CROSS-LINGUAL PHONEME MAPPING FOR LANGUAGE ROBUST CONTEXTUAL SPEECH RECOGNITION

Ami Patel, David Li, Eunjoon Cho, Petar Aleksic

Google Inc., USA
{ampatel,jiayang,ejcho,apetar}@google.com

ABSTRACT

Standard automatic speech recognition (ASR) systems are increasingly expected to recognize foreign entities, yet doing so while preserving accuracy on native words remains a challenge. We describe a novel approach for recognizing foreign words by injecting them with appropriate pronunciations into the recognizer decoder search space on-the-fly. The pronunciations are generated by mapping pronunciations from the foreign language's lexicon to the target recognizer language's phoneme inventory. The phoneme mapping itself is learned automatically using acoustic coupling of Text-to-speech (TTS) audio and a pronunciation learning algorithm. Evaluation of our algorithm on Google Assistant use cases shows we can improve recognition of media-related queries by incorporating English entity pronunciations in French and German recognizers, with wins/losses ratios of roughly 2-3:1, without hurting recognition on general traffic.

Index Terms— cross-lingual, speech recognition

1. INTRODUCTION

Speech systems are typically trained and operate at a per language level. However, there are various applications where the correct handling of foreign entities is important for TTS and ASR. Navigating to foreign geographic locations, querying foreign media entities, and calling personal contacts of foreign origin are scenarios where robustness to foreign entities would prevent recognition errors for ASR and a perceived unnaturalness for TTS services.

For ASR, the focus of our work, building a system that is robust to foreign entities can be challenging. There is prior work on training multilingual acoustic models [1, 2, 3], usually with the objective of building a good baseline recognizer for languages where training data is limited. To improve existing systems with language-specific acoustic models, there are also efforts on building a language robust Grapheme-to-Phoneme (G2P) model [4] to retrieve the right pronunciations for foreign entities. We instead use a phoneme-mapping model that uses a word's actual pronunciation in the source language and finds its closest approximation in the target language's phoneme inventory.

Our objective is to maintain the quality of a traditional per-language recognizer but allow it to accept foreign words with the mapped pronunciation on-the-fly during decoding in contexts where foreign entities are expected to be more prevalent. An example would be when a French-speaking asks to play an English song. We propose a mechanism that adapts dynamic classes [5] to incorporate mapped pronunciations and utilizes contextual biasing [6] to boost the likelihood of various types of foreign entities only in pertinent contexts.

Developing a phoneme mapping between two language pairs has been explored in prior literature. Acoustic-phonetic similarity [7], articulatory feature-based mapping [8], and learning mappings from data [9] are some common approaches. A strategy that learns the phoneme mapping using TTS synthesized audio and the recognizer was explored in [10]. We expand on this data-driven approach by using a pronunciation learning algorithm [11] on TTS audio to learn the mapping between two languages. This algorithm has the advantage of constraining the parameter space with the graphemes, instead of relying purely on the audio signals.

In summary, our contribution is twofold: a process for automatically learning a phoneme mapping with data using pronunciation learning, and a method to contextually inject foreign words with correct pronunciations into the ASR decoder. Sec. 2 describes the contextual ASR system used to recognize foreign words, Sec. 3 describes our cross-lingual mapping algorithm, and Sec. 4 presents evaluations.

2. CONTEXTUAL ASR

In this section we describe how dynamic classes [5] and on-the-fly language model (LM) rescoring [6] are used to incorporate foreign entities' pronunciations into the decoder based on context.

2.1. Dynamic Classes

Dynamic classes, introduced in [5], can be injected into an LM via an arc coming off the unigram state to provide class-based entities at recognition time. Dynamic classes are constructed from a set of entities such as song names, contact names, device names, etc. into finite state transducers (FSTs).

If any entities include out of vocabulary (OOV) words, their pronunciations are obtained on-the-fly and directly included in the dynamic class. Specifically, for a dynamic class d, we build a G_d (FST over words) for all OOV words, and then a d-specific lexicon L_d providing pronunciations for the OOV words in G_d. The dynamic class FST incorporates the OOV construct by building G'_d as follows:

$$G'_d = Det(L_d) \circ G_d$$

To correctly recognize foreign entities, we generalize the dynamic class construction process by additionally considering the source language for each instance. Based on the language information, which can be supplied with the instance or inferred online using a language classifier, we include any foreign word whose language is different from that of the target recognizer into G_d regardless of whether the word is an OOV or not. During the L_d construction, we use the language information to decide from which language's lexicon to fetch the pronunciation. We map the pronunciations of foreign words into the recognizer language's phoneme set using phoneme mapping. The resultant G'_d then encompasses their approximate pronunciations. Figure 1 shows an example of \$SONG dynamic class containing two English song entities to be used in the French recognizer.

After it is constructed, a dynamic class is spliced into the base LM with a certain LM cost on the class open tag arc to avoid over-triggering. As described in the next section, we then adjust the class LM cost on-the-fly based on contextual information.

2.2. On-the-fly LM Rescoring

To ensure that the dynamic class is not pruned out during decoding, we contextually lower the LM cost associated with the dynamic class of interest using on-the-fly LM rescoring (biasing), described in [6].

The context is captured using a set of biasing phrases relevant for a particular dynamic class. For example, in the case of \$SONG dynamic class, relevant phrases are listed in Table 1. These phrases include the class name (\$SONG) as a placeholder for any instance belonging to that class.

French biasing phrases	German biasing phrases
mets \$SONG	spiel \$SONG
jouer \$SONG	spiele \$SONG

Table 1. French and German \$SONG biasing phrases.

The phrases are compiled into a biasing model represented as a weighted FST. The weight of any n-gram in the biasing model represents how much the LM cost of that particular n-gram will be altered. These weights can be learned from logs or explicitly set per context [6]. In this work, we use unigram/bigram method of assigning biasing weights, where all biasing unigrams have identical weights, as well as bigrams.

For each word emitted during decoding, the cost from the original LM, G, and the cost from the biasing model, B, determine the actual cost as follows:

$$s(w|H) = \begin{cases} s_G(w|H), & \text{if } (w|H) \notin B \\ C(s_G(w|H), s_B(w|H)), & \text{if } (w|H) \in B \end{cases}$$

where $s_G(w|H)$ and $s_B(w|H)$ are the costs of the word w with history H from G and B respectively. Using a linear interpolation together with a minimum function for C ensures that the costs can only be decreased with biasing:

$$C(s_G(w|H), s_B(w|H)) = \min(s_G(w|H), \alpha s_G(w|H) + \beta s_B(w|H))$$

Each dynamic class entity is assigned a biasing weight corresponding to the weight associated to \$SONG in biasing phrases.

3. CROSS-LINGUAL PHONEME MAPPING

Our acoustic coupling method for learning a cross-lingual mapping relies on a set of pronunciations in the source language, a TTS system for the source language that can generate audio for these pronunciations, and a pronunciation learning system in the target language. We use these to generate the source/target language pronunciation pairs that are used to learn the phoneme mapping; the mapping can then generate target language pronunciations from novel source language pronunciations. Additionally, in our setup, languages' phoneme inventories are a subset of language-independent X-SAMPA [12]; therefore, we assume that we only need to generate mappings for source language phonemes that are not present in the target language inventory (in practice, 47.5% of phonemes when mapping English to French, and 20% of phonemes when mapping English to German).

Given native source language words for which we have a human-sourced pronunciation, we first synthesize audio of the pronunciation using a TTS voice in the source language. Using TTS audio rather than standard datasets allows us to be certain of the pronunciation used in the audio. To get a representative distribution over phonemes in different contexts, we synthesize the pronunciations of a large set of native words.

To learn the pronunciations from the synthesized audio, we use the method described in [11], which uses FSTs to generate pronunciation candidates based on the graphemes. The pronunciation model score from the FST is combined with the acoustic model score to determine the most likely pronunciation given the audio and the graphemes. The FST used is created from an RNN-transducer - a sequence-to-sequence neural model. This scheme provides an infinite number of weighted pronunciation candidates, with the graphemes serving as a useful additional cue. For example, for an acoustic

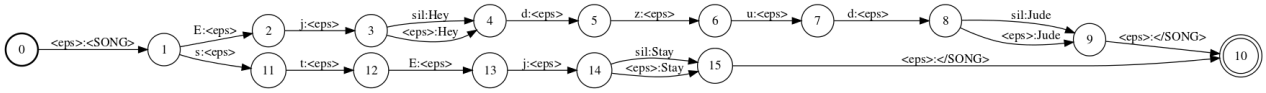

Fig. 1. Example of a $SONG dynamic class with two English song names ("Hey Jude" and "Stay") and their original English pronunciations mapped onto French phoneme set.

model trained only on French data, acoustic similarity alone might suggest a mapping from English to French of the glottal fricative /h/ to the rhotic uvular fricative /R/. However, in practice, we might expect French speakers to assign a high weight to the difference in rhoticity, making a deletion of /h/ the better model.

Given the source and target language pronunciation pairs, we find the best alignments between the constituent phonemes, similarly to the first step of training a traditional joint-sequence grapheme to phoneme model [13]. We allow a one-to-many source phoneme to target phoneme alignment. Each phoneme in the source pronunciation can correspond to 0-2 consecutive phonemes in the target pronunciation. Let

$$q = (s, \mathbf{t}) \in (S \times \bigcup_{i=0,1,2} T_i).$$

where S is the set of source phonemes, T is the set of target phonemes, and T_i is the set of all strings made of elements of T of length i. The set of alignments between a source and target phoneme sequence $\mathbf{s}$ and $\mathbf{t}$ is

$$A(\mathbf{s}, \mathbf{t}) = \{\, q_1 ... q_n \in q^* |\, s_1 ... s_n = \mathbf{s};\ \mathbf{t_1} ... \mathbf{t_n} = \mathbf{t} \,\}$$

where $q_i = (s_i, \mathbf{t_i})$.

We apply the expectation-maximization algorithm on an observation set $\mathbf{O}$ of $(\mathbf{s}, \mathbf{t})$ pairs to iteratively estimate values for $p(q)$ that optimize the likelihood of the training data:

$$\log(p(\mathbf{O})) = \sum_{(\mathbf{s},\mathbf{t}) \in \mathbf{O}} \log(\sum_{q_1 ... q_n \in A(\mathbf{s},\mathbf{t})} \prod_{i=1}^{n} p(q_i))$$

The mapping is then defined as

$$\underset{S \to T}{\text{mapping}}(s) = \begin{cases} s, & \text{if } s \in T \\ \text{argmax}_{\mathbf{t}}\, p(s, \mathbf{t}), & \text{otherwise} \end{cases}$$

The one-to-many mapping is expected to useful in the case of diphthongs - by definition, a combination of two vowels sounds - or in cases like the velar nasal /N/ which might be better approximated by a concatenation of the palatal nasal and velar stop /n g/ than by either of them alone.

We compare acoustic coupling results to a linguistically informed manual mapping with the same constraints: each source language phoneme not in the target language inventory maps to a sequence of 0-2 target language phonemes. Between the manual and acoustic coupling mapping, there were differences in the mappings for 30% of all English phonemes when mapped to French, and 10% when mapped to German.

4. EVALUATION

To evaluate our system we explore recognizing English media queries using the French and German recognizers, as this is a common use case for cross-lingual entities. We collected 1000 popular English songs to construct a $SONG dynamic class to be used during contextual rescoring. For cross-lingual mapping we evaluate both the human-generated mappings and the mappings generated from acoustic coupling.

We show two types of experiments. In Sec. 4.1, we evaluate the the word error rate (WER) of our system on a test set with English media song utterances using the French and German recognizers. In Sec. 4.2, we compare our system to different baselines using real anonymized speech traffic that includes queries to play media entities.

4.1. ASR test sets

To evaluate the effect of contextual rescoring and cross-lingual mapping in our speech recognizer, we set up test sets and measured WERs for our different mapping algorithms. Finding and transcribing utterances for specific cross-lingual scenarios (in our case English media) can be challenging. An efficient alternative is to generate test sets using TTS voices. In our case we used an English TTS voice with a US accent to generate 1000 English media utterances and evaluated them on French and German recognizers. Here, the baseline is a production voice search recognizer trained primarily on single-language data.

Setup	French	German
Baseline	95.7	63.6
Contextual rescoring	27.7	14.3
Contextual rescoring with Human-generated mapping	23.4	7.8
Contextual rescoring with Acoustic coupling mapping	19.9	6.5

Table 2. WERs [%] across different contextual rescoring and cross-lingual mapping setups.

We see in Table 2 that the baseline target recognizers consistently do a poor job at recognizing English media utterances. However, using contextual rescoring and additionally adding mapped pronunciations bring the WER down. We also performed experiments on test sets sampled from general traffic and found no WER regressions.

4.2. ASR SxS

To assess our system on real speech traffic, anonymized utterances were directed to both a baseline system and an experimental system for a SxS comparison. If there was a difference in the recognition result, both transcripts were sent to human raters to rate how well the transcript matched the audio. We collected at least 500 utterances that had differences between the two systems to ensure statistical significance. The two metrics we report are:

- The percentage of traffic that was changed.
- The ratio of wins to losses. A win refers to an utterance where the experimental system was rated higher than the baseline system.

As shown in Table 3 we have a setup where we constrain the input to only media queries. We do this by only selecting utterances that include certain action verbs indicating a request to play media, as is captured with our biasing phrases in Table 1. We also test on general Google Assistant traffic.

Setup #	Input traffic	Baseline system
1	Media	No contextual rescoring
2	Media	Contextual rescoring without cross-lingual mapping
3	General	No contextual rescoring

Table 3. Input traffic types and baselines for SxS experiments

For each SxS setup in Table 3 the experimental system uses contextual rescoring with the human-generated cross-lingual mapping applied to generate the pronunciations. Setup 1 compares our system against a baseline where there is no contextual rescoring. Setup 2 compares it with a system without mapped pronunciations to exclusively evaluate their effect. Setup 3 replicates Setup 1 on general traffic to verify that unrelated utterances are not negatively affected.

Language	Setup #	Percentage changed	Wins/Losses
French	1	1.4%	3.1
	2	0.6%	3.1
	3	0.3%	3.2
German	1	0.3%	2
	2	0.17%	2
	3	0.06%	1.3

Table 4. SxS experiment results for each setup in Table 3

The results for Setup 1 in Table 4 show quality gains produced by our system. The good wins/losses ratio suggests that the combination of cross-lingual mapping and contextual rescoring is a viable method to overcome the absence of these foreign entities in the recognizer. We attribute the smaller impact on the German recognizer to the linguistic similarity between English and German. Some examples of wins are:

- Joue MIGO Piou → Jouer *Help Me Help You*
- mets moi que bye → mets *Rockabye*
- spiel Amazon Mbel → spiel *I'm a Gummy Bear*
- Spiel chen der Wii → spiel *Legendary*

The percentage of traffic changed is expectedly lower when the baseline system already includes the regular dynamic class without phoneme mapping (Setup 2), but a good wins/losses ratio is still obtained by adding the mapped pronunciations. Setup 3 confirms that we are able to achieve the gains for media queries without affecting other utterances.

4.2.1. Human-generated mapping vs. acoustic coupling

We performed a SxS comparing acoustic coupling to a manual mappings baseline. The percentage of changed queries was low (<0.05%), indicating acoustic coupling shows a comparable gain over contextual rescoring with no mapping.

Though the difference in overall performance was not significant, the recognition differences show the ambiguity of the phoneme mapping problem, as both wins and losses result from changing the mapping of rhotacized schwa /@˞/:

- mets la chanson merci → mets la chanson *mercy*
- mets *believer* ... → mets believe her ...

and from altering the mapping of diphthong /oU/:

- Spiel des Passito → spiel *Despacito*
- spiel *Despacito* ... → Spiel des Passito ...

It is possible that acoustic coupling's use of graphemes to inform the mapping is helpful in cases where speakers may also adjust their pronunciation of foreign words based on graphemes. For example, this win resulted from acoustic coupling choosing to map dental fricative /T/ to alveolar stop /t/ instead of alveolar fricative /s/, though the latter may be linguistically more similar:

- Jouer Whitest ... → Jouer *Wild Thoughts* ...

5. CONCLUSION

We've presented an approach to recognize foreign entities based on context without hurting recognition on native words, through dynamic classes with pronunciation mapping and contextual rescoring. This allows us to reduce the WER on foreign media words by an additional 55% on top of the gains achieved using contextual rescoring, and the gain translates to improvement on real media queries. The phoneme mapping learned through acoustic coupling - without ground truth target language pronunciations or linguistic knowledge - is comparable to a human-generated mapping.

We would like to thank Alyson Pitts, Toby Hawker, Tony Bruguier, and Zelin Wu for suggestions and guidance.

6. REFERENCES

[1] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7319–7323.

[2] Liang Lu, Arnab Ghoshal, and Steve Renals, "Regularized subspace gaussian mixture models for cross-lingual speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 365–370.

[3] Kate M Knill, Mark JF Gales, Shakti P Rath, Philip C Woodland, Chao Zhang, and S-X Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 138–143.

[4] Xavi Gonzalvo and Monika Podsiadło, "Text-to-speech with cross-lingual neural network-based grapheme-to-phoneme models," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] Lucy Vasserman, Ben Haynor, and Petar Aleksic, "Contextual language model adaptation using dynamic classes," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 441–446.

[6] Petar Aleksic, Cyril Allauzen, David Elson, Aleksandar Kracun, Diego Melendo Casado, and Pedro J. Moreno, "Improved recognition of contact names in voice commands," in *ICASSP*, 2015, pp. 5172–5175.

[7] Gyorgy Szaszak and Philip N Garner, "Evaluating intra- and crosslingual adaptation for non-native speech recognition in a bilingual environment," in *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*. IEEE, 2013, pp. 357–362.

[8] Leonardo Badino, Claudia Barolo, and Silvia Quazza, "A general approach to tts reading of mixed-language texts," in *Eighth International Conference on Spoken Language Processing*, 2004.

[9] Khe Chai Sim and Haizhou Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4309–4312.

[10] Jan Nouza and Marek Boháč, "Using tts for fast prototyping of cross-lingual asr applications," in *Proceedings of the 2010 International Conference on Analysis of Verbal and Nonverbal Communication and Enactment*, Berlin, Heidelberg, 2011, COST'10, pp. 154–162, Springer-Verlag.

[11] Antoine Bruguier, Danushen Gnanapragasam, Leif Johnson, Kanishka Rao, and Françoise Beaufays, "Pronunciation learning with rnn-transducers," *Proc. Interspeech 2017*, pp. 2556–2560, 2017.

[12] JC Wells, "Computer-coding the ipa: a proposed extension of sampa," 1995, www.phon.ucl.ac.uk/home/sampa/x-sampa.htm.

[13] Maximilian Bisani and Hermann Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Commun.*, vol. 50, no. 5, pp. 434–451, May 2008.