacked in a 'supervector' that is classified by a vector support machine (SVM). The specific implementation used here was previously validated on publicly available datasets.

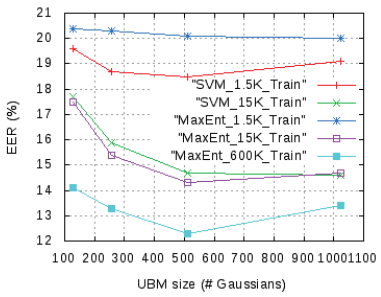


Fig. 1. *Language id EER as a function of the classifier training size.*

We trained and optimized the language identification module for En-Fr-Zh classification assuming equal priors for all three languages. Performance was estimated and optimized on the development sets, and is reported in terms of equal error rate (EER). We found that, even with fairly large amounts of UBM training data (5M utterances total), performance saturated at ∼500 Gaussians. The amount of SVM training data however seemed to have more impact on classification performance (see Fig. 1). Since our SVM

implementation was not scalable, we compared the SVM to a Maximum Entropy (MaxEnt) module which is more readily parallelizeable. Optimizing the UBM size and sweeping the MAP adaptation parameter for increasing amounts of data, we found that by being able to leverage more training data, MaxEnt outperformed our SVM solution by ∼20%.

Nonetheless, the asymptotic EER remains fairly high (over 10%). This is due to the fact that our mobile recognition requests are very short, on average 3.5 seconds (including non-speech frames). Fig. 2 shows the EER distribution as a function of utterance length, together with the distribution of utterance lengths over the dev set. The EER decreases as the sentence length increases, but the mode of the data is at ∼15% EER.

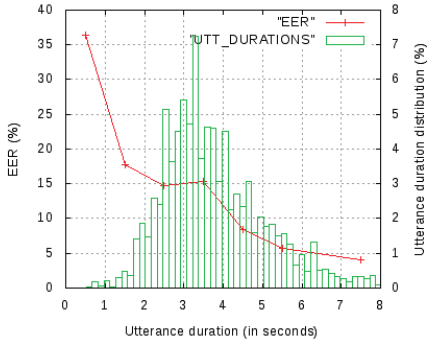


Fig. 2. *Language id EER as a function of utterance length.*

Applying the best language classifier to the trilingual recognition task, we obtained an overall sentence accuracy of 43.5%, better than the single multilingual system, but still much worse than the average monolingual performance of 49.5% (see Table 3).

System	SACC (%)			
	En	Fr	Zh (Zh/En/Mix)	Avg
LangId	50.4	38.8	41.4 (42.5/32.9/29.5)	43.5
LangId (margin)	53.2	41.1	41.0 (42.4/30.1/26.2)	45.1
LangId (oracle)	60.5	46.2	41.9 (42.6/37.3/32.5)	49.5

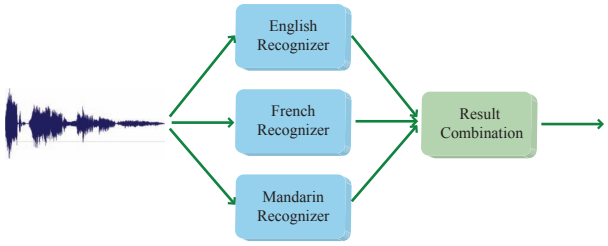
Table 3. *LangId system accuracies.*

Following the approach proposed in [11] for system combination, we considered adding a margin to each language identification score to compensate for possibly biased language decisions. Margins of 0.4, 0.4, 0.0 for En, Fr, Zh, respectively, helped increase the overall system accuracy by 1.6% absolute. Further analysis showed that the En and Fr margins helped roughly 5% of the test utterances in each of these languages 'move' from the Zh system to their own, giving them a better chance of being correctly recognized. With margins, the LangId system remains 4.4% absolute worse than the average monolingual systems (45.1% when the language is estimated, vs 49.5% when the language is known). There just aren't enough phones in an average utterance to estimate its language reliably.

6. SYSTEM COMBINATION

6.1. Combining 3 Monolingual Recognizers

Our work on trying to improve the recognition of English words in Chinese systems [11, 12] taught us that a much easier path to multilingual recognition is that of system combination. Using several recognizers and combining their outputs consistently outperforms any



single recognition system. This approach requires more run-time resources since it performs multiple recognitions for each request, but its latency is only bounded by the slowest model, which itself is limited by timeouts. We tested different flavors of this approach on our trilingual task.

The first approach consists of simply comparing the confidence scores of each monolingual recognizer, and choosing the highest-confidence result. This alone resulted in a slightly better overall accuracy than the LangId system (45.7 vs. 45.1%, see Table 4).

As with the LangId system, we tried applying a confidence margin to each recognizer. Margins of 0.2, 0.3, 0.0 increased the overall accuracy by 1.2% absolute. As with the LangId system, the margins helped move En and Fr utterances off the Zh system, but this only helped the French test set here, likely because the classification criterion in this scenario was recognition confidence, and En utterances with a high confidence score under the Zh system were (almost) equally well recognized by the Zh than the En system.

Adding an extra constant to reinforce results where 2 recognizers agree (third line in the table) didn't help, except on the subset of English queries in the Mandarin test set. This is expected since these utterances can best benefit from recognition redundancy.

Finally, an SVM with Gaussian kernel was trained to choose the best recognition result based on the recognizers confidence scores, the output of a language identification system (running in parallel with the speech recognizers), and additional recognition agreement features (one per language pair). The SVM was trained on the dev sets. This last system performed 1.7% absolute better than the best confidence-based system (from 46.9 to 48.6%), 3.5% better than the best LangId system (45.1%), 10.6% better than the Mix system (38.0%), and only 0.9% worse than the monolingual systems (49.5%). Also, it brought the accuracy on each individual test set close to that of the monolingual recognizers, especially French.

Notice that an oracle system that would pick the best system (from an accuracy viewpoint, not based on knowing the language a priori) does surpass the monolingual accuracy (50.3% vs 49.5%), indicating that the little bit of redundancy between the three systems could possibly be further exploited, especially for English queries in the Mandarin test set.

6.2. Combining 3 Monolingual and 1 Multilingual Recognizers

Although the Mix system has a relatively poor recognition accuracy, it tends to make different errors than the monolingual systems, and it offers redundancy with all three. It is thus not surprising that adding it to the combination boosts the overall accuracy by 1.2% (from 48.6 to 49.8% with an SVM decision logic, see details in Table 5). With the Mix system, the overall accuracy slightly surpasses that of the monolingual systems. In oracle experiments where the best recognizer is chosen, the Mix system brings 3.5% absolute improvement, from 50.3 to 53.8% (though of course this is a slightly misleading as it compares oracles with different number of input sources).

The Mix system seems to most help the French test set (37.0

System	SACC (%)			
	En	Fr	Zh (Zh/En/Mix)	Avg
En+Fr+Zh (conf)	58.0	37.0	42.1 (42.4/43.7/28.6)	45.7
En+Fr+Zh (conf+margin)	58.0	41.6	41.2 (41.9/39.9/22.1)	46.9
En+Fr+Zh (conf+margin+agree)	58.0	41.6	41.3 (41.9/40.5/22.1)	46.9
En+Fr+Zh (conf+agree+LangId,SVM)	58.9	44.9	42.1 (42.4/42.3/29.1)	48.6
En+Fr+Zh (oracle)	60.7	47.1	43.0 (42.6/51.4/32.6)	50.3

Table 4. *Systems combination accuracies.*

System	SACC (%)			
	En	Fr	Zh (Zh/En/Mix)	Avg
En+Fr+Zh+Mix (conf)	57.9	39.9	42.0 (42.0/45.6/30.1)	46.6
En+Fr+Zh+Mix (conf+margin)	57.8	42.4	41.5 (41.8/43.0/27.0)	47.3
En+Fr+Zh+Mix (conf+margin+agree)	59.0	44.3	42.4 (42.5/45.3/28.9)	48.5
En+Fr+Zh+Mix (conf+agree+LangId,SVM)	59.8	46.7	42.8 (42.8/46.9/31.7)	49.8
En+Fr+Zh+Mix (oracle)	63.4	51.5	46.4 (45.8/55.7/37.4)	53.8

Table 5. *Systems combination accuracies, including the Mix recognizer.*

to 39.9% without margin, 41.6 to 42.4% with margin). Indeed, we found that, after application of margins, still 7% of the French test set is recognized by the Mix system rather than the French system.

The agreement feature has a larger effect in this experiment where more system redundancy can be exploited, contributing an increase in overall system accuracy from 47.3 to 48.5%.

Incorporating language identification scores in the combination increases the overall accuracy by 1.3%, from 48.5 to 49.8%, which is slightly superior to the average performance of the monolingual systems.

Overall, the system combination approach is scalable, accurate, and fast. Simply using the monolingual models together with language identification scores and some reasonable decision logic brings the combined system close to monolingual accuracy. The inevitable accuracy loss due to inter-language confusions is not quite compensated by exploiting the small redundancy that exists between the underlying models, unless we also add to the combination a multilingual model. However a 'Mix' multilingual recognizer must overcome significant challenges related to size and scalability to be helpful for different language configurations in a realtime environment. Fortunately, performance is also very strong with purely monolingual combinations.

7. CONCLUSION

We explored several architectures that could be implemented to help multilingual speakers use speech recognition applications in the languages of their choice. We focused our study on a trilingual task of English-French-Mandarin recognition for mobile applications. We found that short speech requests make it difficult to rely on a language identification system to pick the right recognizer, and that large monolingual systems don't perform nearly as well as the monolingual systems when the language is known.

The most promising approach is to allow utterances to be recognized in parallel by several systems, and combine the scores of these systems with a classifier. A simple confidence voting scheme between three monolingual systems brought us closer to monolingual accuracy than any system we previously evaluated, and adding other easily-computed knowledge sources such as language ID scores helped bridge most of remaining accuracy gap.

These results suggest that accurate multilingual automatic speech recognition could be primarily a question of machines and related support costs.

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] J. Kohker, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks", ICASSP 1998.

[3] N.T. Vu, F. Kraus, T. Schultz, "Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training", Interspeech 2011.

[4] M. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models", 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", ICASSP 2008.

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

[8] C. Alberti, M. Bacchiani, "Discriminative Features for Language Identification", Interspeech 2011.

[9] W. Campbell, J. Campbell, D. Reynolds, E. Singer, P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", Computer Speech & Language, vol. 20, no. 2-3, 2006.

[10] T. Schultz, A. Waibel, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition", Speech Communication, Vol. 35, 2001.

[11] H.A. Chang, Y.H. Sung, B. Strope, F. Beaufays, "Recognizing English Queries in Mandarin Voice Search", ICASSP 2011.

[12] J.T. Huang, H. Lin, Y.H. Sung, B. Strope, F. Beaufays, "System Combination to Recognize Mandarin and Accented English", ICASSP 2012.