Interspeech 2018
2-6 September 2018, Hyderabad

Lightly supervised vs. semi-supervised training of acoustic model on Luxembourgish for low-resource automatic speech recognition

Karel Veselý[1], *Carlos Segura*[2], *Igor Szöke*[1], *Jordi Luque*[2], *Jan "Honza" Černocký*[1]

[1]BUT, Speech@FIT group and IT4I Center of Excellence, Brno, Czech Republic
[2]Telefonica Research, Barcelona, Spain

iveselyk@fit.vutbr.cz, carlos.seguraperales@telefonica.com

Abstract

In this work, we focus on exploiting 'inexpensive' data in order to to improve the DNN acoustic model for ASR. We explore two strategies: The first one uses untranscribed data from the target domain. The second one is related to the proper selection of excerpts from imperfectly transcribed out-of-domain public data, as parliamentary speeches. We found out that both approaches lead to similar results, making them equally beneficial for practical use. The Luxembourgish ASR seed system had a 38.8% WER and it improved by roughly 4% absolute, leading to 34.6% for untranscribed and 34.9% for lightly-supervised data. Adding both databases simultaneously led to 34.4% WER, which is only a small improvement. As a secondary research topic, we experiment with semi-supervised state-level minimum Bayes risk (sMBR) training. Nonetheless, for sMBR we saw no improvement from adding the automatically transcribed target data, despite that similar techniques yield good results in the case of cross-entropy (CE) training.

Index Terms: Luxembourgish, call centers, speech recognition, low-resourced ASR, unsupervised training

1. Introduction

In the last years, low-resource Automatic Speech Recognition (ASR) has been a hot research topic in the speech processing community. It is interesting to realize that one of the low-resource languages can be found in the very center of western Europe. Luxembourgish, with its 390k native speakers, has been on UNESCO's list of endangered languages since 2010. Only few works from the ASR related literature are focused on this language from the West Germanic family. On the other hand, the very high GDP of Luxembourg makes it a potentially interesting market for the 'new technologies', such as ASR, which can be used in the customer care contact centers, and one of the H2020 BISON project[1] partners was based in Luxembourg. However, the lack of suitable training data was a limiting factor. Luxembourgish is mainly spoken in the Grand Duchy of Luxembourg, where it is the national language and one of three administrative languages, alongside with French and German. It was traditionally spoken just at home, which makes it difficult to have coherent grammar and spelling. It also has a substantial number of loan words from French and German. Luxembourgish has various distinct regional dialects beside the central Luxembourgish variety, which is seen as the emergent standard language [1].

The pioneering work on Luxembourgish [2] was exploring the possibility of porting the acoustic models trained on similar languages. In [3], a multilingual acoustic model (German + French + English) was used to bootstrap the unsupervised training of Luxembourgish ASR system with 1200 hours of Luxembourgish data (TV-broadcasts). In our work, we do not need the bootstrapping, as transcribed Luxembourgish data (17.7 hours) were collected as part of the BISON project. The data were collected in telephone contact centers of various services (cable TV, etc.).

Given that 17.7 hours is not much data compared to thousands of hours of data for well represented languages, we explore the ways how to further improve our acoustic models with some 'inexpensive' data without transcribing them manually. We compare the following two approaches:

a) Semi-supervised training (SST) with collection of untranscribed in-domain data from contact centers,
b) Training with free out-of-domain data that have imperfect transcripts, the data used for this purpose in our work are the recordings of plenary sessions in Luxembourgish Parliament.

The parliament data are freely available[2], while the untranscribed contact center data were collected by our project partner as anonymized data processed by feature extraction. Both sources of data are inexpensive and would be helpful for developing a commercial system.

In case of the *imperfect transcripts*, we can improve their accuracy by 'assisted decoding' with language model biased towards the imperfect transcripts [4]. Furthermore, it is possible to select the portions of data representing the 'islands of confidence' in which the reference matches the output of the biased decoding [5, 6]. In our work we formed the 'islands' by searching for the 'oracle' path through lattice, i.e. the path with minimum edit distance to the imperfect transcript. The 'islands' are the validated chunks of transcripts and we 're-segment' the training data to be formed from these 'islands'. The rest of speech is discarded.

With the *untranscribed data*, the situation is more difficult. The data can be used in the semi-supervised training [7, 3, 8, 9, 10, 11, 12, 13]. Here, we need to identify the most reliable parts of the automatically generated transcripts to be included into the training. This is typically done according to a lattice-based confidence [14, 7], or eventually according to 'agreement analysis' [15] if multiple ASR systems are available. In our work, we build on top of the word-selection recipe from [7], while we experiment with data selection by frame-masking (scaling gradients with per-frame 0/1 weights), or data re-segmentation (selecting the sub-segments with reliable transcripts). We also focus on sMBR training with untranscribed data, and replicate the experiment with two output layers from [16].

[1] http://bison-project.eu/

[2] http://www.chd.lu/wps/portal/public/Accueil/TravailALaChambre/Recherche/Videos

10.21437/Interspeech.2018-2361

2. Data

2.1. Audio and transcripts

The audio datasets we experiment with are:

CC data 17.7 hours of transcribed telephone data from the contact centers (target domain). We subdivided this set into development set (49 speakers, 2 hours) and training set (368 speakers, 15.7 hours). We tripled the duration of the training set by using speed perturbation with warping factors 0.9, 1.0 and 1.1.

CC-untran data 340 hours of untranscribed data, coming from the target domain of 'contact centers', obtained from our project partner as anonymized feature files.

Parl. data 358 hours of freely available parliament speech recordings, coming from channel different from target domain. The parliament data are not transcribed perfectly, we need to synchronize them with the audio and remove words that are likely to be wrong.

In the transcripts, we verbalize numbers, and we also identify acronyms to match them with their 'spelled' pronunciation in the lexicon.

2.2. Dictionary

The word-list of totally 110k word-items was assembled from sources:

- contact center training transcripts,
- lexicon from 'eSpeak'[3] speech synthesizer,
- parliament transcripts,
- wikipedia corpus (words present at least 4x were added),
- forvo words (on-line database of spoken words),

The dictionary was built according to the pronunciation rules described in the Omniglot[4] website. This led to better results than using the graphemic lexicon or using the lexicon from 'eSpeak' speech synthesizer.

We separate the preposition D' as a standalone token, and we also split the long word forms of numbers at multiples of ten as illustrated by: 'EENANZWANZEG -> EENAN ZWANZEG'

The final word-list was filtered according to the set of valid graphemes for Luxembourgish. This eliminates non-standard symbols (quotes, parenthesis, non-standard tokens containing numbers, etc.) remaining after text data filtering, which can be abundant particularly in the wikipedia corpus.

2.3. Language model

The language model is an interpolated trigram model, built from three source LMs, trained on following corpora:

- transcriptions from the contact center training set (164k words)
- parliament transcripts (3.25M words)
- Wikipedia (6.05M words)

The interpolation weights were tuned on the contact center development set. The actual combination weights were: 0.84 contact-center training set, 0.12 parliament transcripts, 0.04 wikipedia corpus. The final LM was pruned by SRI-LM with `-prune 5e-10`, the final LM size was: 400k 3grams, 1219k 2grams, 109k 1grams. We did not use LM rescoring of lattices.

3. ASR engine

The DNN-HMM system is built with 'nnet1'[5] recipe from Kaldi [17]. The models are trained on top of fMLLR features produced with GMM-HMM system. The fMLLR features are obtained by splicing +/- 4 frames of the 13-dimensional PLPs (includes C0) extended by 3 kaldi-pitch features [18], both feature types are mean normalized on per-speaker basis. The spliced features are projected to 40 dimensions with a global LDA+MLLT [19] linear transform and per-speaker fMLLR [20] linear transform.

The DNN has a feed-forward topology with 6 hidden layers of 2048 sigmoidal neurons. There is 440 dimensional input and $\approx$2500 or $\approx$7000 dimensional softmax output, depending on the amount of training data. The 40 dimensional fMLLR features are spliced by +/- 5 frames and globally re-normalized to have zero mean and unit variance. We used RBM pre-training [21] to initialize the 6 hidden layers. Then, the 'frame CE' training, was done with mini-batch SGD (Stochastic Gradient Descent), in which the learning rate is halved from an epoch with a small improvement of held-out loss till the convergence of the held-out loss. Finally, the network is re-trained by 6 or 2 epochs of sMBR training [22], depending on the amount of training data. In the experiments, we are using NN with single output layer, unless explicitly mentioned otherwise.

4. Experiments

Having a relatively low amount of the transcribed in-domain data (17.7 hours), the accuracy of the acoustic model can be improved by use of additional data. In our scenario we consider adding the 'CC-untran' data (untranscribed data from the target domain = contact centers), or the 'Parl' data (imperfectly transcribed parliament data from different domain). We train either with 'masking' (scaling gradients in NN training with 0/1 per-frame weights) [7], or with 're-segmentatation' (selecting sub-segments with reliable transcripts) [6]. We begin with constructing a seed system, which we use for decoding automatic transcripts and filtering the imperfect transcripts.

4.1. Seed system

The seed system was built with the 15.7 hour subset of CC data with manual transcripts, and evaluated on our development set from the same domain. The model was tuned to have 2500 tied-states. For verification of 'imperfect transcripts' by decoding with biased language model, we used GMM-HMM model. For generating the automatic transcripts, we used an DNN-HMM system with WER of 38.8, this seed system is referenced later in Table 1.

4.2. DNNs with 2500 tied-states

We initially experimented with the 2500 tied-states obtained from the seed system.

4.2.1. Adding parliament data

For the parliament data, we had the imperfect manual transcripts available. We sub-segmented the data by

[3] `http://espeak.sourceforge.net/`

[4] `https://www.omniglot.com/writing/luxembourgish.htm`

[5] `http://kaldi-asr.org/doc/dnn1.html`

finding the 'islands of confidence' with the Kaldi script `steps/cleanup/clean_and_segment_data.sh`. Internally, the lattices are generated with the GMM-HMM acoustic model. Then, an 'oracle path' is found in the lattice, i.e. a path with minimum edit distance to the transcripts. The 'islands' are defined as the chunks where 'oracle path' matches the transcripts. From the original 358 hours of speech, we obtained a cleaned dataset of 297 hours.

A similar method is described in [23], where Smith-Waterman alignment was used to match transcripts with 1-best output from the decoder, or [24] where the matching was done against a confusion network.

The cleaned data are then merged with the manually transcribed data, and new acoustic models are built (line 'CC + Parl,re-segment' with WER 35.6 in Table 1). In this case the per frame mask is not necessary, as the dataset is re-segmented.

4.2.2. Adding untranscribed data

With the untranscribed data, we perform semi-supervised training: We generate automatic transcripts and per-word confidences with the seed system. Next, we train with mixed manually or automatically transcribed data.

In our recent work [7], we found that word selection is beneficial, and we identified a heuristic for finding the optimal amount of accepted words: We select the words with higher MBR confidence (the statistics $\gamma(q, s)$ from the Minimum Bayes Risk decoding [25, section 7.1]). We set the proportion of added words the same as the word-accuracy of seed system on development set.

The word accuracy of our seed ASR system is 61.2%. Therefore, we selected 61% of the automatically transcribed words and mix them with the correctly transcribed data, which were previously augmented by speed perturbation.

This data selection, implemented by frame-masking, works well for the frame-cross entropy training of a DNN. However in [7], we did not use the automatic transcripts for the sMBR training. The sMBR is done only with the manually transcribed data. With this recipe, we obtained WER 36.0 in line 'CC + CC-untran,masking' in Table 1.

4.2.3. sMBR training with untranscribed data

We re-visited the sMBR training with frame masking, we tried to add the same 61% words into the sMBR training with frame-masking of NN gradients. We also applied the mask to the approximate accuracy calculation in the sMBR forward-backward algorithm, effectively excluding the frames with uncertain labels from the calculation of accuracy. However, after adding the masked untranscribed data, the results became worse than training only with the manually transcribed data: WER 36.6 instead of original 36.0 from Table 1. Therefore, either sMBR training is very sensitive to the quality of the automatic transcripts, or the frame-masking is not suitable for sMBR training. Recall that for CE training the frame-masking led to good results [7].

4.2.4. Summary

From the results in Table 1, we see that adding both the untranscribed data (CC-untran,masking) or the parliament data (Parl.,re-segment) separately leads to performance improvements. In the frame cross-entropy training (CE), the 'CC + CC-untran,masking' was better then 'CC + Parl,re-segment' (WER 37.3% vs. 38.5%). However, the ranking switched after the sMBR training: The cleaned parliament data were successfully used for the sMBR training (35.6% WER), while in case of 'CC + CC-untran,masking' the sMBR training was done with the manually transcribed subset 'CC only' leading to WER 36.0%. Recall that sMBR training with masked untranscribed data is not helpful.

Table 1: *DNN systems (nnet1) trained on various types of training data: 'CC' = contact center data, 'CC-untran,masking' = untranscribed contact center data for semi-supervised training with frame masking, 'Parl,re-segment' = re-segmented parliament data. Models with 2500 tied-states.*

Training data	% WER	
	CE	sMBR
DNNs with 2500 tied-states,		
CC only (seed system)	40.7	38.8
CC + Parl,re-segment	38.5	**35.6** (CC + Parl,re-seg.)
CC + CC-untran,masking	37.3	36.0 (CC only)

Table 2: *Semi-supervised training with re-segmented data, in which we tune the amount of added words. The DNN models have 7000 tied states.*

Added words,	% WER	
(CC-untran,re-segment)	CE	sMBR
50% (155h)	36.2	35.3
60% (187h)	36.1	34.9
70% (222h)	36.0	**34.7**
		34.7 (CC only)
80% (260h)	36.3	35.0
61% CC-untran,masking (203h)	36.0	**34.6** (CC only)
CC + Parl,re-segment	37.2	34.9
CC + Parl,re-seg + CC-untran,re-seg	**35.1**	**34.2**

4.3. DNNs with 7000 tied-states

With more training data available, we can build larger acoustic models. To train a DNN-HMM with 7000 tied-states, we first trained GMM-HMM model with the extended dataset (extended by automatically transcribed data without any filtering or cleaned parliament data). This provides the phone-tree and alignment for the DNN-HMM.

4.3.1. Adding parliament data

The parliament data were cleaned and used as described in 4.2.1. The only difference is that the training targets come from a larger GMM-HMM model trained on cleaned parliament database. With this setup, we obtained WER 34.9%, as noted in Table 2 as 'CC + Parl,re-segment'.

4.3.2. Adding untranscribed data

Inevitably, the automatically generated transcripts contain errors. In 4.2.2, we were discarding unlikely words with frame-masking. We have also seen that frame-masking does not work well for sMBR training. An alternative to frame-masking is to 're-segment' the data by selecting the chunks composed of words with higher confidence, which might possibly help in sMBR training. We used the same MBR confidence as in 4.2.2,

Table 3: *Introducing 2nd output layer for sMBR training with untranscribed data. Added 70% automatically transcribed words, the data were re-segmented [%WER].*

Baseline (1 output layer):	34.7
2 output layers:	
trained by transcribed	34.8
trained by untranscribed	34.8

while we tune the amount of added words in Table 2. The best was to add 70% of words, which leads to WER 34.7%.

As a sanity check, we performed sMBR training with manually transcribed data 'CC only', departing from the same 70% CE model. Surprisingly, in both cases we obtained the same WER 34.7. This indicates that adding automatically transcribed data into sMBR training was again not helpful.

As a contrastive experiment, we also trained a model with 'frame-masking' as we did in [7], denoted in Table 2 as '61% CC-untran,masking'. Here, the sMBR training is done with the manually transcribed subset 'CC only'. Surprisingly, this led to 0.1% better WER than we obtained with data re-segmentation. So, we cannot say that data 're-segmentation' is any better than 'frame-masking'. Both approaches lead to results which are on-par.

4.3.3. sMBR with training with 2 output layers

In some works on semi-supervised training, we see models with 2 output layers [12, 13], the 1st output for manually transcribed data, the 2nd output for automatically transcribed data.

Recently, in [16] the authors obtained the best results with NN training done by 'Naive CE + SHL ST' approach: a frame CE training of NN with one output layer followed by sequence training (ST) of NN with the 2 output layers. The training is done with mixed transcribed/untranscribed data, and there are shared hidden layers (SHL) for both types of data.

We replicated this recipe, however, in Table 3, we see that introducing the 2nd output layer for sMBR training did not bring better result than training with 1 output layer. Note that we use lower amount of data than in [16] and unlike [16], we re-segmented the untranscribed data according to word-confidences.

Based on the results in Table 3, we see that it is not crucial to add the 2nd output layer for the sMBR training with the mixed transcribed/untranscribed data, as there was no performance improvement.

4.3.4. Adding parliament and untranscribed data

In the last line of Table 2, we combined all the data: the manually transcribed contact center data (CC), the cleaned parliament data 'Parl,re-seg', and the untranscribed contact center data 'CC-untran,re-seg'. This nearly doubles the amount of training data to 523 hours, while the WER further drops to 34.2%, which is 0.4% better than the previous best result 34.6%.

4.3.5. Summary

With the 7000 tied-state models, it was 0.3% better to add the automatically transcribed data compared to adding parliament data (see Table 2). This is opposite to what we saw in Table 1, where parliament data were better. Thanks to the larger model with more tied-states, the WER dropped from 35.6% to 34.6%.

Despite the effort, we did not find a way to obtain benefit of adding automatic transcripts into sMBR training. However, further performance improvement is possible, if we merge the parliament and untranscribed data.

5. Conclusions

In this paper, we explore two scenarios of adding extra audio data into the training of acoustic model for a low-resource language: adding out-of-domain data with imperfect transcripts (parliament data), or using untranscribed data collected directly from the target domain (contact centers).

Our main observations are summarized as follows:

- Both types of added data improved the WER by ≈4% absolute, from 38.8% to 34.6% (untranscribed data) and 34.9% (parliament data).
- With 7000 states, the system with added untranscribed data was better by 0.3%; with 2500 states, the system with added parliament data was better by 0.4%. Hence, we cannot say which type of data is better to add. It is better to conclude that both types of data help similarly.
- With more training data, it is beneficial to increase the number of tied-states in acoustic model; by increasing the number of tied-states from 2500 to 7000 the best WER improved from 35.6% to 34.6%
- We have seen no difference between frame cross-entropy training with per-frame masks or with re-segmented data, which can be expected
- For sMBR training, the data re-segmentation prevented the degradation from adding the untranscribed data. However, the result was the same as if no untranscribed data were added. In other words, the semi-supervised sMBR training had no advantage over supervised sMBR training, which uses less data and has faster training time.
- We found it unnecessary to add the 2nd output layer for untranscribed data in semi-supervised sMBR training.

We believe that semi-supervised training should be most helpful for generative models like GMMs, where mislabeled data cause little harm. In frame cross-entropy discriminative training, the semi-supervised approach is still helpful, because the language model corrects some of the mistakes of the acoustic model. However, in the sMBR training with automatically decoded reference, we are fostering the reference-paths which are either wrong, or which were already decoded correctly. This might be a good explanation why sMBR training with automatically decoded transcripts is unlikely to further improve the results, which we have seen in our experiments.

6. Acknowledgements

This work was supported by Technology Agency of the Czech Republic project No. TJ01000208 "NOSICI" and by European Union's Horizon 2020 project No. 645523 BISON. This text reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

7. References

[1] P. Gilles and C. Moulin, "Luxembourgish," *Germanic standardizations: past to present*, vol. 18, pp. 303–329, 2003.

[2] M. Adda-Decker, L. Lamel, G. Adda, and T. Lavergne, “A first LVCSR system for luxembourgish, a low-resourced european language,” in *Human Language Technology Challenges for Computer Science and Linguistics - 5th Language and Technology Conference, LTC 2011, Poznań, Poland, November 25-27, 2011, Revised Selected Papers*, 2011, pp. 479–490.

[3] M. Adda-Decker, L. Lamel, and G. Adda, “Speech Alignment and Recognition Experiments for Luxembourgish,” in *4th International Workshop on Spoken Language Technologies for Underresourced Languages*, Saint-Petersbourg, Russia, 2014, pp. 53–60. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01134824

[4] B. Lecouteux, G. Linarès, P. Nocera, and J. Bonastre, “Imperfect transcript driven speech recognition,” in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[5] H. Liao, E. McDermott, and A. W. Senior, “Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription,” in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, 2013.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*.

[7] K. Veselý, L. Burget, and J. Cernocký, “Semi-supervised DNN training with word selection for ASR,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*.

[8] K. Veselý, M. Hannemann, and L. Burget, “Semi-supervised training of deep neural networks,” in *Proceedings of ASRU*, 2013, pp. 267–272.

[9] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, “Deep neural network features and semi-supervised training for low resource speech recognition,” in *Proceedings of ICASSP*, 2013, pp. 6704–6708.

[10] F. Grezl and M. Karafiat, “Semi-supervised bootstrapping approach for neural network feature extractor training,” in *Proc. of ASRU*, 2013.

[11] P. Zhang, Y. Liu, and T. Hain, “Semi-supervised DNN training in meeting recognition,” in *2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014*, 2014, pp. 141–146.

[12] H. Su and H. Xu, “Multi-softmax deep neural network for semi-supervised training,” in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 3239–3243.

[13] V. Manohar, D. Povey, and S. Khudanpur, “Semi-supervised maximum mutual information training of deep neural network acoustic models,” in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2630–2634.

[14] F. Wessel and H. Ney, “Unsupervised training of acoustic models for large vocabulary continuous speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[15] F. de Chaumont Quitry, A. Oines, P. J. Moreno, and E. Weinstein, “High quality agreement-based semi-supervised training data for acoustic modeling,” in *2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, December 13-16, 2016*, 2016.

[16] M. Gibson, G. Cook, and P. Zhan, “Semi-supervised training strategies for deep neural networks,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, 2017, pp. 77–83.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, “The Kaldi speech recognition toolkit,” in *Proc. of IEEE ASRU*, 2011.

[18] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” in *Proceedings of ICASSP*, 2014.

[19] M. J. F. Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, May 1999.

[20] D. Povey and G. Saon, “Feature and model space speaker adaptation with full covariance Gaussians,” in *Proc. of INTERSPEECH*, 2006.

[21] G. E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[22] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” in *Proc. of INTERSPEECH'13*, 2013.

[23] V. Manohar, D. Povey, and S. Khudanpur, “JHU kaldi system for arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, 2017.

[24] L. Chen, L. Lamel, and J. Gauvain, “Lightly supervised acoustic model training using consensus networks,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*, 2004.

[25] H. Xu, D. Povey, L. Mangu, and J. Zhu, “Minimum bayes risk decoding and system combination based on a recursion for edit distance,” *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.