Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 30 – 36

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Performance Improvement of Probabilistic Transcriptions with Language-Specific Constraints

Xiang Kong[a], Preethi Jyothi[b], Mark Hasegawa-Johnson[b,*]

[a]*Computer Science Department, University of Illinois at Urbana-Champaign, IL, 61801, USA*
[b]*Beckman Institute, University of Illinois at Urbana-Champaign, IL, 61801, USA*

Abstract

This article describes a method for reducing the error rate of probabilistic phone-based transcriptions resulting from mismatched crowdsourcing by using language-specific constraints to post-process the phone sequence. In the scenario under consideration, there are no native-language transcriptions or pronunciation dictionary available in the test language; instead, available resources include non-native transcriptions, a rudimentary rule-based G2P, and a list of orthographic word forms mined from the internet. The proposed solution post-processes non-native transcriptions by converting them to test-language orthography, composing with test-language word forms, then converting back to a phone string. Experiments demonstrate that the phone error rate of the transcription is reduced, using this method, by 22% on an independent evaluation-test dataset.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: automatic speech recognition resources ; probabilistic transcription ; mismatched crowdsourcing ; G2P

1. Introduction

Current data-driven speech recognition technologies have proven to be very successful for well-resourced languages (such as English and Mandarin Chinese). However, building these systems involves an expensive and laborious process of transcribing large amounts of speech recordings. This process of acquiring speech transcriptions is the main bottleneck to building automatic speech recognition (ASR) technologies for under-resourced languages[1]. Crowdsourcing, wherein the task of speech transcription is distributed among a large community of online workers, is a viable option to derive speech transcriptions[2]. This technique, however, requires the crowd workers to be native speakers of the language being recognized. This limits crowdsourcing for speech transcriptions to be applicable only to a small fraction of the worlds languages. Mismatched crowdsourcing was proposed to address this limitation[3].

In mismatched crowdsourcing, workers transcribe a language with which they are unfamiliar, using nonsense syllables in their own orthography. Their transcriptions are post-processed, using a noisy channel model of non-native

* Corresponding author.
E-mail address: jhasegaw@illinois.edu

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.026

speech perception, in order to create a probabilistic transcription: a probability distribution over the set of possible phonemic transcriptions of the target speech[3]. Probabilistic transcriptions can be used to train ASR[4]; though they are not as informative as native transcriptions, they are considerably more informative than ASR self-training, cross-language transfer learning, and other semi-supervised techniques applicable in a zero-data scenario.

Section 2 describes probabilistic transcription, and the proposed wordlist-based post-processing algorithm. Section 3 describes the data and the experimental setup, and Section 4 presents results.

We choose Greek to be our target language. Greek has limited ASR support and is one of the many languages that are not well represented on crowdsourcing platforms (as shown in a language demographic study on Amazon Mechanical Turk[6]).

2. Probabilistic Transcription and Target-Language Post-Processing

A probabilistic transcription is a distribution over talker-language phone sequences, $Pr(\pi|T)$, where T is a set of transcriptions in the orthography of the transcriber language. When the transcriber language differs from the talker-language, $Pr(\pi|T)$ is rarely a zero-entropy distribution. Post-processing methods that incorporate side information about the talker language therefore offer the possibility of improved quality, in the form of reduced transcription entropy

2.1. Probabilistic Transcription

A probabilistic transcription is defined to be:

$$\begin{aligned} Pr(\pi|T) &= \sum_{\lambda} Pr(\pi|\lambda,T)Pr(\lambda|T) \\ &\approx Pr(\pi|\lambda)Pr(\lambda|T) \\ &= max_{\lambda}\left(\frac{Pr(\lambda|\pi)}{Pr(\lambda)}Pr(\pi)\right)Pr(\lambda|T) \end{aligned} \tag{1}$$

where λ is in the orthography of the transcriber language, and π is in the phone set of the talker language. $\Pr(\pi|T)$ is the probability that λ is the best transcriber-language orthographic transcription, given the set T generated by crowd workers[5]; $\Pr(\pi|\lambda)$ is the probability that is the best talker-language phone transcription[3]. $\Pr(\lambda)$ is a simple prior over transcriber-language letter sequences, and $\Pr(\pi)$ is a language model prior over talker-language phone sequences.

If native language transcriptions do not exist in the test language, then it is impossible to train the noisy channel model $\Pr(\lambda|\pi)$ using true knowledge of any test-language phone sequence. $\Pr(\lambda|\pi)$ is therefore trained using transcriber-language orthographic sequences and their corresponding talker-language phone sequences in talker-languages other than the test language[4]. In experiments described here, Greek is the test language, but $\Pr(\lambda|\pi)$ is trained using speech in Arabic, Cantonese, Dutch, Hungarian, Mandarin, Swahili and Urdu.

Though mismatched crowdsourcing was invented for zero-resource scenarios, it can only currently be evaluated in languages for which a reference evaluation-test transcription exists. Experiments in this paper evaluate the probabilistic transcriptions by comparison to an orthographic transcription provided to us by a native speaker of Greek. A probabilistic transcription is a distribution over possible phone sequences, therefore it may be evaluated by computing the phone error rate (PER) of the 1-best transcription, or of the N-best transcription, or by computing the probability of the native transcription given the probabilistic transcription. Previous studies[5] showed that these measures are highly correlated, therefore this paper reports PER of the 1-best transcription:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \Pr(\pi|T) \tag{2}$$

2.2. Target-Language Post-Processing

Suppose that the distribution $\Pr(\pi|T)$ is represented in the form of a weighted finite state transducer (FST); use the symbol **PT** to represent this FST. It is possible to reduce the entropy of **PT** by applying language-specific constraints, e.g., side information available as the result of mining the web for information about the test language.

Suppose that a pronunciation lexicon is unavailable in the target language, but that it is possible to acquire a long list of attested orthographic word-forms in the test language. Using this list, it is possible to create an orthographic language model in the target language. If the available text data are sparse, as assumed in this paper, then the language model may even be a 0-gram, in which every attested word form is considered to be equally likely. This language model can also be represented in the form of an FST, which can be denoted with the symbol **LM**.

Though a proper pronunciation model may be unavailable, it is often possible to estimate the set of all possible grapheme-to-phoneme (G2P) mappings for the target language by downloading and appropriately reformatting the Wikipedia page titled XXX Alphabet, where XXX is the test language (in this case, Greek). The resulting G2P can be constructed to generate, for any given grapheme sequence in the test language, all phone sequences that are attested on Wikipedia as possible pronunciations, with equal likelihood for any of the attested pronunciations. An unweighted G2P of this kind is likely to be useless for automatic speech recognition, but may be useful for the purposes described in this paper, as it imposes a loose upper bound on the set of possible phone sequences that might correspond to any given orthographic sequence. Such G2Ps have been constructed for seventy languages, and are available at here[10]. Let us denote the resulting FST mapping graphemes to phonemes as **G2P**.

The transducers **G2P** and **LM** represent a minimal set of information that can be mined from internet sources for at least several hundred different languages; though there is little information in these transducers, it is possible that the constraints they represent can reduce the entropy of the distribution $\Pr(\pi|T)$. The information in these transducers is applied to **PT** by composing the FSTs, creating the modified probabilistic transcription $\widehat{\mathbf{PT}} = \mathbf{G2P}^{-1} \circ \mathbf{LM} \circ \mathbf{G2P} \circ \mathbf{PT}$. Expressing this operation using properly normalized probability distributions is a little tricky, since the edge weights in transducers **G2P** and **LM** are not considered to be useful estimates of the corresponding transition probabilities. It is possible to represent these constraints accurately using zero-exponentiation, defined such that $x \triangleq 1$ if and only if $x^0 \neq 1$ but $0^0 \triangleq 0$. Using this notation, the edge weights in **G2P** are a model of $Pr(W|\pi)^0$ and those in **LM** are a model of $Pr(W)^0$, therefore the edge weights in $\widehat{PT}$ are a model of

$$\widehat{PT}(\pi|T) = \sum_W \Pr(W|\pi)^0 \Pr(W)^0 \Pr(\pi|W)^0 \Pr(\pi|T) \tag{3}$$

Eq. (3) suggests post-processing **PT** using a combination of a language-dependent G2P (representing the set of grapheme-to-phoneme transductions that are attested in the target language) and a language-dependent zero-gram language model (representing attested orthographic word forms). As an intermediate step, it is possible to limit PT by simply pruning away phonemes that dont exist in the test language. Phoneme pruning can be computed by the FST composition $\widetilde{PT} = \mathbf{G2P}^{-1} \circ \mathbf{G2P} \circ \mathbf{PT}$, which computes

$$\widetilde{PT}(\pi|T) = \sum_W \Pr(W|\pi)^0 \Pr(\pi|W)^0 \Pr(\pi|T) \tag{4}$$

3. Experimental Methods

Speech data from podcasts in the test language, and in six other talker languages, were each transcribed by native and non-native transcribers. Probabilistic transcriptions were created from mismatched crowdsourcing using the methods[4]. Auxiliary information including a word list and a rudimentary G2P were mined from the internet in the target language, and used to post-process the probabilistic transcription

3.1. Speech Data

Speech data were extracted from Special Broadcasting Service Australia (SBS) radio podcasts[7]. Native transcriptions were acquired for 40-60 minutes of speech in seven different languages (Arabic, Cantonese, Dutch, Hungarian, Mandarin, Swahili, Urdu). The Cantonese transcriptions were collected at I^2R, Singapore as part of a collaborative research project and transcriptions for the remaining six languages were collected from paid student volunteers at the University of Illinois, Urbana-Champaign who were native speakers of these languages.

Evaluation data included native transcriptions for roughly 20 minutes of Greek speech. These were obtained by finding Greek native speakers on the freelance platform UpWork which explicitly allows online workers to list their language skills[8]. The Greek data were randomly split into a development set and an evaluation set, of roughly equal size. All the native transcriptions were converted into phonemic sequences using a universal phone set. This universal set was constructed manually starting from IPA symbols appearing in canonical descriptions of all seven languages and merging phones (with the closest phone differing in only a distinctive feature) to ensure that each phone was covered by at least two languages.

The native transcribers for all eight languages transcribed short 5-second speech clips that were spliced out of SBS radio podcasts to be largely homogenous in the target language. Before collecting mismatched transcriptions, these clips were further split into 1-second segments to make the transcription task easier for the mismatched crowd workers. The mismatched transcriptions were obtained from workers on Amazon Mechanical Turk[9], using the methods in the paper[4]. Ten distinct crowd workers transcribed each clip; each worker was asked to listen to a speech clip and provide a sequence of nonsense English syllables that is a closest match to what they heard.

3.2. Test-Language Grapheme-to-Phoneme Transducer

Since there is no standard pronunciation dictionary available for Greek, we resort to two simple techniques to derive pronunciation constraints for Greek.

First, a simple set of G2P mappings were compiled for Greek based on the description of its orthography[10]. These rule sets can be fairly easily generated for a range of different languages[10]. Table 1 shows a subset of these mappings for Greek.

Table 1. Subset of G2P mappings for Greek[10].

Greek Orthography	IPA Phonemes (t)
AI	/e/
EI	/i/
η	/i/
AY	/av/
AY	/af/
ΓΓ	/g/

Table 2. Rules to map ASCII-based phone alphabet to IPA for Greek

ASCII Phone Code	IPA Phonemes (t)
u	/i/
x	/ks/
th	/Θ
b	/v/
k	/k/
ph	/f/

Second, a Greek pronunciation dictionary was constructed using data available from the Translation as a Service (Taas) project[11], which has about 200,000 Greek words with corresponding pronunciations available for some words. These pronunciations use an ASCII-based alphabet which we further map to IPA using a deterministic phone mapping. Table 2 shows a subset of these phone mappings.

3.3. Bigram Phoneme Language Model

A bigram phone language model for Greek is needed to represent $Pr(\pi)$ in Equation (1). In order to build this language model, we extracted text documents from Wikipedia for Greek. This text consisted of 100,000 sentences and 20,000,000 Greek word tokens. We used the CMU CLMTK toolkit to train a bigram language model in the ARPA format[12]. This was further represented as a finite state acceptor using the arpa2fst utility in Kaldi[13]. In order to train a phone language model, the Greek word sequences were first mapped to phone sequences using the above-described Greek pronunciation dictionary, along with applying the Greek G2P rules for any out-of-vocabulary words.

4. Results and Discussion

We investigate the importance of Greek pronunciation constraints on the quality of probabilistic transcriptions by measuring the phone error rates on both the development and evaluation sets for Greek. We first use the inverse of the Greek G2P rules to map the phone-based transcriptions into Greek word sequences and then map them back to phone sequences for evaluation (computing the $\widetilde{PT}$ transducer of Eq. (4)). This simple constraint enforces the probabilistic transcriptions to be matched with valid Greek words. This is referred to as "G2P".

For the sake of comparison, we also compute phone error rates using the Greek dictionary (described above) in conjunction with the G2P rules (computing the $\widehat{PT}$ transducer of Eq. (3)). Greek words appearing in the dictionary are mapped to their corresponding pronunciations and any remaining out-of-vocabulary words are mapped to phones using the Greek G2P rules. This constraint is referred to as "G2P + dict".

We also compute the $\widehat{PT}$ transducer after imposing word constraints using a bigram word language model (i.e., obtained by replacing $Pr(W)^0$ in Eq. (3) with $Pr(W)$ derived from a bigram model). Similar to the phone bigram model, we use the Greek Wikipedia text and the CMU CLMTK toolkit to train a bigram word-level language model. We refer to this constraint as "WLM".

Figures 1 and 2 show the 1-best, 2-best, 4-best and 50-best phone error rates (PERs) on the development set and evaluation set, respectively. The 1-best PERs are listed in Table 3 using all the three pronunciation constraints on the probabilistic transcriptions. Table 3 shows that constraining the probabilistic transcription with a language-dependent

Table 3. 1-best probabilistic phone transcription error rates on the development and evaluation sets for Greek.

	Dev set (1-best)	Eval set (1-best)
Original	78.6	81.2
G2P	69.6	70.8
G2P+dict	64.1	66.7
WLM	60.1	62.9

G2P (the $\widetilde{PT}$ transducer of Eq. (4)) significantly improves the 1-best PER of the transcription. Constraining using a language-dependent zero-gram language model (the $\widehat{PT}$ transducer of Eq. (3)) further improves the transcription accuracy. Replacing the zero-gram language model with a bigram model reduces the PERs even further, suggesting that word-level constraints are useful despite the use of sparse text data, as assumed in this paper, to train the language model.

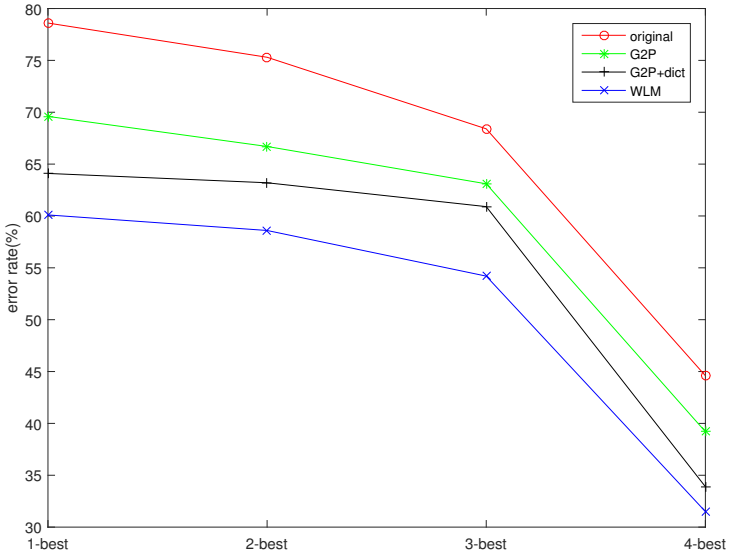

Fig. 1. PERs on the development set.

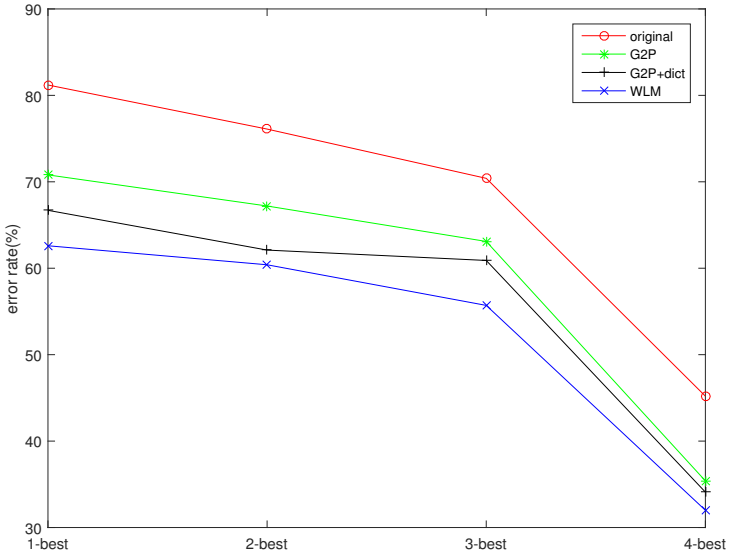

Fig. 2. PERs on the evaluation set.

From Figures 1 and 2, we also observe significant PER improvements moving beyond 1-best PERs to 2-best, 4-best and 50-best PERs, across all four experimental settings. This illustrates the amount of useful information captured by the PTs, which could be leveraged in building an ASR system for Greek[4]. PERs of the 50-best transcription corresponding to Original and WLM on the evaluation set are 44.6% and 31.5%, respectively. PTs rescored using the WLM constraint are significantly improved compared to the original PTs without any pronunciation constraints, when evaluated using both 1-best PERs and N-best (N=2,4,50) PERs.

5. Conclusion

In our experiments, we regard phone sequences directly generated through probabilistic transcriptions (PTs) derived from mismatched transcriptions as a baseline, and use a Greek G2P and a pronunciation dictionary as constraints onto phone sequences to make results more accurate. From experiments, we can see our methods can significantly improve performance of PTs. Phone error rates can be reduced up to about 22%.

References

1. Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, Automatic speech recognition for under-resourced languages: A survey, *Speech Communication* 2014;**56**:85-100.
2. Scott Novotney and Chris Callison-Burch, Cheap, fast and good enough: Automatic speech recognition with non-expert transcription, . *NAACL HLT* 2010
3. Preethi Jyothi and Mark Hasegawa-Johnson, Acquiring Speech Transcriptions using Mismatched Crowdsourcing, in Proceedings of AAAI 2015
4. Chunxi Liu, Preethi Jyothi, Hao Tang, Vimal Manohar, Rose Sloan, Tyler Kekona, Mark Hasegawa-Johnson, and Sanjeev Khudanpur, Adapting ASR for Under-Resourced Languages Using Mismatched Transcriptions To appear in Proceedings of ICASSP 2016.
5. Preethi Jyothi and Mark Hasegawa-Johnson, Transcribing Continuous Speech using Mismatched Crowdsourcing. *n Proceedings of Interspeech,* 2015.
6. Ellie Pavlick, Matt Post, Ann Irvine, Dimitry Kachaev, and Chris Callison-Burch, The Language Demographics of Amazon Mechanical Turk Transactions of ACL 2016; **2**:79-92.
7. Special Broadcasting Service Podcasts in Your Language, Downloaded 1/26/2016 from http://www.sbs.com.au/podcast/yourlanguage
8. Upword Downloaded 1/26/2016 from https://www.upwork.com
9. Amazon Mechanical Turk, Downloaded 1/26/2016 from http://www.mturk.com
10. Mark Hasegawa-Johnson, WS15 dictionary data, Downloaded 11/15/2015 from http://isle.illinois.edu/sst/data/dict
11. Translation as a Service (TASS), Downloaded 1/26/2016 from http://www.taas-project.eu/
12. arnegie Mellon University Language Modeling Toolkit CMUCLMTK, Downloaded 1/26/2016 from http://cmusphinx.sourceforge.net/wiki/
13. arpa2fst Language-Model to FST Converter, Downloaded 1/26/2016 from $http://kaldi.sourceforge.net/arpa2fst_8cc_source.html$"