MGB-3 BUT SYSTEM: LOW-RESOURCE ASR ON EGYPTIAN YOUTUBE DATA

Karel Veselý, Baskar Karthick Murali, Mireia Diez, Karel Beneš

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czech Republic
iveselyk@fit.vutbr.cz

ABSTRACT

This paper presents a series of experiments we performed during our work on the MGB-3 evaluations. We both describe the submitted system, as well as the post-evaluation analysis. Our initial BLSTM-HMM system was trained on 250 hours of MGB-2 data (Al-Jazeera), it was adapted with 5 hours of Egyptian data (YouTube). We included such techniques as diarization, n-gram language model adaptation, speed perturbation of the adaptation data, and the use of all 4 'correct' references. The 4 references were either used for supervision with a 'confusion network', or we included each sentence 4x with the transcripts from all the annotators. Then, it was also helpful to blend the augmented MGB-3 adaptation data with 15 hours of MGB-2 data. Although we did not rank with our single system among the best teams in the evaluations, we believe that our analysis will be highly interesting not only for the other MGB-3 challenge participants.

Index Terms— MGB-3, ASR adaptation, low-resource ASR, Egyptian Arabic, diarization

1. INTRODUCTION

The MGB-3 evaluation [1] consists of two tasks, one is building automatic speech recognition (ASR) system a for low-resource target domain, while the other is dialect identification. We participated in the ASR task and in this paper, we describe the submitted system together with the experiments we did 'on the way'.

Recently, the low-resource ASR was one of the key specifics of the Babel program in OP2 period [2, 3]. During our participation, we have discovered that a cross-lingual transfer of acoustic models is possible and useful: An *initial model* is trained on a large corpus, which can be either mono-lingual [4] or multi-lingual [5]. Then, even with as low amount of the target domain data as 3 hours, we were able to reach the keyword spotting ATWV goal 0.3, which represented a 'practically usable' keyword search. This would not be possible without the pre-trained initial models.

A similar scenario appears also is in the MGB-3 evaluations, in which we have 1200 hours of Al-Jazeera archive available for the initial model construction [6]. The target domain is the Egyptian dialect of Arabic; for the system adaptation we have only 5 hours of speech [1].

Our experiments start by developing the initial system with BLSTM acoustic model. As the original development data contains no speaker information, we integrate diarization [7, 8] to boost the performance of the standard speaker-adaptation techniques in ASR (CMN, fMLLR).

Next, we focus on the ASR adaptation towards the target domain. At first, we perform the adaptation of the language model. Then we proceed to the core of the paper, the adaptation of the acoustic model. We show that improvements can be achieved by data engineering such as speed perturbation and blending the adaptation data with the data from source domain.

However, the most interesting part is dealing with the 4 alternative annotations for each utterance in the target domain, which has no standard orthography. We explored two principal strategies for exploiting this information: The straightforward one is to use them *serially*, i.e. include each sentence 4 times into the training data, once with each annotation. The other approach, denoted as *parallel*, consists of combining all the annotations into a confusion network. The NN training targets are then obtained as the Viterbi path through this network. The approach is similar to the recently published training with 'probabilistic transcripts' [9].

2. DATA DESCRIPTION

The MGB-3 challenge goal was to achieve the best ASR performance, while we have a low-resource database available. The MGB-3 database is compiled from YouTube shows in Egyptian Arabic dialect. We were also given a big corpus of partially mismatched data, an Al-Jazeera archive both in Modern Standard Arabic and dialectical data including some Egyptian Arabic (MGB-2 data).

The work was supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", European Union's Horizon 2020 project No. 645523 BISON, and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602". Mireia Diez was supported by European Union's Horizon 2020 Marie Sklodowska-Curie grant agreement No. 748097.

2.1. MGB-3 data for adaptation, YouTube shows (Egyptian Arabic)

The MGB-3 data were compiled from 80 YouTube shows covering nine genres: comedy, cooking, cultural, environment, family/kids, fashion, drama, sports, and science talks [1]. First 12 minutes were selected from each show and the non-speech segments were removed, which resulted in 16 hours of in-domain (Egyptian Arabic) speech data. These data were split into 3 subsets: 4.8 hours for adaptation, 4.8 hours for development, and 6.2 hours with non-public transcripts for the evaluation. A majority of the topics overlaps between the subsets (seven topics out of nine), and the YouTube shows contain on average more noise than the studio recorded MGB-2 data.

An important feature of Egyptian Arabic is that it has no standard orthography, hence we can have several correct, albeit different, transcriptions per utterance. To accommodate this fact into scoring, the organizers collected transcripts from 4 annotators and implemented a tool for calculating 'Multiple Reference Word Error Rate' (MR-WER) [1]. This tool takes the 'best-case' Levenshtein distance, while considering the mix of several reference transcripts.

For scoring, only the segments without overlapped speech were used. For the adaptation, we used both the overlapped and non-overlapped speech.

2.2. MGB-2 data for initial system, Al-Jazeera archive

This out-of-domain data is the same as provided for MGB-2 challenge [6]. It contains 1200 hours of Arabic speech with imperfect transcripts. Out of this data, 70 % is Modern Standard Arabic (MSA), while the rest is dialectical data covering Egyptian (EGY), Gulf (GLF), Levantine (LEV), and North African (NOR) dialects of Arabic.

The MGB-2 data were prepared from Al-Jazeera archive covering years 2005-2015; the recordings come from 19 distinct TV programs. The 3000 recordings are divided into categories: conversation (>2 speakers, 63 % of recordings), interview (2 speakers, 19 %) and report (1 speaker, 18 %). The data covers as diverse topics as politics (76 %), society (9 %), economy (8 %), media (3 %), law (2 %), and science (2 %).

3. INITIAL MULTI-DIALECT ARABIC SYSTEM (TRAINED ON MGB-2 DATA)

The baseline ASR system has a relatively simple structure: The acoustic model is a neural network composed of 3 BLSTM layers with linear projections [10]. The last BLSTM layer is followed by a tanh non-linearity and a softmax output layer with 3681 triphone tied-states. The BLSTM type is standard: it has input, output, forget gates, the peep-hole connections, and a linear projection on the output. The BLSTM dimensions are following (#inputs, #cells, #projection-outputs): (43,2x350,2x200), (400,2x350,2x200), and (400,2x350,2x320). The softmax layer has thus 640 inputs and 3681 outputs.

The whole network was trained with Stochastic Gradient Descent, in which the model updates were calculated from up to 24 whole sentences in parallel, having a maximum of 4000 frames per update. The training was done with the Kaldi nnet1 training tool `nnet-train-multistream-perutt`, which uses a single GPU and processes sentences of similar lengths simultaneously.

For the supervision in the frame classification BLSTM training, we used alignments from an fMLLR-GMM system. Then, we performed 1 epoch of sMBR training with a fixed learning-rate, in this case the updates were done per single sentence. This acoustic model was trained on the first 500 shows in the MGB-2 training dataset, resulting in 250 hours of audio data.

The input features are 40 dimensional 16kHz log-Mel filterbank features with dithering, extended by 3 Kaldi pitch features [11]. We applied a per-speaker mean normalization and a global variance normalization on all these input features. The frame rate was 10 ms and no temporal splicing was used on the BLSTM input.

The Kaldi decoder uses an HCLG recognition network for generating the hypothesis/lattice from the cross-word tied-state HMMs [12]. The language model was interpolated by SRILM from two tri-gram models: a) a model from all the Al-Jazeera training transcripts (1200 h), b) a model from the 600MB corpus of Arabic texts provided by MGB-3 organizers. The interpolation weights were tuned on initial 10k sentences from the Al-Jazeera training transcripts. The Buckwalter graphemic lexicon was limited to 268k word-symbols, which we obtained by merging the 200k most frequent tokens from the two corpora.

The performance of the initial system is summarized in Table 1. We have obtained a WER 26.3 % for the sMBR BLSTM model on non-overlapped speech.

Table 1: The initial system with and without diarization. BLSTM trained on 250 hours of Al-Jazeera Arabic data, the WER is measured on Arabic development set (MGB-2).

MGB-2 dev (Arabic)	% WER	
	Non-overlap	Overlap
No diarization,		
FMLLR-GMM (tri4)	40.2	77.1
BLSTM CE	28.7	73.6
BLSTM sMBR	26.3	70.7
With diarization,		
FMLLR-GMM (tri4)	37.3	76.5
BLSTM CE	28.1	72.7
BLSTM sMBR	25.9	70.6

4. DIARIZATION

As the original development and test data contain no speaker information (neither MGB-2 nor MGB-3), we integrate diarization to identify the dominant speakers in each utterance.

The used speaker diarization method is based on the Variational Bayes (VB) method described in [7, 8], which makes use of pre-trained eigenvoice bases to facilitate discrimination between speakers.

In order to perform speaker clustering, the original method [7, 8] requires the input speech to be pre-segmented into (preferably) speaker homogeneous segments. The diarization algorithm then iterates between the usual two steps 1) speech segments clustering and 2) re-segmentation.

In our implementation, however, we perform the speaker clustering on frame-by-frame basis, making the re-segmentation step unnecessary. In order to avoid too frequent speaker turns, we represent speakers by HMM states and set transitions probabilities to favor staying in the same speakers.

We used 19 MFCC+Energy coefficients (without any normalization) as features for diarization. We only ran the diarization on segments that contain speech according to our VAD. We used 1024-component, diagonal covariance GMM UBM, and a factor loading matrix with 400 eigenvoices (JFA V matrix). The UBM and the V matrix were trained on the 1200h Arabic training dataset (MGB-2). An agglomerative clustering based on cosine distance between i-vectors estimated on 200 ms segments was performed to initialize the labels for the VB algorithm.

From the ASR point of view, the diarization is used to create the speaker to utterance mapping, which improves performance of ASR adaptation techniques. We keep the segment boundaries as provided by the challenge organizers. We can see in Table 1 that the diarization helped consistently in all the systems. This is because of an improved CMN and fMLLR in GMM systems, and better CMN in BLSTM systems.

5. ADAPTATION TO MGB-3 DATA

Next, we focus on the adaptation of the ASR system to the target domain of YouTube shows in Egyptian Arabic dialect.

5.1. Adaptation of LM (n-gram)

As the first step of adaptation to MGB-3 data, we updated the language model. We added a 3rd trigram model into the interpolation and we tuned interpolation weights on the MGB-3 development set. This 3rd language model was built from the adaptation transcripts, and we concatenated the transcripts from all the 4 annotators.

In total, we used 3 LM data sources: a) transcripts from 1200 hours of MGB-2 data [6], b) 600 MB 'LanguageModelText' from the challenge organizers and c) transcripts of 4x 4.8h of Egyptian. We pruned the LM from 'b)' data and the optimal LM interpolation weights were (0.05, 0.40, 0.55).

The word-lists from 'a)' and 'b)' were limited to 200k most frequent tokens. From 'c)' we added all the words. This extended the original graphemic lexicon by 10k new words.

Table 2: LM adaptation to MGB3 data, non-overlapped speech, BLSTM-sMBR acoustic model.

[% WER]	Unadapted LM	Adapted LM
MGB2 dev (Arabic)	25.9	27.2
MGB3 dev (Egyptian)	69.7	67.2

In Table 2 we see that the LM adaptation helped the system better fit the MGB-3 data (WER reduction by 2.5 % abs.). At the same time, the adaptation caused a slight performance deterioration on the MGB-2 data (WER increased by 1.3 % abs.).

Even though the initial LM was built on multi-dialect Al-Jazeera data (including some Egyptian Arabic), the LM adaptation still brings a solid performance improvement.

5.2. Adaptation of acoustic model (BLSTM), CE training

Next, we begin to adapt the acoustic model. When working with a single annotator, we choose Mohammad's transcript as a ground truth reference for the adaptation. Unless stated differently, the experiments are with the frame Cross-Entropy (CE) loss function.

5.2.1. Tuning the number of epochs with fixed learning-rate

In an initial experiment, the optimal learning rate was found the same as we had for the initial BLSTM training ($2.5 \cdot 10^{-6}$). Because the amount of the adaptation data is very limited (4.8 hours), we fix the learning rate (LR) for the first N epochs. After that, we begin the learning rate halving with the usual per-epoch schedule. We found the number of epochs with fixed LR to be an important parameter of the learning, hence most of the results are presented as a function of it.

Table 3: AM adaptation to MGB-3 data. The learning rate (LR) halving begins after N epochs with fixed LR (we have only 4.8 hours of adaptation data).

# epochs with fixed LR:	10	15	**20**	25	30
% WER (Mohammad)	59.8	59.5	**58.9**	59.0	59.4
% MR-WER (4 refs)	51.9	51.7	**51.0**	51.1	51.7

In Table 3 we see that the best BLSTM adaptation was obtained by fixing LR for 20 epochs. Recall that we previously obtained WER 67.2, when no AM adaptation was applied.

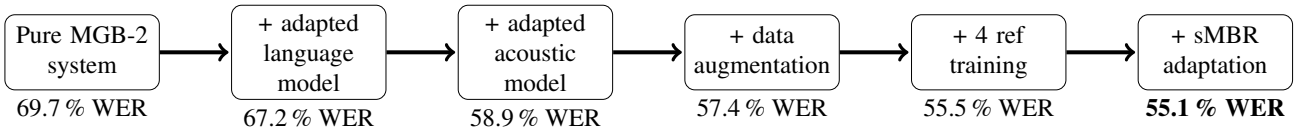

Fig. 1: Incremental adaptation of the system. At first, the BLSTM model was fully trained on 250 hours of multi-dialect MGB-2 data. Then, various ways of incorporating the Egyptian adaptation data (MGB-3) were added to the recipe, obtaining 55.1% WER (46.9 % MR-WER) with the final best system.

5.2.2. Augmenting the adaptation set

Then, to increase the amount of adaptation data, we applied speed perturbation with warping factors 0.9, 1.0, and 1.1, which triples the amount of data. In Table 4 we see that the best WER improved from 58.9 to 58.0.

Table 4: AM adaptation to MGB-3 data with speed perturbation (warps 0.9, 1.0, 1.1). We used the annotation from 'Mohammad' as the reference transcripts.

# epochs with fixed LR:	10	15	20	**25**	30
% WER (Mohammad)	58.2	58.1	58.2	**58.0**	58.3
% MR-WER (4 refs)	50.0	50.2	50.0	**49.9**	50.0

Then, we further increased the amount of adaptation data by adding another 15 hours, which we took from MGB-2 training data (Al-Jazeera). Like this we obtain a balanced adaptation dataset. In Table 5 we see that the WER further improved from 58.0 to 57.4. With this dataset, it seems that we no longer need to keep learning rate fixed; these results are not sensitive to fixing the learning rate within range 0–10, as we already have 'enough data' for reliable LR scheduling.

Table 5: AM adaptation to MGB-3 data, with speed perturbation on MGB-3 data, while we further add 15 hours of MGB-2 Arabic data. The learning rate (LR) halving begins after N epochs with fixed LR.

# epochs with fixed LR:	**0**	5	10	15	20
% WER (Mohammad)	**57.4**	57.4	57.5	57.8	57.8
% MR-WER (4 refs)	**49.4**	49.4	49.6	49.8	50.1

Here we can conclude: it was advantageous to triple the adaptation data with speed-perturbation in combination with balancing with the data from the source domain. These experiments were done with single reference (Mohammad), while we still have references from 3 more annotators available (Ali, Alaa, Omar).

5.3. Supervision from 4 annotators, CE training

As the Egyptian data have multiple 'correct' transcriptions, we were wondering how to use them to further improve our model. So, instead of using Mohammad as the sole transcription, we explored the two approaches, the 'serial' and the 'parallel' combination of transcripts.

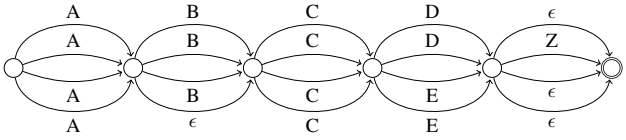

Fig. 2: Example of a training G transducer, which combines all four transcriptions (referred to as a *confusion network* in the text). This specific one corresponds to the output of the hierarchy of Levenshtein alignments from Figure 3.

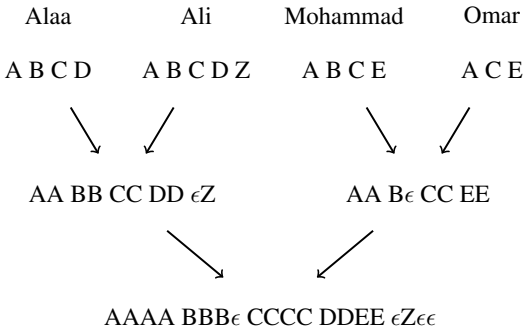

Fig. 3: Example of pair-wise joining of transcripts by a hierarchy of Levenshtein alignments for two strings of symbols. The symbol-pairs are aligned into the final sequence according to minimal edit distance. The average distance within the pairs (Alaa, Ali) and (Mohammad, Omar) is ca. 4 %, while the across-pair difference is ca. 20 %.

5.3.1. 'Serial' combination of annotations

The simplest way of exploiting multiple references is to include each sentence 4 times, once per each transcript from an annotator. This effectively does a 'serial combination' of the reference transcripts. In Table 6 we explored this 'serial' combination, both with and without the MGB-2 Arabic data added. We see that the MR-WER results improved from 49.4 to 48.4, while the MGB-2 data act as a regularizer (the performance is worse without the 'arab15h' data).

Table 6: AM adaptation to MGB-3 data, each adaptation utterance included 4x with all the annotations (adaptation set augmented by speed perturbation)

# epochs with fixed LR:	**0**	5	10
Data: adapt-warps-4refs-serial + arab15h,			
% WER (Mohammad)	**56.6**	56.6	57.3
% MR-WER (4 refs)	**48.4**	48.4	49.2
Data: adapt-warps-4refs-serial,			
% WER (Mohammad)	60.1	60.1	**59.6**
% MR-WER (4 refs)	51.7	51.7	**51.4**

5.3.2. 'Parallel' combination of annotations

The second approach is to combine the annotations into a single reference graph. Then, we can let the acoustic model decide which annotation version is preferable by a force-alignment. We call this method as 'parallel' combination, because we combine the annotations into a 'confusion network' like structure in which the words are in parallel.

To build the confusion network (illustrated in Figure 2), at first, we need to align the word-strings from the annotations, which inserts the epsilons 'ϵ' as 'empty' words. This is done by a hierarchy of Levenshtein alignments for two symbol-strings, as illustrated in Figure 3. Then, from the confusion network we created a training graph, and by a forced-alignment we produced the Viterbi path, which defines the targets for NN-training.

Table 7: AM adaptation to MGB-3 data, supervised by a *confusion network* from all 4 annotators (adaptation set augmented by speed perturbation).

# epochs with fixed LR:	**0**	5	8	10	15
Data: adapt-warps-4refs-parallel + arab15h,					
% WER (Mohammad)	**57.3**	57.3	57.4	57.4	57.6
% MR-WER (4 refs)	**49.0**	49.1	49.2	49.2	49.3
Data: adapt-warps-4refs-parallel					
% WER (Mohammad)	**55.5**	55.5	55.6	55.9	56.2
% MR-WER (4 refs)	**47.3**	47.3	47.3	47.5	47.7

As results in Table 7 suggest, this time, it was better not to include the 'arab15h' data. This is opposite to what was observed with 'serial' combination. Again, we did not need to fix the initial learning rate manually, but it got fixed 'naturally' by a smooth decrease of loss value. When compared to the 'serial' combination, the MR-WER improved dramatically from 48.4 to 47.3.

5.3.3. sMBR training in adaptation

After finishing the CE experiments, we continue with sMBR training from the two best systems trained with the 4 references. We are starting from one representative of the 'parallel' and another with the 'serial' combination.

Table 8: sMBR training starting from the best systems with 'serial' and 'parallel' combination of 4 references

	WER	MR-WER
Adapted system #1, 4-refs serial	56.6	48.4
+ sMBR adapt wrp., 4-refs serial	55.2	47.0
Adapted system #2, 4-refs parallel	55.5	47.3
+ sMBR adapt wrp., 4-refs parallel	55.1	46.9
Submitted system (primary)	55.0	47.3

The results in Table 8 are a bit puzzling. Although there was a significant difference in performance of the CE adapted systems #1 and #2, the gap practically disappeared after the sMBR training. Note that the 'parallel' and 'serial' references were used also in the sMBR training, while we re-generated the alignments with the respective adapted systems. Again, we used the speed-warps to augment the adaptation set 3x.

We also verified, that both sMBR systems have performance similar to the primary system we submitted. In our primary system we used our first prototype of 'parallel' combination, which we refined later. For this paper, we repeated most of the experiments from the evaluations. Thanks to Vimal Manohar who suggested it, we also added the comparison with the 'serial' combination of references.

6. CONCLUSION AND DISCUSSION

In this paper we described the adaptation strategies we used in the MGB-3 evaluations. At first we built an initial BLSTM system on 250 hours of MGB-2 data. We did not particularly strive to select the programs in Egyptian from the Al-Jazeera archive, even though this would very likely boost the performance. Also, the other evaluation participants probably used the full 1200 hour MGB-2 dataset, this seems to explain the $\approx$10 % gap of WER between our system and the three best teams. With this initial system, we demonstrated that diarization is beneficial in ASR for better speaker adaptation, as the WER improved from 40.2 to 37.3 with fMLLR GMM model, and from 26.3 to 25.9 with the sMBR-BLSTM model with CMN.

Then we step-by-step adapted the system to the MGB-3 target domain (YouTube shows in Egyptian Arabic). At first we adapted the language model, then we adapted the acoustic model by SGD re-training with CE objective. We achieved some extra performance improvements from data engineering (speed perturbation, balancing with MGB-2 training data).

And we achieved yet another performance improvements from using all the 4 versions of reference transcripts, either in 'serial' or in 'parallel' way. The overall progress of adaptation is shown in Figure 1, showing the total WER improvement from 69.7 down to 55.1.

We also briefly experimented with speech separation of overlapped speech by Deep Clustering [13]. However, for the moment, we did not achieve satisfactory results with decoding the separated speech of two speakers. It may be caused by a data mismatch as the ASR system was not trained on data processed by the separation algorithm.

7. REFERENCES

[1] Ahmed Ali, Stephan Vogel, and Steve Renals, "Speech recognition challenge in the wild: Arabic mgb-3," in *proceedings of ASRU 2017*.

[2] Le Zhang, Damianos Karakos, William Hartmann, Roger Hsiao, Richard M. Schwartz, and Stavros Tsakalidis, "Enhancing low resource keyword spotting with automatically retrieved web documents," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015.

[3] Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, Abhinav Sethy, Kartik Audhkhasi, Xiaodong Cui, Ellen Kislal, Lidia Mangu, Markus Nußbaum-Thom, Michael Picheny, Zoltán Tüske, Pavel Golik, Ralf Schlüter, Hermann Ney, Mark J. F. Gales, Kate M. Knill, Anton Ragni, Haipeng Wang, and Philip C. Woodland, "Multilingual representations for low resource speech recognition and keyword search," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, 2015.

[4] Frantisek Grézl and Martin Karafiát, "Boosting performance on low-resource languages by standard corpora: An analysis," in *2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, December 13-16, 2016*, 2016.

[5] Martin Karafiát, Murali Karthick Baskar, Pavel Matejka, Karel Veselý, Frantisek Grézl, and Jan Cernocký, "Multilingual BLSTM and speaker-specific vector adaptation in 2016 but babel system," in *2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, December 13-16, 2016*, 2016.

[6] Ahmed M. Ali, Peter Bell, James R. Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang, "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition," *CoRR*, vol. abs/1609.05625, 2016.

[7] Patrick Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," 2008.

[8] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[9] Mark A. Hasegawa-Johnson, Preethi Jyothi, Daniel McCloy, Majid Mirbagheri, Giovanni M. di Liberto, Amit Das, Bradley Ekin, Chunxi Liu, Vimal Manohar, Hao Tang, Edmund C. Lalor, Nancy F. Chen, Paul Hager, Tyler Kekona, Rose Sloan, and Adrian K. C. Lee, "ASR for under-resourced languages from probabilistic transcription," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 25, no. 1, pp. 46–59, 2017.

[10] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014.

[11] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014.

[12] Daniel Povey, Mirko Hannemann, Gilles Boulianne, Lukás Burget, Arnab Ghoshal, Milos Janda, Martin Karafiát, Stefan Kombrink, Petr Motlícek, Yanmin Qian, Korbinian Riedhammer, Karel Veselý, and Ngoc Thang Vu, "Generating exact lattices in the WFST framework," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, 2012.

[13] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016.