Mark Hasegawa-Johnson
University of Illinois
405 N Mathews
Urbana, IL 61801
jhasegaw@illinois.edu

Dr. Tara Sainath
Associate Editor
IEEE Transactions on Speech and Audio Processing
tsainath@google.com

Dear Dr. Sainath,

Please find, attached, a revised draft of our manuscript "ASR for Under-Resourced Languages from Probabilistic Transcription." My co-authors and I have discussed all reviewer comments, and have modified the text of the manuscript in response to each. These modifications are described, in detail, in our itemized responses to reviewer comments below. Please forward the revised manuscript, and this letter, to reviewers of the manuscript. If there are any reviewer concerns that can be addressed quickly in a short correspondence, I hope you will forward them to me so that I may address them without delay.

Faithfully yours,

Mark Hasegawa-Johnson
Professor, ECE, University of Illinois

Responses to reviewer comments

Reviewer comments are in blue font. Responses are in black font.

Reviewer 1

There is only one hour data used for each language in the experiments. It is hard to argue that trying to transcribe one or a few hours speech data for one language by native speakers is really beyond the capability of that language community. What is really beyond the capability is to transcribe massive data by native speakers. So I think the paper would be much more convincing if the experiments could be conducted using more data (e.g. 100 hours or even 10 hours of speech data for each language) and it is also important to see if the performance and impact of PT will change in such scenarios.

The first paragraph of the paper has been changed to explicitly mention the difficulty of finding native transcribers willing to transcribe even one hour of speech audio: "Large corpora are beyond the resources of most under-resourced language communities; we have found that transcribing even one hour of speech may be beyond the reach of communities that lack large-scale government funding. In order to create the databases reported in this paper, we sought paid native transcribers, at a competitive wage, for the 68 languages in which we have untranscribed audio data. We found transcribers willing to work in only eleven of those languages, of which only seven finished the task."

Section V has been changed to explicitly mention the long-term goal of using larger corpora: "It is desirable to test the ideas in this paper with corpora larger than one hour per language, but larger corpora involve problems orthogonal to the purposes of this paper, e.g., the Babel corpora contain telephone speech, and therefore contain far more acoustic background noise than the podcast corpora used in this paper."

Discriminative training of GMM-HMMs. It is a little disappointing that only ML models are investigated for GMM-HMMs as discriminative GMM-HMMs such as MPE or BMMI models are well known to outperform ML-trained GMM-HMMs. Putting aside PT, with deterministic transcriptions discriminative models are the state of the art in GMM-HMM-based acoustic modeling. I would speculate that transcriptions with probabilities probably are more suitable to generative model such as ML-based GMM-HMMs. Since discriminative models are simply better models, it is worth investigating how PT will impact under such conditions and it would be more valuable to show impact on the best models available.

New experiments were conducted to train both CL baseline systems and PT-adapt systems using MMI, MPE and sMBR training criteria. Table IV now presents PERs of the MPE and sMBR systems. Text has been added to describe these results. About the baselines, the following text has been inserted: "Three different types of discriminative training were tested. MMI performs consistently worse than MPE and sMBR, and is therefore not listed in Table IV. Averaged across all languages and systems shown in Table IV, the development-test PERs of ML, MPE, and sMBR training are 73.43%, 73.04%, and 72.98% respectively; differences are not statistically significant, therefore only the ML system was tested on evaluation test data." About PT-adapt, the following text has been inserted: "PT-adapt GMM-HMM systems were trained using four different training criteria: ML, MMI, MPE and sMBR. MMI training consistently underperformed MPE and sMBR, and is therefore not shown. MPE training of PT-adapt systems improves their PER by a little more than 1% on average, comparable to the improvement provided to CL baseline systems." In fact, discriminative training helps the PT-adapt systems a little bit more than it helps the baseline systems, but the difference is small and probably not significant.

Sequence training of NNs. It has been observed that techniques that give improvements at the cross-entropy level may give little gain after sequence training. Therefore, sequence training using PT worth looking into.

Authors agree with the reviewer, but the time allocated for this manuscript revision was insufficient to implement sequence training using PTs. Future research may pursue this possibility.

How are the GMM-HMM configured? How many states and Gaussians? What's the feature space?

The following text has been added to the new section VII.A: "Acoustic features consisted of MFCC (13 features), stacked ±3 frames (13x7=91 features), reduced to 40 dimensions using LDA followed by fMLLR. GMM-HMM systems directly observed this 40-dimensional vector; NN-HMM systems computed fMLLR+d+dd stacked ±5 frames (40x3x11=1320 features/frame). All systems used tied triphone acoustic models, based on a decision tree with 1200 leaves. Each GMM-HMM used a library of 8000 Gaussians, shared among the 1200 leaves."

How are the NN-HMM configured? What's the input? How many hidden layers? How are the activation functions chosen?

The following text has been added to the new section VII.A: "Each NN-HMM used six hidden layers with logistic nonlinearities, and with 1024 nodes per hidden layer, followed by a softmax output layer with 1200 nodes."

Is it possible to also present WERs along with PERs?

The following text has been added to the new section VII.B: "Phone error rates are reported instead of word error rates because, in order to compute a word error rate, it is necessary to have either native transcriptions in the target language (thereby permitting the training of a grapheme-based recognizer) or a pronunciation lexicon in the target language. These resources are used by the monolingual topline, but not by any of the baseline or experimental systems."

On EEG. This is an interesting piece of side information for acoustic modeling. But from what is reported, its impact seems to be quite marginal. To my understanding, it is only used in the interpolation in Table III or am I missing something here? Would appreciate it if the authors can elaborate a little bit more on its impact to ASR.

The reviewer is correct: the results of EEG are only demonstrated in Table III, and are not used at all for ASR training. The article demonstrates (1) that EEG may be used to improve PTs (the new Section VI), (2) that PTs may be used to train ASR (the new Section VII) --- both of these findings are new, and have not been previously published in any journal article (the former result has never previously been published in any paper anywhere; the latter has only been published in the ICASSP paper by Liu et al. cited in the bibliography). We intend eventually to connect these two ideas, but we expect this task to require another two or three years, because EEG is hard to scale. This fact is now explicitly acknowledged by the new structure of sections V, VI, and VII, and by the following text in the article: “Probabilistic transcripts based on EEG were not used to adapt ASR, because it is not yet possible to use EEG to generate probabilistic transcripts on a scale sufficient for ASR adaptation.”

Reviewer 2

Why should we use IPA phone set to achieve model parameter sharing? Looking at results reported in Tables III and IV, IPA universal phone set based model training yields much worse results. Besides, IPA universal phone set based multilingual training is ineffective, leading to worse results compared with language dependent phone set system. What if we use language dependent grapheme letter as phone set? Grapheme based lexicon should be much simpler in terms of building ASR system for low-resource language. Besides, it can also realize model parameter sharing, if we have a transducer system similar to EEG. Grapheme system is also effective. In Babel program, Swahili grapheme ASR system can produce comparable results with the ASR system built with expertise lexicon.

The new section V now includes the following text: “In order to make it possible to transfer ASR from training languages (which have native transcripts) to a test language (that has no native transcripts), the phone set must be standardized across all languages; for this purpose, the phone set was based on the international phonetic alphabet (IPA).” This thought is continued by the following text from the new Section VII: “Phone error rates are reported instead of word error rates because, in order to compute a word error rate, it is necessary to have either native transcriptions in the target language (thereby permitting the training of a grapheme-based recognizer) or a pronunciation lexicon in the target language. These resources are used by the monolingual topline, but not by any of the baseline or experimental systems.”

In addition to introducing IPA phone set, the paper has widely used G2P models, not only for phone language modeling, but also for misperception probability estimate. See Figures 1 and 4. This is not practical. As we know G2P model training needs a lot of supervised lexicon data that should be prepared with expertise knowledge. This contradicts the low-resource condition assumption. Besides, G2P can introduce errors that has never been mentioned in the paper.

There are two different sets of G2Ps per language: one that requires training resources, and one that does not require training resources. The one used to construct the phone bigram language model does not require any test language resources, as described in this new text in Section IV.C: “Because this phone bigram will also be used in ASR testing, it is constructed using a knowledge-based method that requires zero test-language training data: The G2P is constructed by looking up ``Swahili alphabet'' on Wikipedia, downloading the resulting web page, and converting it by hand into an unweighted finite state transducer.”

The multilingual misperception model (the one actually used in ASR experiments) does not require any G2P in the test language, but the EEG-based misperception model does. This point is now specified in Section VI: “Both method (2) and method (3) required the use of a G2P: the Dutch G2P was estimated using the CELEX database, while the Hindi G2P was estimated using the zero-resource knowledge-based method described in Sec. IV.C.”

For all other purposes, G2Ps were avoided in the test language by the simple expedient of assuming that there are no native language transcriptions in the test language. Section VI now specifies: “Similarly, in order to transfer ASR from training languages to a test language, the training transcriptions must be converted to phonemes using a grapheme-to-phoneme transducer (G2P). G2Ps were therefore assumed to be available in all training languages, but not in the test language. Since these G2Ps are only used for training and not test languages, five of them (Arabic, Dutch, Hungarian, Cantonese and Mandarin) were trained using lexical resources, and only two (Urdu and Swahili) were constructed using the zero-resource knowledge-based method described in Sec. IV.C.”

The paper uses EEG to estimate misperception probability, however its effectiveness has not been clearly demonstrated, since no baseline has been provided. I guess, data-driven based machine translation method can be adopted to do the same thing, correct? Shall we compare the efficacy of the two methods?

> The new Section III.B now includes the text "The model learned in this way is essentially a machine translation model, which translates between graphemes in the annotation language to phonemes in any possible utterance language." This baseline is explicitly compared to the EEG-induced misperception model in this text from Section VI: "In order to evaluate the effectiveness of the EEG-induced misperception transducer we looked at the LPER of mismatched crowdsourcing for Dutch when performed using 1) a multilingual misperception model (the machine translation model described in Sec. III.B), 2) feature-based misperception transducer computed using binary weighting, or 3) EEG-induced transducer combined with the feature-based transducer. LPER of the multilingual model was 70.43% (as shown in Table I), of the feature-based model, 69.44%, and of the EEG-interpolated model, 68.61%."

How can we do low-resource speech recognition? I think we should sufficiently take advantage of diversified rich-resource multilingual language data and unlabeled target language data simultaneously (I agree there are lots of low-resource languages, but I also agree there should exist a lot of rich-resource languages as well). Looking into the paper, we find the multilingual data is extremely tiny, about an hour.

> The cross-lingual system was trained using 240 minutes of data: 40 minutes per source language. The limitation on source language data is only important for symmetry of the experiment: it permits each source language to also be treated as a test language, in a leave-one-out paradigm.

Besides, it used IPA phone set for each source language to train multilingual DNN, which yields much worse results than the case with language-dependent phone sets.

> IPA acoustic models are necessary, we believe, in order to allow phone recognition in a test language in which the test-set ASR has (1) no native transcriptions, (2) no lexicon, and (3) no G2P.

Other potential rule that I can think of to develop low-resource ASR: the less human intervention, the better it is. However, the paper contradicts this principle. A lot of human power is used to transcribe the data, and a lot of human power is involved in EEG experiments to estimate the probability of phone transducer. I guess if people use a sharper grapheme based ASR recognizer to transcribe a target language, then use machine translation method to learn the mapping rules between source and target graphemes of languages, it should be using less human intervention.

> Section VIII now includes the text: "The primary conclusion of this article is economic. In most of the languages of the world, it is impossible to recruit native transcribers on any verified on-line labor market (e.g., crowdsourcing). Without on-line verification, native transcriptions can only be acquired by in-person negotiation; in practice, this has meant that native transcriptions are acquired only for languages targeted by large government programs. Native transcription (NT) permits one to train an ASR with PER of 31.58% (average, first column of Table V). Self-training (ST), by contrast, costs very little, and benefits little: average PER is 62.75% (Table V). Probabilistic transcription (PT) is a point intermediate between NT and ST: average PER is 52.29%, cost is typically $500 per ten transcribers per hour of audio. PT is therefore a method within the budget of an individual researcher. We expect that an individual researcher with access to a native population will wish to combine NT (as many hours as she can convince her informants to provide) with PT (on perhaps a much larger scale); future research will study the best strategy for combining these sources of information if both are available.

The paper uses PT to adapt cross-lingual (multilingual trained) DNN. However, little knowledge about how the PT affects the performance is known except for Figure 7. Even in Figure 7, it is not straightforward to understand. For instance, phone alternatives in each ``sausage" on average versus your oracle LPER is not revealed?

The effect of PT on performance of the DNN is measured in Table V. PT adapted DNN-HMM has a 44.73-56.70% PER, compared to 58.76-64.90% for a typical self-trained system, thus the drop in phone error rate is about one sixth.

Figure 7 is not about DNN, it is about the probabilistic transcription derived from mismatched crowdsourcing. The text includes the following description of the content of Fig. 7: “LPER of the 1-best path does not accurately reflect the extent of information in the PTs that can be leveraged during ASR adaptation. Consider, for example, the four Urdu phones [p^h,p,b^h,b]. An attentive English-speaking transcriber must choose between the two letters <p,b> in order to represent any of these four phones. The misperception G2P therefore maps the letters <p,b> into a distribution over the phones [p^h,p,b^h,b]. There is no reason to expect that the maximizer of $\rho(\varphi|\lambda)$ is correct, but there is good reason to expect the correct answer to be a member of a short N-best list. A fuller picture is therefore obtained by pruning the PT to a small number of paths, then searching for the most correct path in the pruned PT. One useful metric is entropy per segment… Fig. 7 shows the trend of LPER (for three languages) obtained by pruning the PT at several different levels of entropy per segment. LPER rates drop significantly across all languages within 1 bit of entropy per phone, illustrating the extent of information captured by the PTs.”

In the last column of Table IV, do you use phone lattice or one-best sequence to tune your DNN? If you use phone lattice, you should provide one-best results as your baseline to show the benefit of using PT.

Section IV includes the following text: “the best path through the PT, and the best alignment of the resulting senones to the waveform, were both computed using forced alignment. The resulting best senone string was used to train a NN.” Results shown in Table IV are also 1-best results.

In VII discussion section, you said you one-best phone sequence has 29-49% PER. From my experience, even this ``worse" hypotheses still can yield improved results, given some simple data selection method in Kaldi. Besides, you should give your oracle results if you use true phone sequence to tune your DNN. Not only will this show your effectiveness of PT training, but also it will show the effectiveness of your multilingual training. Normally, we would expect it will yield better results compared with results in the first column.

Yes, even this “worse” hypothesis certainly improves results. The following sentence has been added to Section VII to make this point clear: “Table IV shows that PT adaptation improves the NN-HMM, but the benefit to a NN-HMM is not as great as the benefit to a GMM-HMM; for this reason, the accuracy of the PT-adapted GMM-HMM catches up to that of the NN-HMM.”

Reviewer 3

II.D: the authors always talk about "the listeners" and "their responses". In the experimental section, it turns out that there was only "one monolingual" listener. It would be interesting to have at least two listeners, to get an idea if the EEG results are generalisable across speakers. Either the authors should add a second listener or correct the text (also abstract) and state "the listener" or "a listener".

Text has been changed to refer to a singular listener.

III.A: did the authors use any confidence score during the self-training?

The following sentence was added: “As in previous work, senones with a posterior probability below 0.7 were removed from the training set, thus the training target was a number between 0.7 and 1.0.

III.A / IV.E: for self-training real-valued targets perform better, but for the PT-NN, forced alignment is preferable? Are the differences between real-valued targets and forced alignment significant? In case of yes, do the authors have an idea why this is the case?

Yes, real-valued targets were better for the NN (in one experiment using the dev set for Swahili), but forced alignment was better for the PT-NN (in one experiment using the dev set for Swahili). Differences were small, and were not tested for significance, but are probably not significant.

III.B: how do the authors estimate the phone prior with a bigram phone language model?

This point has been expanded as follows: “$\rho(\varphi)$ is modeled using either a cross-lingual phone unigram, a language-specific phone zero-gram (the cross-lingual unigram, constrained to take values from the phone set of the target language), or a language-specific phone bigram $\rho(\varphi) = \prod_{m=1}^{M} \rho(\varphi_m|\varphi_{m-1})$. Sec. IV-C describes an algorithm for training the phone bigram without using proscribed test-language resources; Sec. V lists the PT accuracies achieved using each of these three approaches.”

Related to that question, in IV.C: how is the G2P of the under-resourced language obtained/trained?

The following sentence was added: “The G2P is constructed by looking up ``Swahili alphabet'' on wikipedia, downloading the resulting web page, and converting it by hand into an unweighted finite state transducer.”

Sections V and VI seem less well-structured than the first sections of the paper. Probably it would make sense to move the baseline section V.D directly into section VI.C.

These two sections have been restructured into three sections. All results addressing the use of mismatched crowdsourcing to improve PTs are in Section V, all results addressing the use of EEG to improve PTs are in Section VI, and all results of experiments involving ASR are in Section VII.

At the moment, it is not very clear how Table III and numbers at the end of VI.B should be compared. The baseline of Dutch with phone bigram (68.61) is the same as the EEG-interpolated model?

Yes, this number in Table III was a listing of the best score achieved using the bigram phone LM, which was the score achieved using the EEG-derived mismatch model. However, it seems that the EEG-derived mismatch model was only applied to that one entry in the Table; all other entries in the Table use the multilingual misperception model. We consider this to be an error, so that entry in the Table III has been changed to contain the number resulting from the multilingual model.

In the baselines section, the authors talk about oracle experiments. It seems that they are only present in Table IV. It may help to add them to Table III as well. Section V does not talk about the neural network structure at all. The authors may want to add some information about the architecture of the NN that was used.

The word “oracle” has been removed from the manuscript, as it is insufficiently precise. The new Section VII talks instead about “monolingual topline” speech recognizers, and specifies exactly the way in which “monolingual topline” systems have more information than the experimental systems: “Native transcriptions in the target language were used in order to train the monolingual topline system.” Table III addresses the improvement of PTs, not of ASR, so it’s not clear that there is any comparable equivalent topline system that could be applied to Table III.

Minor detail: can the authors say something about where the 24414Hz come from in V.C?

This information has been added to the manuscript: “…downsampled to 24414 Hz (for compatibility with the presentation hardware, Tucker Davis Technologies RP2.1).”

In the analysis of VI.A, the entropy estimates seem not very clear. I.e. how is an entropy of 0.5 bits per segment achieved?

The following more precise definition has been added: “A fuller picture is therefore obtained by pruning the PT to a small number of paths, then searching for the most correct path in the pruned PT. One useful metric is entropy per segment, defined as $H^{\ell}(\Phi) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{u}\log_2 \rho_{\Phi_m}(u)$, e.g., a PT in which every segment has two equally probable options would measure $H^{\ell}(\Phi)$.

VI.B: the system was trained on English. The EER for English seem relatively high for the reader who is not familiar with EEG signal processing. Should such a result be expected?

The following text has been added: EER of the classifier when applied to English phones is comparable to those reported in [9], the only prior work to attempt a recognition of speech phonemes from EEG of the listener.

In the last paragraph: how should the numbers be compared with Table III. What was the optimal value for alpha? Are the differences between the three numbers significant?

The numbers in Table III use the multilingual model, which is the baseline here. These differences are not statistically significant, because the number of tokens we were able to acquire in this experiment is not large enough to make this difference significant. The following text has been added: “The constant $\alpha=0.29$ was chosen as the average of the values selected in all folds of a leave-one-out cross-validation.”

VI.C: Was the neural network trained from scratch on the 40 minutes or was an already trained NN adapted?

The following text has been added to the article: “GMM parameters were initialized using a monophone system trained on the same 40 minutes, NN parameters were initialized using a restricted Boltzmann machine trained on five hours of unlabeled audio in the same language.

Table IV in general: In the abstract and introduction, the authors always separate mismatched crowdsourcing and EEG distribution coding. "Adaptation using mismatched crowdsourcing significantly outperformed self-training": can that statements be seen in Table IV, i.e. is the effect of mismatched crowdsourcing "alone" visible?

The following text has been added to the article: “The system was initialized using the CL system (ML training), then adapted to the target language using probabilistic transcripts based on mismatched crowdsourcing. Probabilistic transcripts based on EEG were not used to adapt ASR, because it is not yet possible to use EEG to generate probabilistic transcripts on a scale sufficient for ASR adaptation.

VII: "In PT adaptation, however, entropy is unavoidable, and quantizing the forced alignment doesn’t necessarily help." in IV.E, the authors say ... forced alignment also improves the accuracy...

The phrase in question has been changed to: “Forced alignment is better than using soft alignment, but is not sufficient to make PT adaptation of the NN-HMM always better than that of the GMM-HMM.”

"Table III showed that the 1-best path through the PT is only correct for 29-49% of all phones, depending on language". Could the authors explain where these numbers come from, and how to read that from Table III?

This sentence has been made more precise: “Table III showed that PTs computed using a text-based phone bigram language model only achieve LPER in the range 50.45-70.88%, depending on the language.”