============================================================================  
                            REVIEWER #1  
============================================================================   
  
  
---------------------------------------------------------------------------  
Comments  
---------------------------------------------------------------------------  
  
Summary of paper:  
This paper describes multilingual automatic speech recognition experiments for  
three African Languages (Amharic, Dinka and Swahili) using crowdsourced  
probabilistic transcriptions. The authors used HMM and DNN modeling tools in  
their experiment.  
  
Strengths of paper:  
Since the paper talks about crowdsourcing, it is interesting for researchers  
and developers in the area pf automatic speech recognition for under-resourced  
languages. The paper is well organized and written clearly. Elaboration on  
terminologies used in the paper are given which makes the paper easily  
understandable.   
  
Weaknesses of paper:  
However, the work is not without limitations. In most of the cases the authors  
do not answer the why question in the paper. They just describe what they did  
and how they did it, but description on why they did something is lacking in  
most of the cases.  
  
Detailed review:  
Explanation and/or clarification on the following issues is essential.  
The authors indicated the fact that they have used a multilingual audio files  
of 8 languages from SBN which include 1000 hours of speech in 68 languages. But  
no explanation is given on why they have considered only the 8 languages. In  
addition, only 3 of the languages are considered as target language again  
without providing any justification as to why they chose the three languages.  
An HMM based language identification system has been used the performance of  
which is not known. Brief explanation of the language identification system and  
its performance is worth reporting.  
  
The authors segmented the speech into smaller (5-second) chunks which might not  
be equivalent to a sentence in the target language. Was it not possible to  
segment the speech at sentence level? Why did you prefer to perform recognition  
at phone level? This approach made the use of word-trigram language models and  
word error rate difficult in the work.  
  
The authors did not use DTs in the target language although it is available and  
they did not give any explanation why they excluded DTs in the target language.  
  
Moreover, the authors claimed that they have used data with zero DTs in target  
languages (Amharic, Dinka and Swahili) and with 2 hours of DTs in the source  
languages (Hungarian, Cantonese, Mandarin, Arabic and Urdu), 40 minutes of  
speech per language. However, 5\*40 gives us about 3 and half hours of DTs in  
the source languages. I think there is some kind of confusion here the DTs of  
the target languages will be equivalent to 2 hours (40 minutes per language)  
instead of that of the source languages.  
  
In the work, language dependent phones were merged into a multilingual phone  
set. However, since speech data is available for each of the languages under  
consideration, the authors could have treated language dependent phones  
separately which may increase the performance of the multilingual speech  
recognition system. That means phone models could be developed for language  
dependent phones.  
  
The paper indicated that, for Swahili, phone-based language models were built  
from text corpora available on the web. But no explanation is given on how the  
authors converted the graphemes of the text to phonemes. Moreover, the authors  
did not indicate the value of “n” in the ngram language models used in the  
experiments. Is it bigram? Or trigram? …  
  
In the paper, phone error rate is reported instead of word error rate (WER)  
which is the standard way of reporting speech recognition systems’  
performance. It is good that the authors provide WER so that one can easily  
make comparison of speech recognition systems developed using manual and  
automatic transcriptions.  
  
There are other works (not cited in this paper) on Amharic and Swahili speech  
recognition system conducted using crowd-sourced transcriptions. Reviewing  
existing works and showing how the current work is different is important and  
this need to be considered.  
  
============================================================================  
                            REVIEWER #2  
============================================================================   
  
  
---------------------------------------------------------------------------  
Comments  
---------------------------------------------------------------------------  
  
Summary of paper:   
In this paper, the authors build on their previous research into use of  
mismatched, crowd-sourced transcriptions to develop ASR systems in a  
low-resource context. The target languages in this study are Swahili, Amharic  
and Dinka. A single, multilingual recogniser is built for the three languages  
using probabilistic transcriptions (PTs) obtained by source language speakers  
with no knowledge of the target languages. Monolingual baseline models are  
built using deterministic transcriptions in the target languages in order to  
establish the lower bound of PERs for these languages. A multilingual model  
without any target language data is used to establish the upper bound, with the  
aim of the research being to use the PTs with a unified phoneset to train a  
system which approaches the lower bound PERs. The best-performing adapted  
system uses a combination of PTs, multilingual DTs (excluding target language  
data), and self-training transcripts which produce an additional set of high  
confidence labels from unlabelled audio for retraining the DNN.  
  
Strengths of paper:   
\*Clear and detailed exposition of the methodology and reasoning behind it  
\* Builds on a strong foundation of related work, particularly by the second and  
third author  
\* Clearly demonstrates that improvements in PERs established without reference  
to deterministic phone rates (i.e. in the absence of DT data in the testing  
stage) correlate well with improvements established with reference to  
deterministic rates.  
  
Weaknesses of paper:  
\* No obvious weaknesses, but a couple of questions:  
- It would be helpful to mention what PER threshold would be the goal  
ultimately: < 30%?   
- How would the authors propose to assess WER for the three languages in  
future? Given the elimination of certain phonemically distinctive,  
language-specific phones from the unified phone set (e.g. ejective vs aspirated  
stops in Amharic), mappings of unified phone sequences to target words are  
likely to be significantly under-determined.  
  
Detailed review: This paper is an important contribution to the growing body of  
work supporting language technology for low-resource languages. One does wonder  
about the potential for use of near-language Africa-based crowd workers (or  
other near-languages for other low-resource languages); there are many writers  
of languages such as Swahili and Amharic based in large, well-connected urban  
centres such as Nairobi and Addis Ababa, who might be drawn on for the sourcing  
of closer PTs. Has this been considered?  
  
============================================================================  
                            REVIEWER #3  
============================================================================   
  
  
---------------------------------------------------------------------------  
Comments  
---------------------------------------------------------------------------  
  
Summary of paper:  
  
This paper describes the development of ASR systems for three sub-Saharan  
African languages (Swahili, Amharic, and Dinka) using inaccurate (approximate)  
transcriptions collected from crowd workers who do not know these languages.  
Probabilistic transcriptions derived from multiple versions of approximate  
transcriptions were used to train the ASR systems. The proposed recipe for low  
resource languages includes: mismatched crowdsourcing, multilingual adaptation  
of GMM and DNN acoustic models using the probabilistic transcriptions (PTs).  
Some phonetic description of the considered languages is given and details on  
phoneme mapping between languages are reported. In addition, a probabilistic  
phone error (PPER) metric is proposed to measure the performance of ASR system  
in absence of deterministic transcripts in the test set.  
  
Strengths of paper:  
  
1. The topic of the paper is relevant and interesting for audience.  
2. A new metric (PPER) for evaluating ASR system on a phoneme level was  
proposed for the case when the exact transcripts for test set are unavailable.   
It was shown that improvements in PPER have similar tendency in behavior as the  
improvements in PER. (However for proving correlation between these two  
measures a more detailed statistical analysis is required.)   
  
Weaknesses of paper:  
1. In this paper several approaches proposed in other papers were used to  
improve the performance of the ASR system for the considered languages, but no  
novel approach for solving the problem has been presented.  
2. The baseline for comparison (the upper bound for PER, as it is called in the  
paper) was chosen as a multilingual model with deterministic transcripts (DTs),  
trained on the source languages, without using training data of the target  
languages. And the recognition results for this model were much worse than the  
results of monolingual training with DTs. This raises the following question.  
Why there are no results with monolingual models, trained only on the data from  
the target languages with PTs (for example, 1-best could be chosen for  
simplicity). This could show the quality of PTs as this result will allows us  
to compare results of acoustic models trained on the same acoustic data, but  
with different types of transcripts.  
On the other hand, the lower bound for PER (i.e., the best possible result) is  
chosen as a monolingual model trained only on the target language with DTs. But  
from other works, for example [15], it is known, that DNN acoustic models using  
multilingual acoustic data can perform better than monolingual DNNs. Hence, the  
question is: why there is no results for DNN acoustic models with multilingual  
shared hidden layers (as described in Section 3.5), but with DTs for target  
language. This probably could provide more reasonable lower bound PER baseline  
in the sense, that first, it could shed light on the quality of PTs (in  
comparison with DTs), and second, give better results in terms of PER  
(establish lower bound for PER).  
3. There is some inaccuracy (for example, “HMM / GMM”, see Detailed review,  
point 1.) and incompleteness (see Detailed review, point 4.)  in the  
description of experiments.  
  
Detailed review:  
1. The incorrect (inaccurate) usage of terminology: “GMM”, “HMM” takes  
place throughout all the paper. Two types of acoustic models are used in  
experiments: GMM-HMM and DNN-HMM, but in the paper, abbreviation “HMM” is  
used for “GMM-HMM” model and abbreviation “DNN” is used for  
“DNN-HMM” model. For example:   
“3.3. Multilingual HMM and DNN”  
“Table 3: PERs of monolingual HMM and DNN models.”  
“3.2. Monolingual HMM and DNN”  
…..  
2. “The forced alignments obtained from the LDA+MLLT model were further used  
for speaker adaptive training (SAT) by computing feature-space maximum  
likelihood linear regression (fMLLR) transforms [9] per subset of speakers.”   
The question refers to the phrase “(fMLLR) transforms [9] per subset of  
speakers”. Does it mean that before applying fMLLR, the speaker clustering  
was performed? If so, could you please give some details on this procedure.  
3. It is not clear from the paper, what amount of data of source languages was  
used for multilingual training. On the one hand, it is written: “The DTs are  
worth about 40 minutes of audio per language. Hence, the total amount of  
multilingual DTs available for training is 2 hours. (40 minutes/language× #  
languages)”.        But on the other hand, it is written that there were 5 source  
languages (Hungarian, Cantonese, Mandarin, Arabic, Urdu), hence the total  
amount should be 5\*40=200 minutes ~ 3.3 hours.  
4. In Section 3.2, in the description of acoustic model training, there is no  
information about the type of input features for DNN models. Also there is no  
information about the number of Gaussians in the GMM models.   
5. In Sections 3.3 it is written: “The steps for building HMM and DNN systems  
were the same as in Section 3.2 except that the training data consists of  
multilingual DTs.”  Does it mean, that the number of parameters for GMM and  
DNN models in monolingual (Sections 3.2) and multilingual (Sections 3.3) cases  
were the same or it was optimized? The amount of acoustic data in monolingual  
case is 5 times less than in multilingual, so probably monolingual models need  
less number of parameters.   
6. In Section 3.6 a probabilistic metric PPER is proposed. To calculate it,  
“the edit distance between the 1-best path in the ASR decoding lattice and  
the unweighted pruned PT is computed”. Could you please comment on the  
algorithm used in this procedure, because unweighted pruned PT is a CN?   
7. In Section 3.5 a new configuration for DNN training is proposed (shortly  
described) – Figure 3.c, but the results for DNNs of this type are not  
reported in the paper.  
8. Sections 1 and 3.1 give the impression that DTs transcripts for training  
corpus for the target languages are not available at all in all experiments:   
-“DTs only in Source Languages: There are DTs from other source languages  
(6=L).”  
-“… since we assume there are no DTs in the target language.”  
However, in Section 3.2 it appeared that these type of transcripts are  
available and are used to build an “ideal” baseline system.  It is not a  
serious problem, and finally it is clear what the authors meant, but it is a  
bit confusing when you read the paper first time.  
  
============================================================================  
                            REVIEWER #4  
============================================================================   
  
  
---------------------------------------------------------------------------  
Comments  
---------------------------------------------------------------------------  
  
Summary of paper:  
  
The authors present ASR system developments for 3 African languages: Swahili,  
Amharic, Dinka. By adding low quality manually transcribed speech data of these  
3 languages to adapt a multilingual system, transcription accuracy is improved.  
The intriguing point is that the manual transcripts are produced by people who  
ignore the languages they transcribe: this recently proposed approach is called  
"mismatched crowdsourcing".  
The paper thus demonstrates that ASR accuracy can be improved using a  
relatively small set of erroneously transcribed language-specific data (40  
minutes per language).  
  
Strengths of paper:  
  
The authors experiment with the recently developed "mismatched crowdsourcing"  
setup, to improve ASR systems (multilingual, adapted MAP-HMM, adapted DNN) in 3  
African languages. Results consistenly show ASR accuracy improvements both for  
adapted MAP and DNN-1 (achieving equivalent results). The DNN-2 condition  
further improves results due to an increased language-specifc adaptation data  
(subsets of automatically transcribed data with highest confidence)  
  
Overall, the paper is very interesting and well written.  
  
Weaknesses of paper:  
  
- better motivate the scientific interest of using  "mismatched crowdsourcing"  
- better describe/motivate your approach as compared to unsupervised and  
cross-lingual training approaches, which do not require human transcripts at  
all.  
- the "quality" (error rates) of the human (PT) transcripts could be presented  
(as there are some data with both PT/DT)  
  
Detailed review:  
  
The manual transcript collection setup sounds weird:  
manual transcriptions to be made by people who completely ignore the language  
they transcribe. Of course, the interest of this setup may be obvious from a  
practical and economical point of view, however, please also relate your  
choices to scientific and technical issues. Please, explain the rationale  
behind this in the introduction - this is well done in other papers on this  
subject.  
  
Another question arising when reading the introduction: how to transcribe an  
unknown language in English? some pseudo-English?  
to clarify, please give an example (spectrogram + aligned transcript).  
  
\* Is this setup meant to mimic automatic phone recognition? Why not explore the  
latter in parallel? Fig.2 typically reminds the outcome of a phonetic speech  
recognizer.  
  
The linguistic description of the 3 languages is unbalanced:  
almost nothing for Swahili and Amharic, lots for Dinka.  
PLEASE give a more balanced, synthetic overview of the 3 languages,  
highlighting those aspects which are most relevant w.r.t. PER measures.   
  
The use of G2P seems particularly error-prone for English, as this language has  
no "deterministic" G2P rules.  
- are PT only due to a merge of multiple manual, diverging, "deterministic"  
transcripts?  
- may G2P produce probabilistic transcripts on 1 single chunk?   
  please give an approximate measure of the quality of the PT production  
process (some inter-annotator agreement, some distance measure between produced  
letter or phone strings)   
  
p.2 the authors mention "...defining equivalence classes for similar  
sounds...".  
  what is "similar" sound?  
  PLEASE: give examples of smallest, longest equivalence classes.  
  PLEASE: indicate the total number of phone/sound classes your system is  
modelling (82?).  
  
\* in section 3, the system training could be clarified:  
maybe introduce an independent DATA section describing        PT/DT production before  
training.  
  
\* Where are DEV data in Table 2? to be added!  
Please clarify the role of development data, DT data in particular.  
  
\* There are only 53 Dinka test chunks (120 expected) - why?  
PLEASE add more information abaout Table 2 in the caption AND in the text.  
PLEASE specify how you compute language-specific PER in th multi-DNN condition?  
mapping these 82 symbols to the language dependent set?  
  
Please, improve captions of Table 2 and Figure 3 (what is ST).  
  
In the DNN-2 system experiment,   
"... the DNN-2 system decodes some additional unlabeled audio in the  
target language and then uses a subset of the decoded labels,  
with high posterior probabilities (confidences), to retrain itself  
in the target language."  
please specify the volume (duration) of "additional unlabeled audio" per  
language, the volume of selected labels (duration and number of labels) added  
per language.  
  
caption of Table 6 is below the table (the others are above).  
  
Please clarify the training/adaptation/role of "language models" for each  
language.