# Leveraging Linguistic Knowledge for Accent Robustness of End-to-End Models

# LEVERAGING LINGUISTIC KNOWLEDGE FOR ACCENT ROBUSTNESS OF END-TO-END MODELS

*Andrea Carmantini, Steve Renals, Peter Bell*

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

## ABSTRACT

Acoustic models are susceptible to the difference in acoustic characteristics between the training distribution and test distributions. Accent variability is a challenging source of variability, and the variations within one accent often do not generalize to others. Consequently, end-to-end models that have only transcriptions as linguistic information need high amounts of data to learn how different accents realize their sounds.

To aid with recognition of accented speech, we make use of an accent independent abstraction of phonemes, often called metaphonemes. We force our models to learn hidden representations that are correlated to metaphonemes using multi-task training. Our aim is to obtain a model that is more robust to accented speech and, can, at the same time, adapt faster to different accents through the learned structure.

Our experiments on the Common Voice corpus show better generalization when making use of this additional linguistic information, with a word error rate reduction of up to 12.6% when compared to the baseline. Furthermore, the relative improvement when adapting an existing model by making use of the metaphonemes is higher than using Byte Pair Encodings alone.

**Index Terms**: speech recognition, acoustic model adaptation, accent adaptation, end-to-end models

## 1. INTRODUCTION

Speech recognition models are highly susceptible to mismatch in the acoustic and language domains between the training and the evaluation data. Dialectal and accented speech is challenging for speech recognition models, as it can present variability in pronunciation, vocabulary and grammar.

Although there is significant literature on automatic dialect identification from speech (e.g. [1, 2, 3]), there has been less work on accent and dialect adaptive speech recognition systems, as the task is often assimilated to acoustic adaptation. The MGB–3 [4] and MGB–5 [5] challenges focused on the identification and transcriptons of dialectal Arabic test sets, with a modern standard Arabic (MSA) training set, using broadcast and internet video data. The best reported results reported on these challenges have used a straightforward model-based transfer learning approach in a Lattice-Free Maximum Mutual Information (LF-MMI) [6] framework, where a baseline model trained on MSA was finetuned with supervised data from specific Arabic dialects [7, 8].

In the context of hybrid systems, such choice seems obvious because the language model and lexicon are separate from the acoustic model, thus the acoustic model only needs to adapt to the new distribution of sounds in the dialect, while the grammar and vocabulary differences are provided to the system separately. In end-to-end models the system is monolithic, thus adaptation techniques aimed specifically at dialect and accent variability are more relevant.

A simple solution to dialectal variability is to create an unified robust model by pooling training data from multiple dialects [9]. In Elfeky et al. [10], this method is expanded by having a secondary output of the network learn to classify the input dialect, making the network explicitly aware of the dialectal variation. In the same vein, Yang et al. [11] train a neural dialect classifier sharing the weights with the lower layers of the ASR model, then uses the classifier's decision to select the dialect-specific output layers to use. Related research combined the multi-task approach with the use of dialectal information as input, either as a one-hot vector or as an embedding generated from a separate model [12, 13, 14].

Grace et al. [15] explored a family of cluster adaptive training and hidden layer factorization approaches and compared them to one-hot auxiliary inputs. They show that using one-hot dialect codes as an input augmentation (corresponding to bias adaptation) proved to be the best approach, and cluster-adaptive approaches did not result in a consistent gain. In related research, Jain et al. [16] explored a mixture of experts (MoE) approach, where separate input subnetworks learn transformations for the input features based on their accent; at test time, an external classifier chooses the right mixture given the input acoustics.

Yoo et al. [17] extended these approaches by applying a method of feature-wise affine transformations on the hidden layers (FiLM), dependent both on the network's internal state and the dialect/accent code. This approach, which can be viewed as a conditioned normalization, differs from the previous use of one-hot dialect codes and multi-task learning in that it has the goal of learning a single normalized model rather than an implicit combination of specialist models. A

related approach is gated accent adaptation [18], with the focus being on using a single transformation inserted in the intermediate layers of the network and conditioned on dialect label.

More recently, Winata et al. [19] experimented with a meta-learning approach for few-shot adaptation to accented speech, where the meta-learning algorithm learns a good initialization and hyperparameters for the adaptation. Techniques for continual learning, aimed at fine-tuning a model while avoiding catastrophic forgetting of what was learned previously, have also been applied to dialect adaptation [20, 21]

While end-to-end models try to make away with external linguistic information, previous research has shown that making use of phonetic transcriptions can help with the accuracy of these models, both when used as the primary source of information or a secondary task. [22, 23]. Furthermore, multitask configurations making use of different representation levels resulted in slight improvements for end-to-end models when using words as the primary task and characters as a secondary task [24, 25].

In our paper, we will discuss a method making use of external linguistic information to help the model learn better representations of the acoustics. Furthermore, we analyze an higher abstraction of phonemes, referred to as metaphonemes, developed to have an accent invariant representation of word pronunciations. We show that this expert knowledge can help end-to-end models learn general patterns describing accent variability, improving generalization and adaptation power.

## 2. METAPHONEMES

For our experiments, we make use of the Unisyn lexicon [26], a pronunciation resource for the English language built with accent invariance in mind. Unisyn maps words of the English language to accent agnostic transcriptions by using metaphonemes. Metaphonemes are an higher level abstraction of phonemes. A metaphoneme transcription is accent independent: for each metaphoneme and accent, there is a rule that describes how the metaphoneme is realized in the specific accent.

In Unisyn, all words in the English language are classified under certain "keywords" that exemplify the set of rules a specific metaphoneme follows to be mapped to its accent-specific phoneme. Examples of keywords, their realization in different accents and their Unisyn symbol for the open vowel are in table 1.

The use-case for which Unisyn was built is different than our aims. The aim of Unisyn is to use mapping rules written by linguists on the metaphonemic transcription to generate accent-specific lexicons. In our case, we want our models to learn the rules governing the shifts in accent pronunciation internally during training. Assuming the model can learn some of the patterns that govern accent variability, this should also help with faster adaptation of the models to new accents.

| Keyword | RP | American | Australian | Unisyn symbol |
|---|---|---|---|---|
| Tr**a**p | æ | æ | æ | a |
| B**a**th | $\alpha$: | æ | a: | ah |
| P**a**lm | $\alpha$: | $\alpha$* | a: | aa |

**Table 1**. Example of Unisyn metaphonemes for three keywords used to classify open vowels.

## 3. METHOD

To determine whether bringing separate linguistic information related to pronunciation can help end-to-end models learn better representations, we experimented with a multitask training configuration where both the graphemes and sounds of a word are used as targets.

The objective of the two tasks calculated from the attentional decoder is to minimize the cross-entropy of the ground truth given the acoustic observations:

$$\mathcal{L} = -\log P(Y|X) = -\sum_{t} \log P(y_t|X, Y_{0:t-1}) \qquad (1)$$

where $X$ is a vector of acoustic observations, $Y$ is the sequence of ground truth labels and $t$ is the step in the sequence. The two tasks will use different representations of the ground truth, where one of the representations is based on expert knowledge in the form of a lexicon while the other is based on the textual transcriptions.

We also use a secondary decoder trained using the Connectionist Temporal Classification objective function:

$$\mathcal{L}_{CTC} = -\log P(Y|X) = \sum_{\pi \in \mathcal{B}^{-1}(Y)} P(\pi|X) \qquad (2)$$

where $\mathcal{B}^{-1}(Y)$ is a set containing all possible alignments $\pi$ of the sequence $Y$ when including the blank label and allowing label repetitions. As shown in literature, this helps the model learn a monotonic alignment between the acoustics and the labels, resulting in faster convergence [27].

The multitask loss of our model is a weighted combination of the losses from the primary and secondary task and the CTC decoder:

$$\mathcal{L}_{mt} = (1 - \alpha - \beta)\mathcal{L}_1 + \alpha\mathcal{L}_2 + \beta\mathcal{L}_{CTC} \qquad (3)$$

For our secondary task, we use labels with a higher degree of granularity representing the sounds composing the utterance. To help the model with learning this representation as a lower lever abstraction of the main task's representation, we compute the output of the secondary task using the hidden

representation of a lower layer in the decoder instead of the last one. Fig.1 gives a visual representation of our architecture.

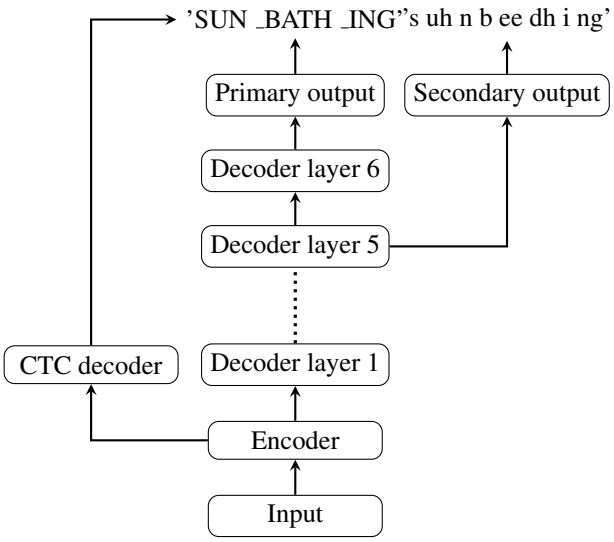


**Fig. 1**. Architecture of the multitask transformer model.

## 4. EXPERIMENTAL SETUP

Our experiments were carried out using a transformer model [28]. All the feed-forward layers in the transformer have a dimension of 2048. We use 8-headed attention layers with a dimension of 512. The encoder uses 12 transformer layers, while the decoder has 6. The primary target labels are byte pair encodings (BPE), with a vocabulary size of 5000 BPEs. A CTC decoder is trained jointly with the transformer decoder on the BPE labels. Our input features are 83 dimensional Filterbanks features with pitch. Our architecture follows the one in Karita et al. [29].

To include the information from the lexicons, we trained models in a multitask configuration. The secondary task in this configuration is recognizing the metaphonemic transcription of the audio. The output layer for the secondary task is connected to the penultimate layer of the decoder. This is to force the model to learn the metaphoneme transcription of the acoustics as a lower level abstraction of the BPEs.

To determine whether the accent invariance of the metaphoneme transcriptions brings useful information to our models, we also experimented with phonemic transcriptions. For this, we used the CMU pronunciation dictionary, containing pronunciations of English words as they're commonly realized in North American accents. Both the phoneme and metaphoneme dictionaries were expanded using grapheme-to-phoneme conversion to contain all words in our datasets. The metaphonemic lexicon uses 632 unique symbols, while the phonemic one has 72. The secondary output is used in training and adaptation, but is not used during the decoding process.

The baseline models were trained for 120 epochs. When running adaptation on the separate accents, we fine-tuned all parameters of the model for 15 epochs. The CTC loss had a weight of 0.3 and, where used, our secondary task loss had a weight of 0.2.

During decoding, a language model trained on the textual data released with LibriSpeech was combined with the models using shallow fusion. Our language model consists of 16 transformer layers using the same hyperparameters as the speech recognition model. We used the ESPnet toolkit for our experiments.

### 4.1. Data

Our experiments make use of the English portion of Mozilla Common Voice, a crowd-sourced and crowd-validated corpus of prompted speech [30]. During the collection of the speakers were asked to self-report various information, including accent spoken. Reporting the accent information by the participants was made optional; we made use of the data where accent labels are available.

Note that Common Voice has multiple releases and we are using the second version of the corpus. Data collection and labeling conventions vary between releases. In the second version of the corpus only accents from nations where English is an official language could be reported.

We split the English data for which accent labels are present into new train, development and test sets using the Mozilla CorporaCreator tool, so as to make our setup reproducible. Details of this setup are in Tables 2 and 3, listing the amounts of data and speakers for each accent.

For our experiments, we created a mixed accent set by pooling the England, US and Australian accented sections of the training set, resulting in ~145 hours of data.

Some of our experiments use LibriSpeech, a corpus of read speech from audiobooks [31]. Similarly to Common Voice, the recordings are crowd-sourced but the validation process used hybrid speech recognition models to align and filter the data. As the hybrid models used for the validation process were based on VoxForge English and Wall Street Journal data, we can expect LibriSpeech to have recordings of accents closer to North American English [32, 33].

## 5. RESULTS

### 5.1. Common Voice

For models trained on ~145 hours of Common Voice English, US and Australia pooled data, our results show the multi-task training is, on average, slightly more robust to accent variability than the single task model. The single task model has better results than the metaphoneme model only on the Irish,

| Accent | Duration | Utterances | Speakers |
|---|---|---|---|
| US | 100.5 hrs | 68527 | 886 |
| England | 26.1 hrs | 18280 | 291 |
| Australia | 18.0 hrs | 12046 | 102 |
| Canada | 13.4 hrs | 8883 | 127 |
| Scotland | 6.6 hrs | 3716 | 25 |
| Ireland | 2.5 hrs | 1612 | 18 |
| African | 2.0 hrs | 1366 | 31 |
| Philippines | 59.0 min | 691 | 12 |
| Singapore | 45.1 min | 489 | 6 |
| Malaysia | 16.7 min | 205 | 6 |
| Hong Kong | 1.5 min | 21 | 5 |
| Bermuda | 0.4 min | 6 | 4 |
| **TOTAL** | 171.4 hrs | 115729 | 1523 |

**Table 2**. Amount of data in relation to number of speakers and utterances for different accents in our Common Voice train set. The table is ordered by hours of data in the train set.

| Accent | Duration | Utterances | Speakers |
|---|---|---|---|
| US | 13.2 hrs | 9026 | 2050 |
| England | 3.7 hrs | 2488 | 590 |
| Australia | 55.2 min | 631 | 154 |
| Canada | 1.5 hrs | 998 | 215 |
| Scotland | 15.5 min | 178 | 41 |
| Ireland | 28.0 min | 253 | 57 |
| African | 31.9 min | 344 | 77 |
| Philippines | 16.2 min | 173 | 43 |
| Singapore | 5.5 min | 58 | 18 |
| Malaysia | 11.9 min | 134 | 30 |
| Hong Kong | 7.5 min | 82 | 22 |
| Bermuda | 5.3 min | 59 | 19 |
| *India* | 3.1 hrs | 2003 | 624 |
| *New Zealand* | 10.9 min | 120 | 34 |
| **TOTAL** | 24.8 hrs | 16671 | 4000 |

**Table 3**. Amount of data in relation to number of speakers and utterances for different accents in our Common Voice test set. The table is ordered by hours of data in the train set. In italic, accents not present in the train set.

Scottish and African data. A possible explanation is the phonetic Levenshtein distance of those accents from the ones in our pooled training set [34]. For the accents seen in the training set, using metaphonemes as a secondary task results in consistently better word error rate than the single task training and, on average, better results than the model trained on phonemes as targets, showing the model can make good use of the expert knowledge given. Full results are in table 4.

### 5.2. LibriSpeech

We trained baseline models on the full ~960 hours of training data in the LibriSpeech corpus. Our results for the single task model mirror those from Karita et al. [29], which uses the same transformer architecture. The models trained on phonemes or metaphonemes as the secondary task have a very similar performance to the single task architecture on the LibriSpeech development and test sets (table 5).

When decoding the accented speech in Common Voice, the multitask models trained on the full LibriSpeech data had, on average, worse results than the single task model (table 6).

After adapting the models trained on the full LibriSpeech data to the accents in Common Voice separately, the single task model is, on average, more accurate than the multitask models. In relative terms, the model adapted using metaphonemes as a secondary task has an average relative improvement of 11.5% over its baseline, a slight edge over the single task model's improvement of 9.6%.

The metaphonemic model consistently outperforms the phonemic model. Furthermore, the metaphonemic adaptation is the only one having accuracy improvements over all accents, while the single task and phonemic models degrade in two of the lower resourced accents, Hong Kong and Singapore. As the unadapted multitask models have similar performance and identical architecture and training, the higher and

| Accent | BPE | MT (BPE + MPH) | MT (BPE + PHN) |
|---|---|---|---|
| African | 18.1 | 18.4 | 18.1 |
| Australia | 12.5 | 11.8 | 13.1 |
| Bermuda | 15.1 | 15.6 | 14.5 |
| Canada | 13.4 | 12.9 | 12.5 |
| England | 15.7 | 15.4 | 15.0 |
| Hongkong | 28.7 | 27.7 | 28.9 |
| Indian | 27.5 | 26.0 | 26.6 |
| Ireland | 16.1 | 18.4 | 16.5 |
| Malaysia | 20.4 | 20.1 | 18.8 |
| New Zealand | 16.5 | 15.8 | 15.9 |
| Philippines | 25.3 | 24.9 | 24.9 |
| Scotland | 16.4 | 17.3 | 16.8 |
| Singapore | 17.5 | 15.3 | 18.7 |
| US | 14.4 | 13.6 | 13.6 |
| **Avg. seen** | 14.2 | 13.6 | 13.9 |
| **Avg. unseen** | 19.5 | 19.3 | 19.3 |

**Table 4**. WER (%) results for models trained either using only BPEs or in a multitask configuration with metaphonemes or phonemes. The models are trained on CommmonVoice US, Australian and England pooled training data.

| Model | Dev | Dev – other | Test | Test – other |
|---|---|---|---|---|
| BPE | 2.3 | 5.6 | 2.6 | 5.7 |
| MT (BPE + MPH) | 2.3 | 5.6 | 2.7 | 5.8 |
| MT (BPE + PHN) | 2.2 | 5.6 | 2.6 | 6.0 |

**Table 5**. WER (%) results for models trained either using only BPEs or in a multitask configuration with metaphonemes or phonemes. The model is trained on ~960 hours of LibriSpeech data.

| Accent | BPE | MT (BPE + MPH) | MT (BPE + PHN) |
|---|---|---|---|
| African | 19.8 | 18.9 | 20.2 |
| Australia | 12.3 | 12.7 | 12.6 |
| Bermuda | 13.8 | 14.0 | 14.7 |
| Canada | 14.2 | 14.1 | 14.4 |
| England | 16.2 | 16.7 | 17.0 |
| Hong Kong | 32.3 | 34.5 | 32.5 |
| Indian | 29.8 | 30.5 | 31.3 |
| Ireland | 16.7 | 17.6 | 17.0 |
| Malaysia | 22.6 | 24.2 | 23.9 |
| New Zealand | 14.3 | 16.5 | 16.8 |
| Philippines | 22.9 | 25.2 | 25.0 |
| Scotland | 17.4 | 19.1 | 18.1 |
| Singapore | 17.6 | 18.9 | 19.6 |
| US | 15.2 | 15.4 | 15.9 |
| **Average** | 18.9 | 19.9 | 19.9 |
| **Adapt average** | 18.4 | 19.3 | 19.2 |

**Table 6**. Baseline WER (%) results for models trained either using only BPEs or in a multitask configuration with metaphonemes or phonemes. The models are trained on ~960 hours of LibriSpeech data and tested on Common Voice data. Adapt average is the average of accents that do have a train set; this average is comparable to the one in table 7.

| Accent | BPE | MT (BPE + MPH) | MT (BPE + PHN) |
|---|---|---|---|
| African | 16.2 | 17.1 | 17.1 |
| Australia | 10.1 | 9.8 | 10.4 |
| Bermuda | 13.8 | 14.0 | 14.7 |
| Canada | 12.1 | 12.4 | 12.9 |
| England | 13.3 | 13.1 | 14.2 |
| Hong Kong | 33.9 | 34.2 | 33.2 |
| Ireland | 14.9 | 15.6 | 15.1 |
| Malaysia | 19.2 | 20.9 | 21.4 |
| Philippines | 20.8 | 22.9 | 22.9 |
| Scotland | 15.9 | 16.2 | 16.9 |
| Singapore | 18.7 | 17.5 | 21.6 |
| US | 10.8 | 11.0 | 11.5 |
| **Average** | 16.6 | 17.1 | 17.7 |

**Table 7**. WER (%) results for models adapted either using only BPEs or in a multitask configuration with metaphonemes or phonemes. The seed models are trained on ~960 hours of LibriSpeech data and fine-tuned on the Common Voice accents separately.

more consistent gains during adaptation can be attributed to the choice of secondary targets. This indicates that the accent independence of the metaphonemes brings useful and relevant information for accent adaptation. Full results for the adaptation experiments are in table 7.

### 5.3. Ablation study

To have a clearer picture of how the information from the secondary tasks is used by the model, we adapted the LibriSpeech baseline models on two accents of the Common Voice corpus, Australian and Canadian, experimenting with adapting the three seed models in single and multitask configurations.

| Model | WER Australia | WER Canada |
|---|---|---|
| Single task base | 12.3 | 14.2 |
| Adapted – BPE only | 10.1 | 12.1 |
| Adapted – MPH only | —- | —- |
| Adapted – CMU only | —- | —- |
| Adapted – BPE + MPH | 9.8 | 12.2 |
| Adapted – BPE + PHN | 9.8 | 11.9 |
| MT base (BPE + MPH) | 12.7 | 14.1 |
| Adapted – BPE only | 10.4 | 13.1 |
| Adapted – MPH only | 11.2 | 13.3 |
| Adapted – BPE + MPH | 9.8 | 12.4 |
| MT base (BPE + PHN) | 12.6 | 14.4 |
| Adapted – BPE only | 10.9 | 13.5 |
| Adapted – PHN only | 12.5 | 15.0 |
| Adapted – BPE + PHN | 10.4 | 12.9 |

**Table 8**. WER (%) results for single and multi-task models adapted in different configurations. The seed models are trained on ~960 hours of LibriSpeech data and fine-tuned on Common Voice England English data. Missing results indicate models that diverged during adaptation.

The results in table 8 show that the single task models did not converge when trying to fine-tune using only the secondary task; this is expected, as the model hasn't seen this information during the original training. When fine-tuning the single task baseline with an additional secondary task, the model has slight fluctuation in accuracy over using only the BPE labels, which we believe to be a regularization effect.

More interestingly, the metaphonemic models rely on both tasks. Both sources of information have a positive impact on the accuracy of the model when used separately, and the combination of the tasks outperforms their separate improvement. In comparison, the phonemic models don't seem to rely on the secondary task as much, with the performance on Canadian degrading when adapting only on phonemes.

Note that the secondary task is ignored during the beam search and, as it is output from a lower layer, it doesn't adapt the last decoder layer. Thus, the improvements in WER when adapting a multitask model relying only on the secondary task show that the models managed to learn a structured representation that informs the final BPE output, and that is possible to adapt this representation with a positive impact on the output.

We then tried to determine why the single task model has better absolute performance than the multi-task model on the LibriSpeech dataset. We hypothesized this degradation to be an effect of the lower accent variability in the training data, as the model can't learn a more robust representation without seeing the shifts in pronunciation in different accents.

| Model | Dev | Dev – other | Test | Test – other |
|---|---|---|---|---|
| BPE | 5.3 | 14.2 | 5.9 | 14.8 |
| MT (BPE + MPH) | 4.9 | 13.7 | 5.3 | 13.9 |
| MT (BPE + PHN) | 4.9 | 13.4 | 5.5 | 14.1 |

**Table 9**. WER (%) results for models trained either using only BPEs or in a multitask configuration with metaphonemes or phonemes. The model is trained on ~100 hours of clean LibriSpeech data.

| Accent | BPE | MT (BPE + MPH) | MT (BPE + PHN) |
|---|---|---|---|
| African | 45.6 | 41.2 | 41.2 |
| Australia | 28.9 | 27.8 | 26.3 |
| Bermuda | 29.6 | 26.1 | 29.4 |
| Canada | 25.9 | 24.7 | 24.7 |
| England | 37.8 | 36.3 | 36.4 |
| Hongkong | 55.1 | 49.3 | 54.5 |
| Indian | 65.4 | 62.1 | 63.3 |
| Ireland | 31.8 | 30.6 | 30.5 |
| Malaysia | 47.2 | 42.7 | 45.5 |
| New Zealand | 38.0 | 37.1 | 36.8 |
| Philippines | 43.1 | 41.8 | 41.7 |
| Scotland | 36.3 | 32.5 | 32.7 |
| Singapore | 41.3 | 37.6 | 42.5 |
| US | 29.6 | 28.2 | 28.0 |
| **Average** | 39.7 | 37.0 | 38.1 |

**Table 10**. Baseline WER (%) results for models trained either using only BPEs or in a multitask configuration with metaphonemes or phonemes. The models are trained on ~100 hours of clean LibriSpeech data and tested on Common Voice data.

To determine how the accent variability seen by the model and the hours of training data affect our models, we decided to train on the clean training set of LibriSpeech. The ~100 hours of clean data in LibriSpeech were originally selected by decoding using a model trained on the WSJ corpus and taking the utterances with the lowest WER. This means that the clean set is closer in realization to North American English, as the utterances matching the domain of the WSJ corpus will have lower error rates.

The results in tables 9 and 10 show the multitask models outperforming the single task models. This improvement is consistent on both the LibriSpeech and Common Voice test sets, with the metaphonemic model having a relative gain of about 6% over its single task counterpart. Additionally, while the two multitask models perform similarly on LibriSpeech data the phonemic model is, on average, less robust to the accents present in the Common Voice corpus.

Since the clean set of LibriSpeech has less accent variability, the fact that the model using metaphonemes outperforms the single task model on the Common Voice data is surprising when compared to the reverse results when using the full training set. The full training set sees more variability in accent realization, so we'd expect to see an increase in performance at least on the accented data.

This indicates that with higher amounts of data, the model reaches a better internal representation when unconstrained by the secondary task. As our other experiments show that the auxiliary information is valuable and synergistic with the primary task, we believe the degradation on the full LibriSpeech set might be due to inadequate use of the information provided, either by under-fitting or lacking modelling power. A better matching architecture or hyperparameters could lead to gains in performance on larger datasets.

## 6. CONCLUSION AND FUTURE WORK

We presented a multitask method making use of additional linguistic information to help structure the internal representation of the model. We showed how this representational constraint can help, when used in conjunction with relevant expert knowledge, to increase model robustness and adaptation power, especially when training on lower amounts of data.

In the future, we plan to expand the experiments to better determine how the accent variability and data amounts seen by the model affect results, and whether model and hyperparameter choices can lead to better accuracy. Further exploration will pertain whether it's possible to obtain similar information to metaphonemic transcriptions in an unsupervised way, such as finding accent independent mappings through sound class frequencies, phonotactics and word level transcriptions.

## 7. ACKNOWLEDGMENTS

This work was supported by a project funded by Samsung Electronics Co., Ltd. (Samsung Research).

## 8. REFERENCES

[1] Carlos Teixeira, Isabel Trancoso, and António Serralheiro, "Accent identification," in *ICSLP*, 1996.

[2] Ghinwa Choueiter, Geoffrey Zweig, and Patrick Nguyen, "An empirical study of automatic accent classification," in *ICASSP*, 2008.

[3] Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals, "Automatic dialect detection in Arabic broadcast speech," in *Interspeech*, 2016.

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

[5] Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri, "The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech," in *ASRU*, 2019.

[6] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Interspeech*, 2016.

[7] Peter Smit, Siva Reddy Gangireddy, Seppo Enarvi, Sami Virpioja, and Mikko Kurimo, "Aalto system for the 2017 Arabic multi-genre broadcast challenge," in *ASRU*, 2017.

[8] Sameer Khurana, Ahmed Ali, and James Glass, "DARTS: Dialectal arabic transcription system," *arXiv preprint arXiv:1909.12163*, 2019.

[9] Kanishka Rao and Haşim Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *ICASSP*, 2017.

[10] Mohamed Elfeky, Meysam Bastani, Xavier Velez, Pedro Moreno, and Austin Waters, "Towards acoustic model unification across dialects," in *SLT*, 2016.

[11] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *ICASSP*, 2018.

[12] Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *ICASSP*, 2018.

[13] Abhinav Jain, Minali Upreti, and Preethi Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Interspeech*, 2018.

[14] Thibault Viglino, Petr Motlicek, and Milos Cernak, "End-to-end accented speech recognition," in *Interspeech*, 2019.

[15] Mikaela Grace, Meysam Bastani, and Eugene Weinstein, "Occam's adaptation: A comparison of interpolation of bases adaptation methods for multi-dialect acoustic modeling with LSTMS," in *SLT*, 2018.

[16] Abhinav Jain, Vishwanath P Singh, and Shakti P Rath, "A multi-accent acoustic model using mixture of experts for speech recognition," in *Interspeech*, 2019.

[17] Sanghyun Yoo, Inchul Song, and Yoshua Bengio, "A highly adaptive acoustic model for accurate multi-dialect speech recognition," in *ICASSP*, 2019.

[18] Han Zhu, Li Wang, Pengyuan Zhang, and Yonghong Yan, "Multi-accent adaptation based on gate mechanism," in *Interspeech*, 2019.

[19] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung, "Learning fast adaptation on cross-accented speech recognition," in *Interspeech*, 2020.

[20] Brady Houston and Katrin Kirchhoff, "Continual learning for multi-dialect acoustic models," *Interspeech*, 2020.

[21] Samik Sadhu and Hynek Hermansky, "Continual learning in automatic speech recognition," *Interspeech*, 2020.

[22] Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu, "Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition," in *Interspeech*, 2017.

[23] Weiran Wang, Yingbo Zhou, Caiming Xiong, and Richard Socher, "An investigation of phone-based subword units for end-to-end speech recognition," in *Interspeech*, 2020.

[24] Amit Das, Jinyu Li, Guoli Ye, Rui Zhao, and Yifan Gong, "Advancing acoustic-to-word ctc model with attention and mixed-units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1880–1892, 2019.

[25] Sei Ueno, Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara, "Acoustic-to-word attention-based model complemented with character-level ctc-based model," in *ICASSP*, 2018.

[26] Susan Fitt, "Documentation and user guide to UNISYN lexicon and post-lexical rules," Tech. Rep., Centre for Speech Technology Research, University of Edinburgh, 2000.

[27] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, 2017.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[29] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al., "A comparative study on transformer vs RNN in speech applications," in *ASRU*, 2019.

[30] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[31] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[32] "Free speech recognition - voxforge.org," http://www.voxforge.org/.

[33] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *Workshop on Speech and Natural Language*, 1992.

[34] Søren Wichmann and Matthias Urban, "Toward an automated classification of englishes," in *The Oxford Handbook of the History of English*, Terttu Nevalainen and Elizabeth Closs Traugott, Eds. Oxford University Press, 2012.