SELECTION AND COMBINATION OF HYPOTHESES FOR DIALECTAL SPEECH RECOGNITION

Victor Soto[⋆] *Olivier Siohan*[†] *Mohamed Elfeky*[†] *Pedro Moreno*[†]

[⋆] Columbia University, Computer Science Department, New York
[†] Google Inc., Languages Modeling Group, New York

ABSTRACT

While research has often shown that building dialect-specific Automatic Speech Recognizers is the optimal approach to dealing with dialectal variations of the same language, we have observed that dialect-specific recognizers do not always output the best recognitions. Often enough, another dialectal recognizer outputs a better recognition than the dialect-specific one. In this paper, we present two methods to select and combine the best decoded hypothesis from a pool of dialectal recognizers. We follow a Machine Learning approach and extract features from the Speech Recognition output along with Word Embeddings and use Shallow Neural Networks for classification. Our experiments using Dictation and Voice Search data from the main four Arabic dialects show good WER improvements for the hypothesis selection scheme, reducing the WER by 2.1 to 12.1% depending on the test set, and promising results for the hypotheses combination scheme.

***Index Terms*—** speech recognition, dialects, system combination, system selection

1. INTRODUCTION

Dialects are defined as variations of the same language, specific to geographical regions or social groups. Dialects of the same language are differentiated at various linguistic levels. For example, at the prosodic level, Arabic dialects differ in intonational and rhythm cues [1]. At the orthographical level, the same word can have different spellings like *color* vs. *colour* and *center* vs. *centre* in Standard American English and British English respectively. Vocabularies can evolve quite differently between dialects too, depending among other factors on the interactions with other languages, eg. in Castilian Spanish the word for cell phone is *móvil*, whereas Latin American speakers will use *celular*. In some cases, these linguistic differences do not impact speech intelligibility among speakers of different dialects, as is the case for English or Spanish speakers, whereas for other languages like Arabic, dialect speakers will understand each other with much difficulty [2].

Inevitably, these variations impact the development of the acoustic model, language model, pronunciation model and lexicon of an Automatic Speech Recognizer (ASR). Therefore, a decision must be made whether to develop an ensemble of dialect-specific recognizers or a single unified recognizer. So far, the strategy at Google has been to build dialect-specific recognizers. This decision was based on linguistic facts as well as rigorous cross-dialect experimental analysis (e.g., [3]). In an application like VoiceSearch, the issue becomes how to choose a recognizer to decode an arbitrary spoken query. In the past, queries have been directed to a dialectal recognizer based on the user's language / country preferences, whereas presently the recognizer is selected based on location information extracted from the query's IP address [4]. For example, voice queries that originate from Egypt are directed to the Egyptian Arabic recognizer. Table 1 shows the Word Error Rate (WER) performance of Google's production ASR for the main four Arabic Dialects (Egyptian, Gulf, Levantine and Maghrebi) on Dictation (IME) and VoiceSearch (VS) data. These systems, despite sharing similar architectures and algorithms, show significantly different WER behaviors that can be attributed to the data-dependent nature of these algorithms. Unified Arabic ASRs in [3] were shown to underperform compared to its dialect-specific counterparts. Since based on these experiments using dialect-specific is the optimal choice, but their performance show such variance between them, the question then becomes whether we can use better dialectal speech recognizers to leverage dialectal systems of the same language.

ASR (Dataset)	# Utts	WER
Egyptian (IME)	8439	37.4
Egyptian (VS)	8633	34.7
Gulf (IME)	3536	29.4
Gulf (VS)	8739	21.5
Levantine (IME)	10138	33.7
Levantine (VS)	9677	28.4
Maghrebi (IME)	7829	38.4
Maghrebi (VS)	8090	34.7

Table 1. *Number of utterances (column 2) and WER for each dialect-specific ASR on its own test sets (column 1). These are used as baseline results throughout the rest of the paper.*

We run a series of experiments to explore the potential of such idea. Table 2 contains two subtables. The first subtable contains the cross-dialectal performance of the four dialectal Arabic ASRs (columns two to five) evaluated on the eight test sets. The WER numbers clearly show that on average the best performing system for each test set is its own matched ASR. The second subtable shows the performance of two oracle system experiments. The first oracle (column six) performs what we refer to as *hypothesis selection*. It consists of decoding each voice query from each test set using the four dialect-specific ASR and then hand-picking the one with the lowest WER. The second oracle (column seven) performs *hypotheses combination*. It combines the four decoded hypothesis into a word alignment similar to the procedure used in ROVER and then selects the correct word from each bin. In both cases, the WER improvements are very large, reaching between 23.7% and 59.1% relative improvements over the dialect-specific ASRs. Unsurprisingly, the word-level combination oracle outperforms the utterance-level selection in every test set. From these experiments we can conclude that, even though on average dialect-specific systems are the best option for dialectal speech recognition, there is potential for vast improvement

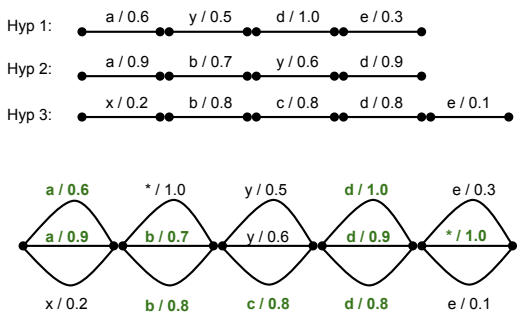

Fig. 1. *Hypotheses Word Alignment. Each arc contains a pair of word token and confidence score. The epsilon token is represented as an asterisk. The majority vote function would return the hypothesis "a b y d e", whereas the Maximum Sum Confidence hypothesis would be "a b y d". iROVER can potentially return the true reference transcript "a b c d".*

if we employ selection and combination techniques across dialectal hypothesis.

In this paper, we propose selection and combination schemes that use Machine Learning (ML) classification to improve speech recognition performance across dialects of the same language. The paper is organized as follows. In section 2, we give an overview of work on dialectal speech and hypothesis combination in the speech recognition field. In section 3, we give an overview of the ASRs and corpora used in our experiments. Sections 4 and 5 explain the feature extraction pipeline, the experiments conducted and results for the hypothesis selection and combination schemes respectively. Section 6 concludes the paper.

2. PREVIOUS WORK

Similar to the topic of *hypothesis selection*, there has been a considerable effort from the speech recognition community towards reranking N-best lists of speech recognition outputs. In [5], the authors proposed a scheme that consisted of producing an N-best list, performing a rescoring pass on each hypothesis, and then combining both sets of scores. Later on, [6] used the weighted sum of learned discriminative scores as the decision function to choose from the N-best list; [7] used syntactic, semantic and acoustic features and trained a Linear Regression classifier to rerank N-best hypotheses; [8] used Comprehensive Information Theory to calculate the amount of comprehensive information in each utterance and reorder N-best lists according to it; and [9] used Maximum Entropy classifiers with speech recognition, semantic and search results features and selected the utterance with highest prediction confidence as correct. Finally, in the same spirit of our hypothesis selection task presented here, the authors in [10] propose to join several N-best lists of Machine Translation output and use sentence-level features to select on hypothesis.

On the topic of *hypothesis combination* for speech recognition output, ROVER [11] is a widely used algorithm for system combination. It works in two basic steps. First, it finds a linear alignment between the words of each sentence using the Levenshtein distance algorithm, and then proceeds to choose a word from each word alignment following an aggregation function. Typically, two aggregation functions are used in the literature: Majority voting, which chooses the token that appears most often in the word alignment; and Maximum Weighted Voting, which sums up the confidence of every arc with the same token in the word alignment and picks the one with the highest confidence. This process is illustrated in Figure 1.

Several modifications have been proposed to improve ROVER. In [12], the authors proposed N-best ROVER, an extension that merges N-best hypothesis of several systems, and reported relative gains of 0.8% WER points with respect to ROVER. Following a ML approach, iROVER [13] proposes to first compute the ROVER alignment of several 1-best hypotheses and then use a ML classifier to choose the token from each alignment slot. For their experiments, they used Adaboost ensembles of decision tree stumps and feature vectors containing confidence, durational, character-distance and top error features. They reported between 2.8 and 13.4% relative WER points with respect to ROVER, depending on the number of systems being combined.

Another approach to system combination different from combining 1-best and N-best hypotheses has been combining Confusion Networks. A Confusion Network (CN) [14] is a compact linear representation of a speech recognition lattice, designed to optimize word error rate. CNs are created by clustering lattice edges into an ordered series of slots based on time similarity. Confusion Network Combination [15] finds alignments between the word slots of several CNs and unifies them into a single graph by substituting the string match function used during the alignment process with a function that computes the probability of word match given two word slots. Some studies have also used ML to select [16] or rerank [17] the CN word arcs. An alternative approach to CNs for minimizing WER decoding on word graphs is Time Frame Error decoding [18], in [19] an algorithm is presented to combine multiple systems producing word graphs of different structure and with different segmentations.

In this paper, we build on previous hypothesis selection and combination algorithms and focus on its application and impact on dialectal speech recognition. We also introduce the use of Neural Networks and Word Embeddings for this task.

3. ASR AND CORPORA

We have four dialect-specific ASRs for Egyptian, Gulf, Levantine and Maghrebi Arabic. Each ASR was trained on corpora of its own dialect comprising around 3M anonymized user utterances. The acoustic model of our speech recognizers is a very similar architecture as the one described in [20]. It is a fully-connected feed-forward bottleneck network trained on minibatch stochastic gradient descent. Our network is composed of an input layer, eight hidden layers, a bottleneck layer and a softmax layer. The input layer to the network is a 1040 dimensional vector composed of 26 frames of 40 log filterbanks each. The 26 frames contain 20 past frames, the current frame and 5 following frames. The eight hidden layers have 2560 ReLU units each [21]. The bottleneck layer [22] has 256 linear activations and a soft-max layer holds 14336 units, one for each context-dependent state in our inventory. In total, the DNN holds around 53 million parameters.

The ASR test sets consist of 25K of anonymized and manually-transcribed utterances each. Manual transcribers were asked to follow specific guidelines that take into account the spoken form of the Arabic dialects, and the plethora of colloquial words in them. The test sets are 50% balanced between VoiceSearch and dictation logs. Hence, there are eight test sets: two test sets per dialect of dictation and VoiceSearch. For our hypothesis selection and combination experiments we used the decoded output of those ASR test sets, and

Dataset	Production ASRs				Oracles	
	Egyptian	Gulf	Levantine	Maghrebi	Selection	ROVER
Egyptian (IME)	**37.4**	43.5	44.3	53.1	26.9 (+28.1%)	23.1 (+38.2%)
Egyptian (VS)	**34.7**	38.2	42.2	48.2	23.6 (+47.0%)	19.4 (+44.1%)
Gulf (IME)	36.2	**29.4**	34	47.4	20.8 (+29.3%)	18.7 (+36.4%)
Gulf (VS)	27.6	**21.5**	26.3	37.3	14.3 (+33.5%)	12.7 (+59.1%)
Levantine (IME)	41.2	38	**33.7**	48.9	25.7 (+23.7%)	23.1 (+31.5%)
Levantine (VS)	34.7	29.9	**28.4**	41	19.9 (+29.9%)	17.7 (+37.7%)
Maghrebi (IME)	44.2	41.5	41.6	**38.4**	24.6 (+35.9%)	21.1 (+45.1%)
Maghrebi (VS)	42.6	38.2	41.5	**34.7**	21.9 (+36.9%)	18.6 (+46.4%)

Table 2. *Left subtable shows the cross-dialectal performance of each ASR in two test sets. Right subtable contains the performance of the hypothesis selection and hypotheses combination oracles. Relative improvements (%) with respect to the matched systems are between parentheses.*

Dataset	Best Hyp Selection	Rel. Imp	+ BWE	Rel. Imp.
Egyptian (IME)	36.1	+3.4	35.4	+5.3
Egyptian (VS)	31.8	+8.4	31.7	+8.6
Gulf (IME)	28.6	+2.7	28.3	+3.7
Gulf (VS)	20.7	+3.7	20.4	+5.1
Levantine (IME)	33.3	+1.2	33	+2.1
Levantine (VS)	26.4	+7.0	26.3	+7.4
Maghrebi (IME)	34	+11.5	33.7	+12.2
Maghrebi (VS)	30.7	+11.5	30.5	+12.1

Table 3. *Left subtable shows the WER performance of the best hypothesis selection systems trained on the baseline feature set and its Relative Improvements (%). Right subtable contains WER results and Relative improvements after adding the Bag-of-Words embedding (BWE) layer. Relative improvements are calculated with respect the matched-dialect systems WER.*

ran 5-fold cross-validation to train and test on it.

4. HYPOTHESIS SELECTION

Hypothesis selection consists of predicting the hypothesis to pick from a set of recognition hypotheses (from multiple systems) that will have the lowest WER. In this setting each voice query is decoded by the dialect-specific ASRs, a feature vector is created that describes every hypothesis at the utterance level, and a classifier predicts the best one.

We pose the classification problem as a multi-label learning task. Since more than one hypothesis can minimize the WER, each feature vector can have more than one label. The utterance-level features are: frame-averaged acoustic model cost, frame-averaged language model cost, minimum, maximum and average word confidence and word posterior, number of words, and the lattice density defined as the number of arcs in the decoding lattice divided by the duration of the utterance in seconds. Each feature has as many instances as the number of dialectal hypotheses (four for Arabic). Finally, we add the Levenshtein distance between each pair of hypotheses, making a total of forty-two features. All features are normalized to a mean of zero and a standard deviation of one. We train fully-connected feed-forward neural networks on this feature dataset. The architecture of the networks is the following: an input layer of forty-two dimensions, a single hidden layer of 512 ReLU units, and an output layer of four Logistic Regression units (one per hypothesis). The network is trained using minibatch stochastic gradient descent and a learning parameter with exponential decay.

The results of this first batch of experiments, show significant improvements, as can be seen in columns 2 and 3 of Table 3. The WER is reduced for every test set with respect to the matched system baseline, and we obtain relative improvements ranging between 1.2 and 11.5%. The largest improvements are obtained for Maghrebi.

We ran another batch of experiments using the baseline feature set and adding bag-of-word embeddings to our neural network input layer. Our word embedding implementation works as follows: each word in the corpus lexicon that appears more than five times is assigned an ID and the rest of the words are hashed into shared IDs. For a bag-of-words a vector is created with ones on the IDs of the words in the utterance and zeros in the rest of them. This vector is then fully connected to a hidden layer of 64 dimensions which is learnt during training. The embedded layer is then fully connected to the main hidden-layer. Running experiments with this setup and a hidden layer with 2048 ReLU units we obtain additional improvements with respect to the previous results and raise the relative improvement to 2.1 to 12.1% with respect to the matched system baseline, as shown in columns 4 and 5 of Table 3. Experiments using extra hidden layers on the neural networks were conducted and showed performance reductions.

5. HYPOTHESES COMBINATION

The idea behind hypotheses combination is to fuse all hypotheses into one that potentially has lower WER. We follow the approach taken by iROVER and create word alignments of the four dialect hypotheses. We do so by iteratively aligning hypothesis into a word finite-state transducer (FST) [23]. The alignment is done using the tokens in the hypothesis and ignoring any time information.

After the word alignment is created, the task becomes selecting the arcs from each word bin, which when combined together, create the best possible hypothesis. In order to do so, we create a feature vector per word bin and label it with the indices of each correct word arc. Finding the correct word arcs is done by aligning the FST and the reference transcript. Once again the task is a multi-label classification problem since more than one arc can be correct. The features extracted for each word bin are: acoustic model cost and language model cost of the FST arc and its frame-averaged values; weighted value of language model and acoustic model cost; word confidence and lattice posterior; number of phones in the token from the pronunciation lexicon; mean, standard deviation, best, worst and difference between mean and best acoustic model scores at the frame level; and epsilon arc flag. Notice that as in the hypothesis selection case, each feature has as many instances as the number of dialectal hypothe-

	ROVER			i-ROVER			
Dataset	Maj.Vote	Max.SumConf.	Rel. Imp	Best iROVER	Rel Imp.	+Context	Rel. Imp.
Egyptian (IME)	39.9	38.4	-2.7	37.6	-0.5	37.6	0.0
Egyptian (VS)	35.7	34.5	+0.6	32.7	+5.8	32.9	-0.6
Gulf (IME)	31.6	30.7	-4.4	29.8	-1.3	29.4	+1.3
Gulf (VS)	23.3	22.5	-4.7	21.2	+1.4	21	+0.9
Levantine (IME)	35.8	34.6	-2.7	35.2	-4.5	35.2	0.0
Levantine (VS)	28.5	27.9	+1.8	27.6	+2.8	27.6	0.0
Maghrebi (IME)	37	34.4	+10.4	35.3	+8.1	35.2	+0.3
Maghrebi (VS)	34.4	32.6	+6.1	31.2	+10.1	31.4	-0.6

Table 4. *ROVER (left subtable) and iROVER (right subtable) WER performance. For ROVER we try the Majority Voting and Maximum Confidence Sum aggregation functions and the relative improvement of the last one with respect the baseline matched systems. For iROVER we report the best trained systems on the baseline features along with its relative improvements and the WER of the contextual systems and its relative improvement with respect the baseline iROVER systems.*

ses. However, the features here are extracted at the word level rather than at the utterance level. We also add boolean features that indicate whether these features are ranked first in the word bin, and the lattice density defined as the time overlap of all FST arcs with the reference segment divided by the duration of that arc. Finally we add four layers of word embeddings, one per dialectal token.

We first test the performance of ROVER on our dataset (Table 4, columns 2-4) and find the following. Using the majority voting function, ROVER only has WER improvements on the Maghrebi test sets. For maximum sum confidence, it achieves additional WER points on Egyptian voice search and Levantine voice search datasets. In our experiments ROVER's performance is impacted severely by the number of systems being combined. Specifically, the Maghrebi hypotheses are causing performance drops and even though experiments conducted using ROVER with only three systems show small improvements, the Oracle experiments also show that the four system combination has the potential for great improvements too. Therefore, all iROVER experiments will be conducted using the four-system alignments.

For the combination experiments we use the same type of feed-forward neural networks and learning setup with two exceptions: the single hidden layer has 2048 ReLU units and the output layer consists of five logistic regression units, where four of the units predict one of the word tokens as correct and the other unit predicts all of them as incorrect and outputs an epsilon token. Columns 5-7 in Table 4 contain the WER, relative improvements by the best hypothesis combination systems. Comparing these results to Table 2, we observe that: a) iROVER not always improves the WER over the baseline ROVER, e.g., for Levantine IME and Maghrebi IME, iROVER actually yields slightly worse results; b) with respect to the matched systems, we obtain relative improvements of between 1.4 and 10.1% for five out of eight test sets and relative declines of 0.5 to 4.5%; and c) these relative improvements are smaller than the ones we obtained for hypothesis selection. As with hypothesis selection, adding extra hidden layers did not improve WER. To gain a deeper understanding of why our implementation of iROVER underperforms compared to the baseline and hypothesis selection we looked at the prediction performance class by class and found out that in all our experiments the F-score of the all-arcs-incorrect class had very low F-score (between 0.29 and 0.47, depending on the test set), which adds extra deletions to the WER measure.

In an effort to improve our hypothesis combination systems, we added context features, i.e., added the feature vectors of the two previous word bins. This gave us an extra 0.3%, 0.9% and 1.3% relative improvements on three datasets, no improvements in other three datasets and caused performance drops in two test sets. Therefore, it cannot be concluded that context helps iROVER performance.

6. CONCLUSIONS AND FUTURE WORK

We have presented a hypothesis selection and combination schemes and run experiments with them on the context of dialectal speech recognition. We found that using a limited amount of features from the speech recognition pipeline, the selection scheme achieved between 1.2 and 12.1% relative WER improvements (depending mainly on the performance of the dialect-specific ASR). Adding a word-of-bags embedding layer to the Neural Network has further improved WER by 2.1 to 12.2% relatively. For the combination experiments, we tried an implementation of iROVER with our own set of features and word embeddings. The combination system, despite having greater potential for improvements, underperformed in every test set when compared to the selection systems. We tried adding contextual features to the classification task but these did not seem to help the classification performance.

For future work we plan on exploring ways of improving the hypotheses combination systems. We also plan on using the hypotheses combination scheme to deal with utterances with instances of code-switching by combining the output of ASRs trained on different languages and mixed-language into time-aligned alignments.

7. REFERENCES

[1] Fadi Biadsy and Julia Hirschberg, "Using prosody and phonotactics in arabic dialect identification," in *Interspeech*, Brighton, UK, 2009.

[2] Margo E. Wilson, "Arabisc speakers: Language and culture, here and abroad," *Topics in Language Disorders*, vol. 16, no. 4, 1996.

[3] Fadi Biadsy, Pedro Moreno, and Martin Jansche, "Google's cross-dialect arabic voice search," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012, pp. 4441–4444.

[4] Mohamed Elfeky, Pedro Moreno, and Victor Soto, "Multi-dialectical languages effect on speech recognition: Too much choice can hurt," in *International Conference on Natural Language and Speech Processing (ICNLSP)*, 2015.

[5] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek, "Integration of diverse recognition methodologies through reevaluation of n-best sentence hypotheses," in *Proceedings of the Workshop on Speech and Natural Language*. 1991, HLT '91, pp. 83–87, Association for Computational Linguistics.

[6] Manny Rayner, David Carter, Vassilios Digalakis, and Patti Price, "Combining knowledge sources to reorder n-best speech hypothesis lists," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 217–221.

[7] Ananlada Chotimongkol and Alexander I Rudnicky, "N-best speech hypotheses reordering using linear regression," in *Proceedings of EuroSpeech*, 2001, pp. 1829–1832.

[8] Jianyi Liu and Yixin Zhong, "N-best speech hypothesis reordering based on comprehensive information theory," in *Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on*. IEEE, 2003, pp. 29–32.

[9] Fuchun Peng, Scott Roy, Ben Shahshahani, and Françoise Beaufays, "Search results based n-best hypothesis rescoring with maximum entropy classification.," in *ASRU*, 2013, pp. 422–427.

[10] Almut Silja Hildebrand and Stephan Vogel, "Combination of machine translation systems via hypothesis selection from combined N-best lists," in *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, 2008, pp. 254–261.

[11] Jonathan G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 347–352.

[12] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *In Proceedings of the NIST Speech Transcription Workshop*, 2000.

[13] Dustin Hillard, Björn Hoffmeister, Mari Ostendorf, Ralf Schlüter, and Hermann Ney, "iROVER: Improving system combination with classification," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, Rochester, NY, USA, Apr. 2007, pp. 65–68.

[14] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[15] Gunnar Evermann and PC Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*. Baltimore, 2000, vol. 27.

[16] Björn Hoffmeister, Ralf Schlüter, and Hermann Ney, "iCNC and iROVER: The limits of improving system combination with classification?," in *Interspeech*, Brisbane, Australia, Sept. 2008, pp. 232–235.

[17] Victor Soto, Erica Cooper, Lidia Mangu, Andrew Rosenberg, and Julia Hirschberg, "Rescoring confusion networks for keyword search," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7088–7092.

[18] Frank Wessel, Ralf Schlter, and Hermann Ney, "Explicit word error minimization using word hypothesis posterior probabilities.," in *ICASSP*. 2001, pp. 33–36, IEEE.

[19] Björn Hoffmeister, Tobias Klein, Ralf Schlüter, and Hermann Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *Interspeech*, 2006, pp. 537–540.

[20] Hank Liao, Erik McDermott, and Andrew Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 368–373.

[21] George E Dahl, Tara N Sainath, and Geoffrey E Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8609–8613.

[22] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6655–6659.

[23] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.