EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Pronunciation augmentation for Mandarin-English code-switching speech recognition

Yanhua Long[1]*, Shuang Wei[1], Jie Lian[1] and Yijie Li[2]

Abstract

Code-switching (CS) refers to the phenomenon of using more than one language in an utterance, and it presents great challenge to automatic speech recognition (ASR) due to the code-switching property in one utterance, the pronunciation variation phenomenon of the embedding language words and the heavy training data sparse problem. This paper focuses on the Mandarin-English CS ASR task. We aim at dealing with the pronunciation variation and alleviating the sparse problem of code-switches by using pronunciation augmentation methods. An English-to-Mandarin mix-language phone mapping approach is first proposed to obtain a language-universal CS lexicon. Based on this lexicon, an acoustic data-driven lexicon learning framework is further proposed to learn new pronunciations to cover the accents, mis-pronunciations, or pronunciation variations of those embedding English words. Experiments are performed on real CS ASR tasks. Effectiveness of the proposed methods are examined on all of the conventional, hybrid, and the recent end-to-end speech recognition systems. Experimental results show that both the learned phone mapping and augmented pronunciations can significantly improve the performance of code-switching speech recognition.

Keywords: Code-switching, Phone mapping, Pronunciation variation, Lexicon learning, Speech recognition

1 Introduction

Code-switching (CS) phenomenon is prevalent in many multilingual communities. It is defined as the switching of two or more languages at the conversation, utterance, and sometimes even word level [1–3]. There are two different forms of code-switching, one is the inter-sentential switching with the alternation is between sentences, and the other is the intra-sentential with the switching is within one sentence or word [3].

The code-switching phenomenon is quite common around the world. For example, in India, it is very common to see the Bengali-English or Bengali-Hindi-English in most people's daily speech [4]; in USA and Switzerland, people can often hear Spanish-English and French-Italian code-switching speech [3]; and in Hong-Kong, the combination of English and the native Cantonese is also very common [5]. Particularly, in East Asia, the Mandarin-English code-switching is extremely popular, such as in Singapore, Malaysia, Mainland China, and Taiwan [6, 7]. In addition, the code-switching is also now frequently found in our daily life, such as in some professional activities, social media, consumer goods, or entertainment, it is fairly common to hear people borrowing words from one language to use them in another [8, 9]. In recent years, the research of code-switching automatic speech recognition (ASR) has received increasingly attention. This is because with the rapid development of speech technology, variety speech-driven interfaces to smart devices, and other real AI applications become mainstream, most state-of-the-art monolingual ASR systems fail when they encounter code-switched speech.

*Correspondence: yanhua@shnu.edu.cn
[1] SHNU-Unisound Joint Laboratory of Natural Human-Computer Interaction, Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai, China
Full list of author information is available at the end of the article

Compared with the recent significant success achieved in monolingual speech recognition [10, 11], the ASR systems still have problem to deal with the code-switching speech, especially the intra-sentential switching. To build a good code-switching ASR system, several challenges need to be handled, either in acoustic or language modeling. One of the major challenge is the pronunciation variation phenomenon of the embedding language at the code-switches. Unlike the matrix language for native speakers, in many CS scenarios, most code-switching speakers may not be familiar with the embedding language, the words borrowed from the embedding language may be pronounced with a spectrum of accents and may be systematically or randomly mispronounced [8]. For example, in the Mandarin-English code-switching utterances that collected from Mainland of China, most of those embedded English words may be Chinglish (Chinese English). There is a significant pronunciation variation between the Chinglish and the standard British or American English. Although there were many previous works have been proposed to deal with the discrepancy and co-articulation effects between different CS mixed languages, such as the units merging [8, 12–14], language-universal acoustic modeling targets, or framework [15–17]. Works related to handle the pronunciation variation of embedding words are very limited.

In this study, we concentrate on exploring pronunciation augmentation techniques for acoustic modeling. These techniques are applied and examined to improve a Mandarin-English intra-sentential code-switching ASR system. We aim at dealing with the pronunciation variation and alleviating the sparse problem of code-switches by using pronunciation augmentation methods. Our contributions are as follows: (1) an English-to-Mandarin mix-language phone mapping is proposed. We first validate the effectiveness of conventional data-driven phoneme sharing. However, the direct one-to-one unit mapping only helps to alleviate the embedding language training data sparsity at some extent. The acoustic discrimination between different languages is ignored. Therefore, we propose a new English-to-Mandarin phone mapping to enhance the pronunciations of universal code-switching lexicon. (2) An acoustic data-driven lexicon learning framework is proposed to learn new pronunciations to cover the accents, mis-pronunciations, or pronunciation variations of those embedding English words. Only using the phone mapping still can not well handle the pronunciation variation with the mispronounced, accented embedding words or phrases. Because the standard pronunciations in the monolingual lexicon can not cover these variation cases, these words would typically need to be expressed with another new pronunciation or phone set. Therefore, motivated by the acoustic data-driven lexicon learning in [18], we propose a novel pronunciation augmentation approach to produce the possible new pronunciations for those embedding words at the code-switches. Based on the merged universal code-switching phone set, this approach integrates both of the information from the expert knowledge, and acoustic evidences in training corpus. Effectiveness of these proposed techniques are not only examined to improve the conventional, hybrid ASR system, but also validated to enhance the state-of-the-art end-to-end ASR systems. Our experiments on real code-switching ASR task show that the proposed methods are very effective to improve the performance of CS speech recognition, and without any performance degradation of the matrix language recognition (Mandarin test set), this is very important for the real ASR applications. Because in real ASR scenarios, a CS ASR system may be used not only for recognizing the Mandarin-English code-switching speech, but also used for monolingual Mandarin speech recognition simultaneously.

The rest of the paper is organized as follows. A review of previous works is presented in Section 2. Section 3 presents the framework of the proposed pronunciation augmentation method. Section 4 describes the details of three speech recognition systems. Experimental configurations are presented in Section 5. Results and performance analysis are presented in Section 6. Finally, we conclude and present future works in Section 7.

2 Review of previous works

The code-switching phenomenon is very natural for people's communication; however, it throws several interesting challenges to the speech recognition community. Three major challenges have been focused in the literature: (i) the heavy sparsity of code-switching training data, especially for the data of intra-sentential code-switched points in both the acoustic and language modeling; (ii) the significant language discrepancy and co-articulation effects in code-mixed utterances, it imposes a big gap between acoustic modeling units of different languages; and (iii) the above mentioned pronunciation variation of embedding language at the code-switches. All of the stages in an ASR system could be significantly affected by any of these challenges, including the acoustic modeling, language modeling, and decoding.

To handle the code-switching data sparsity problem, the most straight forward way is to create code-switching speech corpus. However, for the Mandarin-English CS ASR, only a few small publicly available code-switching corpus can be found, such as the 80 h OC16-CE80 corpus that provided for the Chinese-English mixlingual speech recognition challenge (MixASR-CHEN) [19] and the SEAME corpus [20, 21] with 63 h spontaneous intra-sentential and inter-sentential code-switch speech. The matrix language in both OC16-CE80 and SEAME is Mandarin; the data amount of code-switch speech events in both corpus is extremely sparse. In addition, with the mix-language property and high cost of time and money, it is arduous to create large-scale CS corpus with golden standard manual transcription [22]. Therefore, some works start to focus on developing automatic CS event detection system to extract speech utterances with language switches. For example, in [23], a latent language space model and delta-Bayesian information criterion were proposed to detect the code-switching event. Rallabandi et al. [24] proposed to use an ASR system to detect code-switching style utterance from acoustics. In [25], the authors used frame-level language posteriors generated from a CS ASR system to detect the code-switches. Most of these works were highly dependent on the performances of ASR systems. They may not practical for extracting CS utterances from large real audio archives; therefore, some recent works start to focus on building CS event classifiers based on the deep neural networks directly [26]. In addition, we know that traditional data augmentation techniques, such as the speed or volume perturbation [27, 28], the SpecAugment [29] or audio synthetic [30] have shown their effectiveness to alleviate the data sparsity at some extent in various speech and sound processing tasks, however, these techniques can not handle the pronunciation variation problem of the embedding language in a CS ASR scenario.

Towards the code-switched text data augmentation for language modeling, there are also many works in the literature. Works in [14] used machine translated text to augment the available code-switched text and found that those synthesized CS texts achieved significant reductions in perplexity. And in [31], the authors increased the CS texts by integrating both the syntactic and semantic features into the language modeling process. The recent generative adversarial networks with reinforcement learning was proposed in [32] to create CS text from monolingual sentences. Pratapa et al. [33] proposed to generate grammatically valid artificial CS data using parallel monolingual sentences with linguistic equivalence constraint. In [4], a simple transliteration-based data augmentation approach was proposed to augment the Bengali-English code-switched transcripts. The results showed that transliterating the code-mixed textual corpus to the matrix language and adding it to training data significantly improved the CS ASR performance. All of these previous works showed that the artificial generated CS texts were very effective for alleviating the data sparsity problem in language modeling at some extent. However, from our previous observation of acoustic data augmentation presented in [26], it seems that it is more difficult to generate effective synthesized CS speech than CS text; this may due to the fact that effective real speech are normally with complicated acoustic environment and intra and inter speaker variabilities. These variabilities are very challenging to speech synthesize community.

To alleviate the data sparsity and co-articulation effect problem in code-switching acoustic modeling, previous works mainly focused on (1) exploring an universal acoustic modeling units, (2) developing new acoustic modeling strategies with multi-task or transfer learning, and (3) unsupervised or semi-supervised learning. For improving the CS ASR systems with conventional architectures, such as the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) or the Deep Neural Network-HMM (DNN-HMM)-based hybrid framework, these works mainly focused on mix-language phone mapping and unit merging, such as, in [34–36], the unit merging on state, senone, and Gaussian Levels was proposed for Mandarin-English ASR tasks. In [8, 12, 14], different data-driven and knowledge-based phone merging and clustering algorithms were investigated to get a compact bilingual phone set. For the recent popular end-to-end (E2E) acoustic modeling, new universal acoustic modeling units for CS ASR were proposed to minimize the co-articulation and discrepancy between different languages, such as, in [7, 15], the *Character-Subword* units with Chinese characters for Mandarin and Byte-pair Encoding (BPE) [37] subword for English were constructed. Shan et al. [38] adopted the Mandarin characters plus English letters and wordpieces as its E2E modeling units. For the new strategies of CS ASR, most works aimed at integrating the individual language information to improve the final CS models. For example, two language-specific DFSMN subnets with a shared output layer was proposed to model the CS acoustic information in [15]. And in [7, 16, 38], multi-task joint training of language identification and CS E2E ASR tasks were investigated, and in some works, the transfer learning was also used to provide a good initialization of the E2E encoders using large-scale monolingual corpus. In addition, to exploit the large-scale untranscribed code-switching data, many other efforts have been paid on using the unsupervised and semi-supervised learning for improving the CS acoustic model, such as in [26, 39, 40], etc. All of these previous works have been proved to be effective, either for alleviating the data sparsity or co-articulation effects at some extent. However, they still can not solve the embedding language pronunciation variation problem, because most current hybrid acoustic modeling approaches still highly rely on the monolingual lexicons with standard pronunciations. And moreover, in E2E Mandarin-English ASR scenarios, most sub-words or wordpiece extraction approaches only consider character sequence frequencies instead of acoustics, which at times produce inferior subword segmentation that might lead to erroneous speech recognition output [41].

Therefore, in few latest works, people start to focus on dealing with the pronunciation variation problem in acoustic modeling. For example, to handle the accented pronunciation problem in conventional DNN-based hybrid Mandarin-English code-switching ASR system, [17] proposed to generate native pronunciations representation of embedding language words in the matrix language phoneme set, using a combination of existing acoustic phone decoders and a LSTM-based grapheme-to-phoneme (G2P) model. However, for the popular E2E architectures in CS ASR, we have not found any approach focusing on the pronunciation variation issue in the literature, although in some recent works for monolingual E2E ASR, they have found that the high quality pronunciation lexicons developed by linguists can potentially improve the performance of E2E systems, such as the detail investigations on evaluating the value of pronunciation lexicon in E2E models in [42] and the pronunciation-assisted sub-word modeling method in [41].

In this study, we also focus on dealing with the pronunciation variation problem in the Mandarin-English CS ASR tasks. We aim at achieving better recognition accuracy of the embedding languages phrases without any performance scarify of the matrix language speech, by using pronunciation augmentation techniques. Unlike the work in [17], we not only consider merging the acoustic similarity between mixed languages, but also consider enhancing the discriminative information between different languages. New possible pronunciations for those embedding words at code-switches will be automatically generated. Combined with the expert information and G2P, effectiveness of these new pronunciations are examined in both hybrid and E2E CS ASR systems.

3 The proposed pronunciation augmentation approach

In this section, we present the details of the proposed pronunciation augmentation approach. An English-to-Mandarin (E2M) mix-language phone mapping approach is first proposed to obtain an universal code-switching phone set. Based on this phone set, we further investigate a pronunciation augmentation strategy for embedding language words using acoustic data-driven lexicon learning.

3.1 E2M phone mapping

It is well known that there is big language difference between Mandarin and English; however, the existence of the co-articulation and CS data sparsity problem make it important to use an universal phone set for building a success DNN-HMM based hybrid ASR system. Instead of mapping all the embedding language phones to the matrix language phones as in previous works [8, 17], here we choose to cast light on the balance between the similarities and differences of the two mixed Mandarin and English languages. Only part of phones with high similarity measure are merged together.

Figure 1 illustrates the proposed framework of E2M mix-language phone mapping. In this framework, we combine two effective conventional data-driven phone clustering methods with an expert correction to generate the final universal phone set for Mandarin-English code-switching ASR. Specifically, given two monolingual lexicons and speech corpus, we first obtain a set of English to Mandarin phone mapping pairs Φ_{Tag}, using the Tag model-based phone mapping method that has been proposed in [8]. Then, we perform the TCM phone clustering to obtain another set of phone mapping pairs Φ_{TCM} as proposed in [43]. These two sets of pairs are further combined and merged using following rule:

$$\Phi_{Com} = \{(P_E, P_M) | (P_E, P_M) \in \{\Phi_{Tag} \cap \Phi_{TCM}\}\} \quad (1)$$

where P_E is the English phone, P_M is Mandarin phone, and (P_E, P_M) is the E2M phone mapping pair, Φ_{Com} contains those E2M pairs that lie in both the Φ_{Tag} and Φ_{TCM} sets. These pairs are taken as the high confidence similarity phone mapping pairs, because they derived from two different phone clustering methods.

The principle of the Tag model-based method aims at sharing individual Gaussians across languages. All the Gaussians in the Tag model are clustered in a single, phone-independent, language-independent, Kullback-Leibler divergence-based, Vector Quantization (VQ) code-book. If any two phones in the Tag model have the

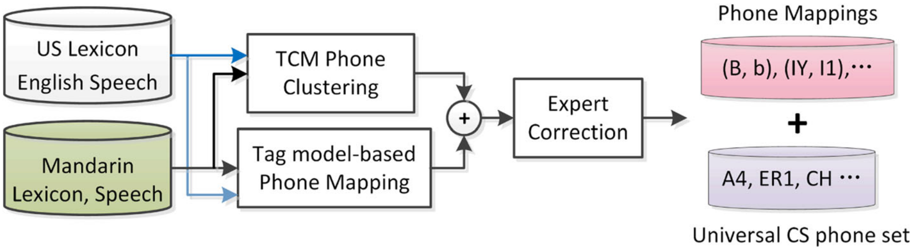
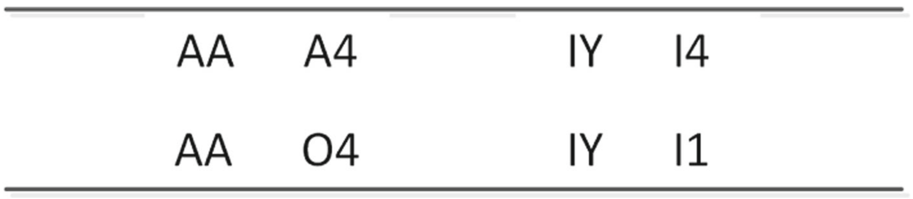


Fig. 1 The proposed E2M mix-language phone mapping framework

majority of their Gaussians lying in common VQ clusters, then these phones are assumed to be similar [8]. On the other side, the TCM-based method in [43] is a two-pass phone clustering method that is based on a co-occurrence confusion matrix. In each pass, Mandarin and English take turns as the source and the target language. The counts of co-occurrence between force-aligned target phone strings and the corresponding source phonetic transcriptions are then arranged to calculate the confusion probability between phone pairs. That is to say, these two methods merged the characteristics of two languages in totally different aspects. Therefore, we hope that comparing and integrating the outputs of two different methods can not only assure the confidence of data-driven phone mapping but also can provide some potential guidance for our further expert correction.

Besides the Φ_{Com}, as shown in Fig. 1, we also add an expert correction stage to improve the overall quality of E2M phone projection. The motivation of this stage is that, in our extensive experiments, we find the Φ_{Com} contains most of non-vowel phone mappings, but with only few vowel phone-mappings, most vowel phone-mappings produced by the TCM and Tag-model based methods are very different. Therefore, we invite three linguistic experts from Shanghai Normal University to perform the vowel phone mapping correction. In this stage, all pairs in Φ_{Com} are directly fed into the final phone mapping set Φ_{Final}; only other vowel phone mapping pairs in Φ_{Tag} and Φ_{TCM} are then checked and corrected by linguists. The majority voting rule (2/3) is used to measure the reliability of experts' correction. This correction process not only dependent on the linguistic knowledge of experts, but also guided by the statistics of confusion matrices achieved from the TCM and Tag model-based phone clustering processes. It worth noting that we do not perform any mapping for those pairs with low similarity measurements (e.g., the `DH`, `ZH` in CMU English lexicon); keeping these language-dependent phones may help to integrate the large acoustic and language discriminative information during acoustic modeling. Finally, all of the corrected phone mapping pairs, the pairs in Φ_{Com}, and the few language-dependent English phones are combined to produce the final universal CS phone set, perform the English lexicon mapping, and obtain the final universal CS lexicon.

In addition, according to the experts' knowledge and different outputs of the above two phone clustering methods, we speculate that it may be useful to perform a one-to-two or one-to-many mapping for those English vowels with a similarity measure larger than a threshold. This may help to efficiently handle the Mandarin accents in English words because all the mappings from both phone clustering methods are data-driven results. A mapping pair has a very high confusion probability may still indicates a similar acoustic characteristics, even it is not lie in Φ_{Com} set. Therefore, in this study, besides the one-to-one mapping, the one-to-two mapping cases with a further expert correction are also investigated. Two one-to-two mapping examples for English phones `AA` and `IY` are illustrated in Fig. 2.



AA	A4	IY	I4
AA	O4	IY	I1

Fig. 2 One-to-two phone mapping cases

3.2 Pronunciation augmentation using Lexicon learning

The idea of our pronunciation augmentation approach is motivated by the algorithm of acoustic data-driven lexicon learning in [18]. In [18], this algorithm was proposed to automatic generate pronunciations for the OOV words in monolingual English ASR task. However, in this study, we borrow the lexicon learning idea to handle the pronunciation variation problem in Mandarin-English code-switching ASR. It is developed to generate informative new pronunciations only for the embedding language (English) words. These new pronunciations are then appended to the universal CS lexicon for ASR acoustic modeling.

In fact, it is also possible for the system to produce new pronunciations for Mandarin words using lexicon learning. However, in our Mandarin-English code-switching tasks, the Mandarin is the matrix language, while English is the embedding language, all the Mandarin words are spoken by the native speakers, and there is no pronunciation variation for a native speaker to say native speech. Therefore, we only considering the new pronunciations of English words and ignore those ones produced for the Mandarin words.

Figure 3 presents the whole framework of the proposed pronunciation augmentation using lexicon learning. It includes three main modules: (a) the CS lexicon preparation for all the words in acoustic training data, (b) the new pronunciation candidates collection, and (c) the

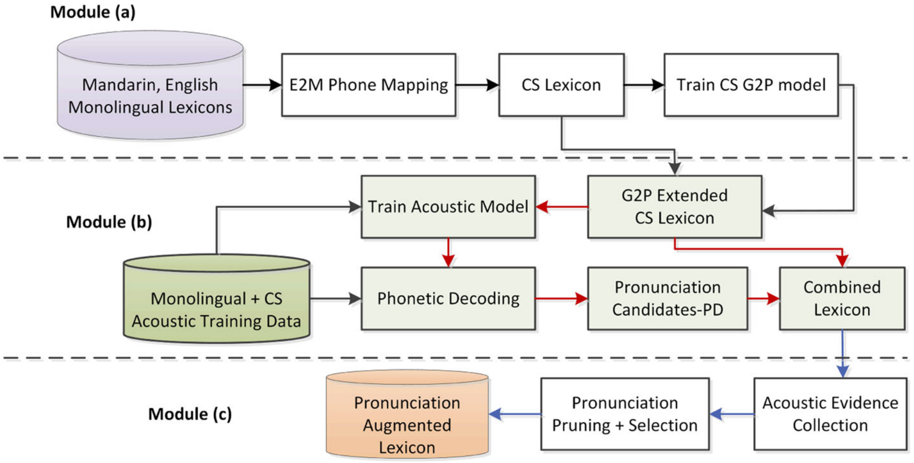

Module (a)
Mandarin, English Monolingual Lexicons
E2M Phone Mapping
CS Lexicon
Train CS G2P model
Module (b)
Train Acoustic Model
G2P Extended CS Lexicon
Monolingual + CS Acoustic Training Data
Phonetic Decoding
Pronunciation Candidates-PD
Combined Lexicon
Module (c)
Pronunciation Augmented Lexicon
Pronunciation Pruning + Selection
Acoustic Evidence Collection

Fig. 3 The proposed pronunciation augmentation framework of embedding language words in CS ASR using lexicon learning

pronunciation pruning and selection. This framework is very similar to the one proposed in [18], however, there is many implementation details that are specially designed for the focused code-switching speech recognition task.

In module (a), we first create a Mandarin-English CS lexicon by mapping the standard monolingual English lexicon using those phone mappings obtained in Section 3.1. Then, unlike the small seed lexicon in [18], here we extend the whole CS lexicon by training a Sequitur G2P [44] model on it to produce initial pronunciation only for OOV words in the training data. Based on the G2P extended lexicon, in module (b), we train an acoustic model using all of the monolingual Mandarin, English, and code-switching speech data to perform the training data forced-alignment, build the phone language model, and further construct a phonetic decoder. This decoder is then used to generate the phonetic transcription for each specific word w exist in the training data. Finally, for each individual word, we can obtain many new pronunciations candidates (PD) generated from the phonetic decoding, by aligning the phone sequence of forced-alignment and phonetic transcription using a normalized relative frequency measurement. These PDs can be combined with the G2P extended lexicon into a large CS lexicon (Combined Lexicon) for the next acoustic evidence collection in module (c). The "acoustic evidence" is defined as $\tau \triangleq p(\mathcal{O}_u|w, b)$; it is the acoustic conditional data likelihood of utterance $\mathcal{O}_u$, given the pronunciation of word w being b. This "acoustic evidence" is derived from the per-utterance lattice pronunciation-posterior statistics, and these statistics are computed using lattices of training utterances that are produced based on the `Combined Lexicon` and existing acoustic model in module (b). Given a set of pronunciation candidates for a specific word w, and the acoustic evidence τ per utterance, the pronunciation pruning and selection are performed using an iterative greedy pronunciation selection (IGPS) procedure with a per-utterance likelihood reduction criterion. Finally, with this procedure, all the least important pronunciation candidates will be iteratively removed in an efficient greedy fashion. All the details about lexicon learning algorithm and other implementation tricks, please refer to the work [18].

Furthermore, unlike the motivation to generate pronunciations for OOV words in [18], the lexicon learning idea in this study just plays a pronunciation augmentation role. Therefore, based on the greedy pronunciation selection, we added an additional PD selection constraint to assure a higher quality of new pronunciations as below:

$$PD(w)_\tau \geqq \rho \; Avg.R(w)_\tau \tag{2}$$

where $PD(w)_\tau$ is the "acoustic evidence" soft counts of pronunciation for word w derived from phonetic decoding after the last iteration of IGPS. $Avg.R(w)_\tau$ is the average soft counts of word pronunciations in the reference source lexicon (G2P Extended CS Lexicon), and $\rho \in [0, 1]$ is the statistical pruning factor.

It worth noting that only the training utterances contain English words are taken into account during the acoustic evidence collection and PD pruning stages, because in our CS task, we only expect to augment the pronunciations of English words for the heavy pronunciation variation issue. After the greedy process of PD pruning, only the informative subset of PDs for each word with acoustic evidence is selected. These informative pronunciations are then used to augment the source CS lexicon for acoustic modeling. For those words in target vocabulary that are not seen in the acoustic training data, or no pronunciation produced during lexicon learning, we choose to generate their pronunciations by re-training a CS G2P model using the already augmented lexicon, instead of the initial G2P pronunciation candidates in module (a).

4 Speech recognition systems

Three types of ASR system are used to evaluate the proposed method. They are the conventional GMM-HMM system, the state-of-the-art lattice-free maximum mutual information (LF-MMI)-based hybrid system and the latest Transformer-based E2E system.

4.1 GMM-HMM system

Our GMM-HMM systems are built using the open source Kaldi speech recognition toolkit [45]. The 13-dimensional mel-frequency cepstral coefficients (MFCC) plus one-dimensional pitch with their first- and second-order differential coefficients are used as the input acoustic features to train the initial GMM-HMM acoustic models. Based on the initial model, all of the acoustic features are then spliced over 9 frames and projected to 40-dimensional subspace using the linear discriminant analysis (LDA). A further maximum likelihood linear transform (MLLT) is applied to transform the projected features for a better orthogonality. These transformed features are then used to refine the GMM-HMM parameters. After the decision tree clustering, the final models have around 6000 context-dependent tied states with around 32 Gaussians per state (different lexicon leads to different number of states). Based on this LDA+MLLT model, we further perform the speaker adaptive training (SAT) with constrained maximum likelihood linear regression to adapt the Gaussian mixture model parameters. After adapting the parameters, a re-alignment is performed to improve the LDA+MLLT+SAT system. The framework and implementation details of our GMM-HMM system training are followed the example recipe `egs/swbd/s5c` in Kaldi main branch.

4.2 LF-MMI based hybrid system

The LF-MMI based ASR hybrid framework was first proposed in [46]. Because of its good performances and excellent generalization ability, it has been becoming a mainstream technology for speech recognition, either in industry or in academic community. The LF-MMI based hybrid acoustic model is a time-delay neural network (TDNN) with multi-splice sub-sampling topology. Povey et al. [46] proposed to train it in a purely sequence-discriminative way using the lattice-free version of the MMI criterion. Compared with the classical TDNN training with cross-entropy criterion, three major modifications have been introduced to the LF-MMI TDNN training:

- Training from scratch without initialization from a cross entropy system.
- The use of a threefold reduced frame rate and a simpler HMM topology.
- Limiting the range of time frames where supervision labels can appear by using finite state acceptors (FSA).

In addition, unlike the denominator lattices in classical MMI, the lattices in the LF-MMI architecture are first generated from a phone-level n-gram language model, and then compiled into utterance-specific FSA graphs for TDNN training. Furthermore, to avoid over-fitting during training, the cross-entropy objective function as well as the leaky HMM are also applied as extra regularization techniques in this architecture. In this study, we choose to use a TDNN-LSTM hybrid structure presented in [47] as our acoustic model because of its better performances.

4.3 Transformer-based E2E system

With the great success of no-recurrence sequence-to-sequence model-Transformer proposed in machine translation [48], more and more research works in speech community start to focus on it. Recently, a Speech-Transformer [49] was successfully proposed by introducing the Transformer to ASR task. With the encoder-decoder architecture and multi-head self-attention mechanism to learn the context and positional dependencies, the Transformer has proven to be very successful to achieve competitive speech recognition performances, and it has already become the state-of-the-art E2E ASR system.

Compared to the above LF-MMI based hybrid system that consist of separate pronunciation, acoustic, and language models, the Transformer-based ASR system is a single neural-network which implicitly models all three. Due to lack of pronunciation lexicon, most E2E systems choose to model the output text sequence in finer units instead of the whole words, such as the characters, BPE subwords, and wordpieces. Therefore, in this study, we also investigate how to improve a Transformer-based E2E system for Mandarin-English code-switching ASR, by introducing the augmented CS pronunciations to assist the BPE subword modeling. This is motivated by the recent work of pronunciation-assisted subword modeling (PASM) proposed in [41]. We hope the PASM can generate linguistically meaningful subwords for the embedding language English by analyzing the training text corpus and our augmented CS lexicon.

The recipe of PASM word segmentation[1] was used in our experiments. All the experiments of E2E ASR system building were performed using the open-source end-to-end speech recognition toolkit ESPnet [50].

5 Experimental configurations

5.1 Datasets

The corpus used to build our code-switching ASR systems is provided by Unisound Corporation[2], including 186 hours (hrs) Mandarin-English code-switching speech,

[1] https://github.com/hainan-xv/PASM.git
[2] https://www.unisound.com/

500 hrs Mandarin, and 100 hrs English (with Chinese accent) monolingual speech. We use "Chilish" to term this accented English set. All of the utterances are conversational speech or speech collected from voice search, and the speakers are all from the mainland of China. In our training set, there is a heavy imbalance between the data amount of Mandarin and Chilish, because in real applications, it is much harder to collect Chilish speech than Mandarin.

As our goal is to improve the ASR performance of the embedding language without any performance scarify of the matrix language, we designed three test sets for performance evaluation, one is a 3 hrs pure Mandarin speech test set (Mandarin), one is 3.6 hrs Mandarin-English code-switching test set (1.6 hrs are from voice search, 2.0 hrs are general conversational speech), and the third one is a pure 1.6 hrs Chilish test set.

5.2 Neural network structures and acoustic features

The experimental configurations of GMM-HMM systems have been already presented in Section 4.1. Unlike the typical TDNN, the LF-MMI based TDNN-LSTM hybrid system is a mixture architecture of LSTMPs and sub-sampled TDNNs, using 3 fast-LSTMP layers interleaved with 7 spliced TDNN layers. More details of this architecture can refer to the TDNN-LSTMP structure used for SWBD corpus in [47]. The frame-level alignments and lattices in TDNN-LSTM model training were directly generated from the GMM-HMM system. The recipe of `swbd/s5c/local/chain/run_tdnn_lstm.sh` in Kaldi repository is used for our hybrid model training, but without any i-vectors or other speaker adaptation techniques. The language model used in both GMM-HMM and hybrid system is the same trigram LM that built on all of the training data texts.

As most ASR example recipes in ESPnet, for the Transformer-based E2E systems, we use 12 Encoder and 6 Decoder blocks with 2048 feed-forward inner dimension. The model dimension d_{model} is set to 256 and the attention head number h was set to 4. Both the hybrid and E2E acoustic models use the same 80-dimensional filter-bank features, plus 3-dimensional pitch (pitch and its first and second derivatives). All of the input acoustic features are extracted using a 25-ms Hamming window with a 10-ms frame shift. In addition, to enhance the model robustness, we perform both the `SpecAugument` and `speed perturbation` acoustic data augmentation during all the hybrid and E2E system training.

5.3 Lexicon and performance measure

The monolingual Mandarin lexicon (MLex) used in this study is an ARPAbet-based tonal lexicon provided by Unisound Corporation. It has 109 phoneme/toneme phone set, covering more than 200k Mandarin words. The monolingual English lexicon is the CMU open source English dictionary with 39 phonemes[3]. In the experiments, these two monolingual lexicons are first combined and mapped into an universal CS lexicon, and then further augmented using the proposed E2M mix-lingual phone mapping and lexicon learning approaches to handle pronunciation variation problem in the code-switching ASR tasks. For the system performance measure, we use token error rate (TER), where the "token" refers to the unit of Mandarin character and English word respectively.

Table 1 TER% results on pure Chilish test set of GMM-HMM systems

Training data	Chilish
LibriSpeech (460hrs)	57.3
Chilish (100hrs)	25.9

6 Results and discussions

6.1 Acoustic difference between native and non-native English speech

It is well known that there is big acoustic difference between native English and non-native English speech. In Table 1, we also performed an experiment to validate this difference. In this experiment, the 1.6 hrs Chilish test set is taken for evaluation. The 100 hrs Chilish training data is used to train the GMM-HMM model for non-native monolingual English. The native English model is directly taken from the Kaldi LibriSpeech repository on LibriSpeech.

Both systems in Table 1 are trained using the CMU lexicon. From the big TER gap on Chilish test set, it is clear to observe the significant acoustic difference between native English and Chilish, even the native acoustic model is trained on a larger corpus. However, in the literature, most ASR systems related to English are still using the CMU pronunciations which are specially designed for the native English. This indicates that there might be chance to improve the ASR performance by integrating the acoustic variations in the lexicon pronunciations.

6.2 Evaluating E2M phone mapping in conventional ASR systems

Table 2 compares the effectiveness of the proposed E2M mix-lingual phone mapping with different conventional phone mapping methods. The first two systems with only MLex and CMU lexicon are monolingual Mandarin and English system respectively; they are only separately trained using the 500 hrs Mandarin and 100 hrs Chilish training data. Other systems in this table are all coder-switching GMM-HMM systems and trained on the total 786 hrs training data. The "CMU+MLex" represents directly concatenating the phone sets and lexicons

[3] CMU lexicon:http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict

Table 2 Performance comparison(TER%) on and code-switching test sets from GMM-HMM systems with different phone mapping methods

Phone mapping method	Mandarin	Code-switching	Chilish
Only MLex	18.2	–	–
Only CMU	–	–	25.9
CMU+MLex(baseline)	17.6	32.3	26.7
TCM [43]	17.9	25.6	27.2
Tag model-based [8]	17.5	25.2	27.0
E2M-po2o	17.5	23.7	27.0

of CMU and MLex without any phone sharing or mapping. The "TCM" and "Tag model-based" are the conventional data-driven phone clustering methods proposed in [43] and [8] respectively. The "E2M-po2o" is our proposed E2M phone mapping, in which only part of English phonemes are involved in the English to Mandarin phone mapping, and in this case, they are just an one-to-one mapping.

From TERs in Table 2, three observations can be found: (a) phone-mapping is very effective to improve the code-switching ASR performances, either by using the TCM, Tag model-based phone clustering or the proposed E2M. Both conventional phone clustering methods achieved more than a relative 20% TER reductions over baseline, a further 6% TER reduction is obtained by the proposed E2M. In the proposed E2M framework, this further gain may benefit from the tradeoff mechanism between language-independent acoustic commonness and language-dependent characteristics. (b) More Chilish words are mis-recognized by the universal code-switching acoustic model than the pure English model; it indicates that phone mapping brings acoustic confusion between Chilish and Mandarin than monolingual ASR. (c) On the Mandarin test set, all of the CS models achieve better performances than the system trained on pure Mandarin data. Relative 1.6–3.8% TER reductions has been obtained. This is due to the increased Mandarin data included in the CS training set. In addition, we can see that there is almost no performance degradation on the Mandarin test set, by using phone mapping or clustering methods over the baseline.

Figure 4 demonstrates the TERs calculated separately on the Mandarin and English part of the code-switching test set. "CS" is the whole code-switching test set as shown in Table 2, "CS-Mandarin" and "CS-English" represent the Mandarin part and English part of CS. Compared with the CMU+MLex baseline, both the performances of Mandarin and English speech are improved by the proposed E2M. And moreover, it is clear to see that the performance gain on English part is much larger than the one on Mandarin part. This indicates that the Chinese accent characteristics of embedding English words are well learned and the acoustic events of code-switches are heavily increased by the proposed E2M; the recognition error within code-switches are significantly reduced.

Table 3 shows the performance comparison of the state-of-the-art LF-MMI based hybrid systems with different E2M phone mapping strategies. From the TER numbers in Tables 2 and 3, we can see that these hybrid systems achieve significant ASR performance gains (more than

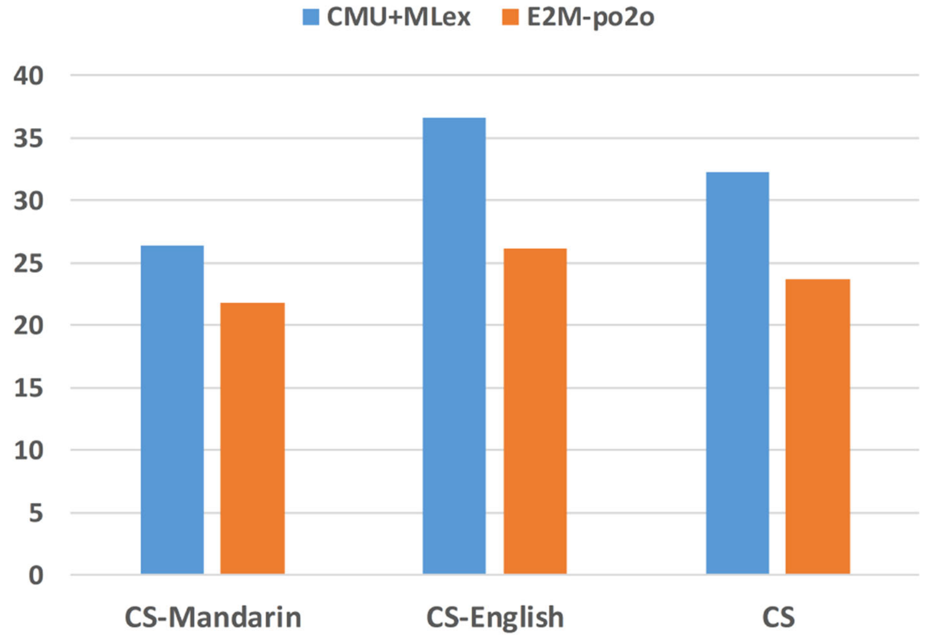


Fig. 4 TER% between Mandarin and English part of code-switching test set with different phone mapping methods

Table 3 Performance (TER%) comparison of the proposed E2M in LF-MMI based hybrid systems

Phone mapping method	Mandarin	Code-switching	Chilish
CMU+MLex (baseline)	5.5	12.9	14.9
E2M-po2o	5.2	11.5	15.4
E2M-po2m	5.8	12.2	15.9
E2M-po2o-T, E2M-po2m-D	5.4	10.8	15.4

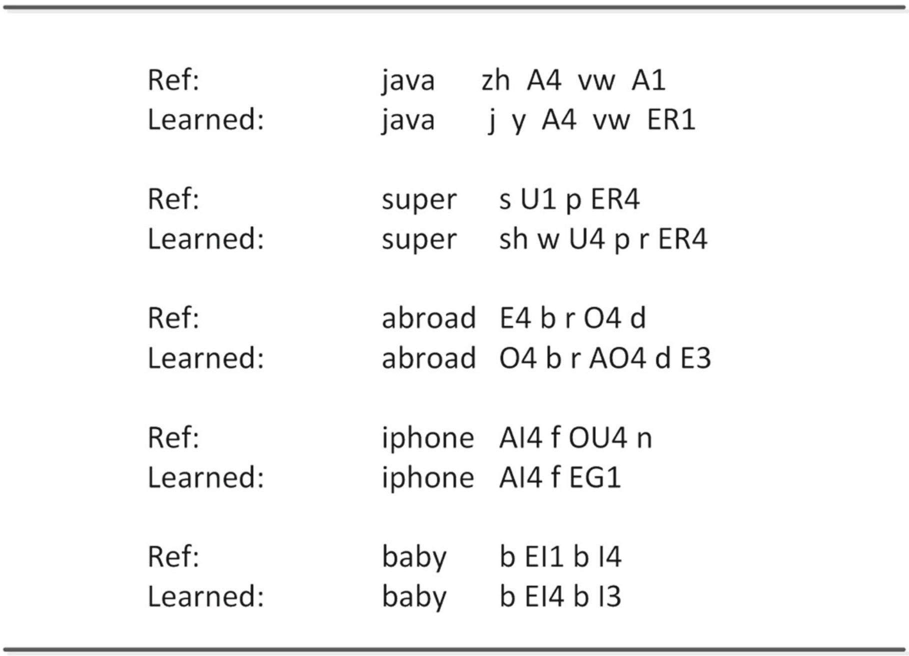
relative 45%) over the conventional GMM-HMM systems. The "E2M-po2m" is the proposed E2M phone mapping with part of English phones ($\sim$ 38%) are mapped to two or three Mandarin phones. "E2M-po2o-T, E2M-po2m-D" means we use the "E2M-po2o" for acoustic model training, while "E2M-po2m" is used for decoding. Except for the last line of Table 3, all other systems use the same lexicon for both the acoustic model training and decoding.

Compared with the large gains obtained from phone mapping in GMM-HMM systems, the E2M-po2o only obtains relative 10.8% TER reduction on code-switching test set over the CMU+MLex baseline. In addition, by comparing the results on code-switching test set of last two lines in Table 3, it is interesting to find that the one-to-many mapping in E2M brings more mix-lingual acoustic confusion than the one-to-one phone mapping. And given the fixed universal phone set, only using proper one-to-many phone mapping in the decoding stage can bring further 6.1% TER reduction over the E2M-po2o method. This may due to the fact that more pronunciation entries provided more possible competitive candidates in the WFST paths. Moreover, it is clear to see that the recognition of pure Mandarin speech does not significantly affected by different phone mapping methods, while on the pure Chilish test set, a relative 6.7% performance degradation is obtained. This indicates that the acoustic model with a Mandarin dominant phone set and training data is not suitable for a pure English ASR task. In addition, the TER gains on code-switching test set indicates that the training data of acoustic modeling bi-phone units for code-switches is also significantly increased in the hybrid acoustic modeling through the phone mapping.

6.3 Examination of Lexicon learning-based pronunciation augmentation

6.3.1 Performances in conventional ASR systems

After achieving the universal E2M-po2o mix-lingual lexicon, we then augment the pronunciations of English words in training data set using the proposed lexicon learning-based framework in Fig. 3. The recipe of `wsj/s5/steps/dict/learn_lexicon_greedy.sh` in Kaldi main branch is modified to perform our CS ASR lexicon learning. During the iterative greedy pronunciation selection, we tune the scaling factor $\alpha = 0.05, 0.01, 0.001$ and smoothing factor $\beta = 10, 5, 10$ to compute the likelihood reduction threshold for controlling the pruning degree of pronunciations from the phonetic decoding, G2P, and source lexicon respectively. Figure 5 demonstrates new pronunciation



Ref:	java	zh A4 vw A1
Learned:	java	j y A4 vw ER1
Ref:	super	s U1 p ER4
Learned:	super	sh w U4 p r ER4
Ref:	abroad	E4 b r O4 d
Learned:	abroad	O4 b r AO4 d E3
Ref:	iphone	AI4 f OU4 n
Learned:	iphone	AI4 f EG1
Ref:	baby	b EI1 b I4
Learned:	baby	b EI4 b I3

Fig. 5 Samples of learned pronunciations for embedding English words. "Ref" and "Learned" is the word-pronunciation pair in the source ("G2P Extended CS Lexicon") and learned lexicon in Fig. 3 respectively

Table 4 Performance(TER%) comparison of the proposed lexicon learning-based pronunciation augmentation on conventional ASR systems

System	ρ	Lexicon(avg.#prons)	Mandarin	Code-switching	Chilish
GMM-HMM	–	Source Lexicon	17.5	23.7	27.0
	0.1	+ LexLearn (3.65)	17.8	22.8	26.4
	0.4	+ LexLearn (3.07)	17.6	22.1	26.2
	0.7	+ LexLearn (2.41)	17.6	21.6	26.0
LF-MMI hybrid	–	Source Lexicon	5.2	11.5	15.4
LF-MMI hybrid-1	0.7	+ LexLearn (2.41)	5.7	10.9	15.2
LF-MMI hybrid-2	0.7	+ LexLearn (2.41)	5.5	10.4	14.8

"avg.#prons" means the average pronunciations per English word included in the training data. ρ is the pruning factor of acoustic soft counts in Eq.(2)

samples learned for five embedding English words. It is clear to see that the learned pronunciations based on the acoustic evidences tend to be more Chilish than the ones from only one-to-one E2M mapping. It may provide a chance for acoustic model to capture the accent and mis-pronounced property of the embedding English words by augmenting the lexicon with these learned pronunciations.

Table 4 presents the performance comparison of the proposed lexicon learning-based pronunciation augmentation on conventional ASR systems. The "Source Lexicon" is the universal CS lexicon with E2M-po2o phone mapping. "+ Lexlearn" means the "Source Lexicon" augmented with the new pronunciations of English words learned from the proposed CS lexicon learning framework. The same acoustic modeling and decoding lexicon is used for each GMM-HMM and "LF-MMI hybrid-1" system as shown in this table. However, the "LF-MMI hybrid-2" system used different lexicons with E2M-po2o and E2M-po2m phone mapping respectively for the acoustic modeling and decoding, and both lexicons are also augmented with the same new pronunciations as in "LF-MMI hybrid-1" system. System "LF-MMI hybrid" is the same as system with "E2M-po2o" in Table 3.

From the results of Table 4, we can see that by using the augmented CS lexicon, the GMM-HMM systems can obtain a relative 3.8 to 8.8%, and 2.2 to 3.7% TER reduction on the code-switching and Chilish test set respectively. By taking the source pronunciations as reference, introducing the acoustic soft-count pruning factor can effectively help to select better new pronunciations with enough acoustic evidences. Based on the outputs of standard iterative greedy pronunciation selection, we find that the $\rho = 0.7$ achieves the best results. Furthermore, by comparing the results of hybrid systems in Table 4 with their baselines in Table 3, we still can obtain relative 5.2% (hybrid-1) and 3.7% (hybrid-2) TER reduction on the code-switching test set by using the augmented lexicon, and consistently, the TER of Chilish is also reduced from 15.4 to 14.8%. However, it is clear to see that the gains obtained on LF-MMI based hybrid systems are much smaller than the ones obtained on GMM-HMM systems; it indicates that the hybrid system has a better acoustic modeling ability to deal with the pronunciation variations of embedding language than the traditional GMM-HMM system.

6.3.2 Performances in transformer-based ASR systems

Table 5 compares the results with different target acoustic modeling units in our E2E Transformer-based code-switching ASR systems. The targets of our baseline system are a set of Mandarin characters and English letters plus blank symbol which leads to an output dimension of 5257. In addition, we also tried to adopt the widely used BPE subword segmentation to generate 2000 subwords as acoustic modeling units for English. Therefore, in the system with "Character-BPE," there is a total of 7230 acoustic modeling targets. The systems with "Character-PASM(*)" modeling units are performed to examine the effects of our augmented CS lexicon for a better English subword generation. In these system, the Mandarin characters units are the same as in the baseline, but the English targets are produced from the universal CS lexicons using the pronunciation-assisted subword modeling (PASM) method. Three CS lexicons are investigated, the basic "CMU+MLex," the lexicon with E2M-po2m phone mapping, and the final augmented CS lexicon with acoustic lexicon learning ("Source Lexicon+Lexlearn (2.41)" in

Table 5 Results (TER%) of Transformer-based E2E CS ASR systems with different target acoustic modeling units

Modeling units	Mandarin	Code-switching	Chilish
Baseline (character-letter)	5.2	11.6	15.9
Character-BPE	5.3	11.1	15.3
Character-PASM (CMU+MLex)	5.2	10.7	15.3
Character-PASM (E2M-po2m)	5.1	10.5	15.1
Character-PASM (Source Lexicon+Lexlearn (2.41))	5.2	10.1	14.8

Table 4). For a clear comparison, we keep the number of subword units to be the same in BPE and PASMs.

By comparing the E2E results in Table 5 with the best result of LF-MMI hybrid systems in Table 4, we can see a little bit performance improving on pure Mandarin test set while the performance of code-switching and Chilish speech are significantly degraded, such as for CS test set, the TER is degraded from 10.4 to 11.6%. However, by using the BPE subwords instead of simple letters as English E2E acoustic modeling targets, a relative 4.3% and 3.8% TER gains are obtained on the code-switching and Chilish test set respectively. This indicates that the BPE subwords can provide a better acoustic representation than simple letters. When we use the PASM to produce English targets instead of the BPE, the TER of CS is reduced from 11.1 to 10.7% with only CMU pronunciations for English. This observation is consistent with the one achieved in [41]. And it indicates that unlike the BPE only taking the spelling into consideration, leveraging the pronunciation information of a word during the subword segmentation can produce better E2E acoustic modeling units. Furthermore, when we use the phone mapped universal CS lexicon, a further 1.8% relative TER reduction on CS is obtained. And in addition, by comparing the results on both code-switching and Chilish test sets between last two lines of Table 5, the E2E CS ASR system can also benefit from the learned new pronunciations. These observations indicate that the learned and phone mapped new pronunciations provided more phonetically meaningful subwords for the embedding English words.

7 Conclusion

This paper presents a pronunciation augmentation framework based on the universal Mandarin-English code-switching lexicon. This framework is proposed to handle the accented pronunciations or randomly mispronounced words of the embedding English in the code-switching speech recognition task. We first examine the proposed English-to-Mandarin phone mapping on both the conventional GMM-HMM and the state-of-the-art LF-MMI-based hybrid ASR systems. Experimental results show that we obtain more than relative 10% TER reduction on the code-switching test set, by using the universal CS lexicon with the proposed phone mapping strategy. In addition, from the comparison results on the GMM-HMM systems, we can see that this strategy provided much more ASR performance gains than two conventional phone mapping methods in the literature, by considering the balance of acoustic similarity and language discrimination between Mandarin and English.

Furthermore, based on the universal CS lexicon, we validate the proposed pronunciation augmentation framework from two main aspects. One is directly using the augmented pronunciations to train conventional GMM-HMM and LF-MMI hybrid systems. The other is using the augmented pronunciations to assist the subword segmentation to generate better acoustic modeling targets for end-to-end Transformer-based ASR system. The extensive results in Section 6.3.2 show that both the proposed phone mapping and pronunciation augmentation framework can also be taken as effective solutions for improving the E2E CS ASR performances.

More investigations about proposing new pronunciation-assisted methods to fully exploit the phonetically meaningful information for improving E2E CS ASR system and validating the proposed methods on other CS corpus will be our future works.

Acknowledgements

The authors thank the associate editor and the anonymous reviewers for their constructive comments and useful suggestions

Authors' contributions

The authors' contributions are equal. All authors read and approved the final manuscript.

Authors' information

Yanhua Long received her Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2011. From July 2008 to February 2009, she studied as a visiting student in the department of Human Language Technology at the Institute for Infocomm Research, Singapore. From September 2009 to February 2010, she has been an intern worked in the speech group of Microsoft Research Asia. From October 2011 to May 2013, she worked with the Machine Intelligence Lab as a Research Associate in Cambridge University, UK. Since June 2013, she has been with Shanghai Normal University, Shanghai, China, as an Associate Professor, Her research interests include speech signal processing and pattern recognition. She has published about 40 papers and she is a ISCA member.

Shuang Wei received her B.S. degree and the M.S. degree from the Huazhong University of Science and Technology, China, in 2005 and 2007 respectively. And she received the Ph.D. degree in electrical and computer engineering from the University of Calgary, Canada, in 2011. Her current research interests include speech signal processing, signal estimation and detection, computational intelligence, and compressed sensing.

Jie Lian is an Assistant Professor in the Computer Science department at Shanghai Normal University where she has been a faculty member since 2017, and she obtained the "Sailing" Talent Program of China in 2019. Jie Lian completed her doctor degree at Towson University in 2017. Her research interests lie in the area of spatio-temporal data mining, deep learning and big data, ranging from theory to design to implementation.

Yijie Li received his M.S. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2009. From July 2009 to Jan 2012, he joined Shanda Innovations as an Associate Researcher. Since February 2012, he has been with Unisound AI Technology Co., Ltd. Beijing, China, as a Senior Researcher.

Funding

This research was supported by the National Natural Science Foundation of China under grant No.62071302 and No. 61701306.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]SHNU-Unisound Joint Laboratory of Natural Human-Computer Interaction, Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai, China. [2]Unisound AI Technology Co., Ltd., Beijing, China.

Received: 10 May 2021 Accepted: 18 August 2021
Published online: 30 August 2021

References

1. D. Sankoff, S. Poplack, A formal grammar for code-switching. Res. Lang. Soc. Interact. **14**, 3–45 (1981)
2. L. Milroy, P. Muysken, *One speaker, two languages: cross-disciplinary perspectives on code-switching*. (Cambridge University Press, New York, 1995)
3. P. Auer, *Code-switching in conversation: language, interaction and identity*. (Routledge, New York, 2013)
4. M. Ma, B. Ramabhadran, J. Emond, A. Rosenberg, F. Biadsy, in *Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Comparison of data augmentation and adaptation strategies for code-switched automatic speech recognition, (Brighton, 2019), pp. 6081–6085. https://doi.org/10.1109/icassp.2019.8682824
5. D. C. Li, Cantonese-english code-switching research in Hong Kong: a Y2K review. World Englishes. **19**(3), 305–322 (2000)
6. Y. Khassanov, H. Xu, V. Pham, Z. Zeng, E. S. Chng, C. Ni, B. Ma, in *The 20th Annual Conference of the International Speech Communication Association(Interspeech)*. Constrained output embeddings for end-to-end code-switching speech recognition with only monolingual data (ISCA, Graz, 2019), pp. 2160–2164
7. Z. Zeng, Y. Khassanov, V. Pham, H. Xu, E. S. Chng, H. Li, in *The 20th Annual Conference of the International Speech Communication Association(Interspeech)*. On the end-to-end solution to Mandarin-English code-switching speech recognition (ISCA, Graz, 2019), pp. 2165–2169. https://doi.org/10.21437/interspeech.2019-1429
8. H. Chang, Y. H. Sung, B. Strope, F. Beaufays, in *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Recognizing english queries in Mandarin voice search, (Praque, 2011), pp. 5016–5019. https://doi.org/10.1109/icassp.2011.5947483
9. C. Baker, *Foundations of bilingual education and bilingualism*. (Multilingual Matters, Tonawanda, 2011)
10. G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition. Sig. Process. Mag. **29**, 82–97 (2012)
11. D. Yu, J. Li, Recent progresses in deep learning based acoustic models. IEEE/CAA J. Autom. Sin. **4**(3), 396–409 (2017)
12. S. Yu, S. Zhang, B. Xu, in *Proceedings of the 29th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Chinese-English bilingual phone modeling for cross-language speech recognition, (Montreal, 2004), pp. 917–20. https://doi.org/10.1109/icassp.2004.1326136
13. H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, C.-H. Lee, in *Proceedings of the 34th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. A study on multilingual acoustic modeling for large vocabulary ASR, (Taipei, 2009), pp. 4333–4336. https://doi.org/10.1109/icassp.2009.4960588
14. N. Vu, D. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E. Chng, T. Schultz, H. Li, in *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. A first speech recognition system for Mandarin-English code-switch conversational speech, (Kyoto, 2012), pp. 4889–4892. https://doi.org/10.1109/icassp.2012.6289015
15. S. Zhang, Y. Liu, M. Lei, B. Ma, L. Xie, in *The 20th Annual Conference of the International Speech Communication Association(Interspeech)*. Towards language-universal Mandarin-English speech recognition, (Graz, 2019), pp. 2170–2174. https://doi.org/10.21437/interspeech.2019-1365
16. K. Li, J. Li, G. Ye, R. Zhao, Y. Gong, in *Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Towards code-switching ASR for end-to-end CTC models, (Brighton, 2019), pp. 6076–6080. https://doi.org/10.1109/icassp.2019.8683223
17. Z. Huang, X. Zhuang, D. Liu, X. Xiao, Y. Zhang, S. M. Siniscalchi, in *Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Exploring retraining-free speech recognition for intra-sentential code-switching, (Brighton, 2019), pp. 6066–6070. https://doi.org/10.1109/icassp.2019.8682478
18. X. Zhang, V. Manohar, D. Povey, S. Khudanpur, in *The 18th Annual Conference of the International Speech Communication Association(Interspeech)*. Acoustic data-driven lexicon learning based on a greedy pronunciation selection framework, (Stockholm, 2017), pp. 2541–2545. https://doi.org/10.21437/interspeech.2017-588
19. D. Wang, D. Tang, Z. Tang, Q. Chen, in *The 19th Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. OC16-CE80: a Chinese-English mixlingual database and a speech recognition baseline, (Bali, 2016), pp. 84–88. https://doi.org/10.1109/icsda.2016.7918989
20. D. Lyu, T. Tan, E. Chng, H. Li, in *The 11th Annual Conference of the International Speech Communication Association (Interspeech)*. SEAME a Mandarin-English code-switching speech corpus in south-east Asia (ISCA, Makuhari, 2010), pp. 1986–1989
21. G. Lee, T.-N. Ho, E.-S. Chng, H. Li, in *The 21th International Conference on Asian Language Processing (IALP)*. A review of the Mandarin-English code-switching corpus: SEAME, (Singapore, 2017), pp. 210–213
22. M. Sperber, G. Neubig, J. Niehues, S. Nakamura, A. Waibel, Transcribing against time. Speech Commun. **93**, 20–30 (2017)
23. C. Wu, H. Shen, C. Hsu, Code-switching event detection by using a latent language space model and the delta-Bayesian information criterion. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(11), 1892–1903 (2015)
24. S. Rallabandi, S. Sitaram, A. Black, in *The Third Workshop on Computational Approaches to Linguistic Code-Switching(CALCS)*. Automatic detection of code-switching style from acoustics, (Melbourne, 2018), pp. 76–81. https://doi.org/10.18653/v1/w18-3209
25. Q. Wang, E. Yilmaz, A. Derinel, H. Li, in *The 20th Annual Conference of the International Speech Communication Association(Interspeech)*. Code-switching detection using ASR-generated language posteriors, (Graz, 2019), pp. 3740–3744. https://doi.org/10.21437/interspeech.2019-1161
26. Y. Long, Y. Li, Q. Zhang, S. Wei, H. Ye, J. Yang, Acoustic data augmentation for Mandarin-English code-switching speech recognition. Appl. Acoust. **161** (2020). https://doi.org/10.1016/j.apacoust.2019.107175
27. T. Ko, V. Peddinti, D. Povey, S. Khudanpur, in *The 16th Annual Conference of the International Speech Communication Association(Interspeech)*. Audio augmentation for speech recognition (ISCA, Dresden, 2015), pp. 3586–3589
28. Y. R. Pandeya, D. Kim, J. Lee, Domestic cat sound classification using learned features from deep neural nets. Appl. Sci. **8**(10) (2018). https://doi.org/10.3390/app8101949
29. D. S. Park, W. Chan, Y. Zhang, et.al., Specaugment: a simple data augmentation method for automatic speech recognition, 2613–2617 (2019). https://doi.org/10.21437/interspeech.2019-2680
30. Y. R. Pandeya, B. Bhattarai, J. Lee, in *IEEE 2020 International Conference on Information and Communication Technology Convergence (ICTC)*. Sound event detection in cowshed using synthetic data and convolutional neural network, (Singapore, 2020), pp. 273–276. https://doi.org/10.1109/ictc49870.2020.9289545
31. H. Adel, D. Telaar, N. Vu, K. Kirchhoff, T. Schultz, in *The 15th Annual Conference of the International Speech Communication Association (Interspeech)*. Combing recurrent neural networks and factored language models during decoding of code-switching speech (ISCA, Singapore, 2014), pp. 1415–1419
32. C. T. Chang, S. P. Chuang, H. Y. Lee, in *The 20th Annual Conference of the International Speech Communication Association(Interspeech)*. Code-switching sentence generation by generative adversarial networks and its application to data augmentation, (Graz, 2019), pp. 554–558. https://doi.org/10.21437/interspeech.2019-3214
33. A. Pratapa, G. Bhat, M. Choudhury, S. Sitaram, S. Dandapat, K. Bali, in *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Language modeling for code-mixing: the role of linguistic theory based synthetic data, (Melbourne, 2018), pp. 1543–1553. https://doi.org/10.18653/v1/p18-1143
34. C. Yeh, C. Huang, L. Sun, L. Lee, in *The 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. An integrated framework for transcribing Mandarin-English code-mixed lectures with improved acoustic and language modeling, (Tainai, 2010), pp. 214–219. https://doi.org/10.1109/iscslp.2010.5684908

35. Y. Qian, J. Liu, in *Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Phone modeling and combining discriminative training for Mandarin-English bilingual speech recognition, (Florence, 2010), pp. 4918–4921. https://doi.org/10.1109/icassp.2010.5495112
36. C. F. Yeh, L. S. Lee, in *Proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Transcribing code-switched bilingual lectures using deep neural networks with unit merging in acoustic modeling, (Florence, 2014), pp. 220–224. https://doi.org/10.1109/icassp.2014.6853590
37. R. Sennrich, B. Haddow, A. Birch, in *The 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Neural machine translation of rare words with subword units, (Berlin, 2016), pp. 1715–1725. https://doi.org/10.18653/v1/p16-1162
38. C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, L. Xie, in *Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Investigating end-to-end speech recognition for Mandarin-English code-switching, (Brighton, 2019), pp. 6056–6060. https://doi.org/10.1109/icassp.2019.8682850
39. E. Yilmaz, M. McLaren, H. van den Heuvel, D. A. van Leeuwen, Semi-supervised acoustic model training for speech with code-switching. Speech Commun. **105**, 12–22 (2018)
40. P. Guo, H. Xu, L. Xie, E. S. Chng, in *The 19th Annual Conference of the International Speech Communication Association(Interspeech)*. Study of semi-supervised approaches to improving English-Mandarin code-switching speech recognition, (Hyderabad, 2018), pp. 1928–1932. https://doi.org/10.21437/interspeech.2018-1974
41. H. Xu, S. DIng, S. Watanabe, in *Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Improving end-to-end speech recognition with pronunciation-assisted sub-word modeling, (Brighton, 2019), pp. 7110–7114. https://doi.org/10.1109/icassp.2019.8682494
42. T. N. Sainath, R. Prabhavalkar, S. Kumar, S. Lee, A. Kannan, D. Rybach, V. Schogol, P. Nguyen, B. Li, Y. Wu, Z. Chen, C.-C. Chiu, in *Proceedings of the 43th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. No need for a lexicon? Evaluating the value of the pronunciation lexica in end-to-end models, (Calgary, 2018), pp. 5859–5863. https://doi.org/10.1109/icassp.2018.8462380
43. Q. Zhang, J. Pan, Y. Yan, in *Proceedings of the 33th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Mandarin-English bilingual speech recognition for real world music retrieval (IEEE, Las Vegas, 2008), pp. 4253–4256
44. M. Bisani, H. Ney, Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. **50**(5), 434–451 (2008)
45. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, et.al., in *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. The Kaldi speech recognition toolkit (IEEE, Waikoloa, 2011)
46. D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, et.al., in *The 17th Annual Conference of the International Speech Communication Association(Interspeech)*. Purely sequence-trained neural networks for ASR based on lattice-free MMI, (San Francisco, 2016), pp. 2751–2755. https://doi.org/10.21437/interspeech.2016-595
47. , in *The 18th Annual Conference of the International Speech Communication Association(Interspeech)*. An exploration of dropout with LSTMs, (Stockholm, 2017), pp. 1586–1590. https://doi.org/10.21437/interspeech.2017-129
48. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, in *Advances in Neural Information Processing Systems (NIPS)*. Attention is all you need (NIPS, Long Beach, 2017), pp. 5998–6008
49. L. Dong, S. Xu, B. Xu, in *Proceedings of the 43th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition (IEEE, Calgary, 2018), pp. 5884–5888
50. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, T. Ochiai, in *The 19th Annual Conference of the International Speech Communication Association(Interspeech)*. ESPnet: end-to-end speech processing toolkit (ISCA, Hyderabad, 2018), pp. 2207–2211

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.