Normalization of Gender, Dialect and Speaking style using Probabilistic front-ends

Udhyakumar Nallasamy[1], Florian Metze[1], Thomas Schaaf[2]

[1] *InterACT Labs, Carnegie Mellon University, Pittsburgh, PA, USA, E-Mail:udhay@cmu.edu, fmetze@cs.cmu.edu*

[2] *M*Modal technologies, Pittsburgh, PA, USA, E-Mail: tschaaf@mmodal.com*

Abstract

This paper analyzes the capability of probabilistic Multilayer Perceptron (MLP) front-end to perform various normalizations for robust Automatic Speech Recognition (ASR). We find decision trees to be a useful tool for investigating the normalization of the feature space achieved by various front-ends. We introduce additional questions for different environmental conditions to the training of the phonetic context decision tree, and count the number of splits dedicated to lexical discrimination using context, and to these environmental conditions. We compare (1) Bottle-Neck (BN) features and (2) standard stacked Mel Frequency Cepstral Coefficients (MFCC) with LDA. In previous work, we found the BN front-end to be effective in reducing the number of gender questions than MFCC, which may be part of the reason why BN front-ends can achieve significant improvements. In this work, we extend this approach to the analysis of dialect on a large database of Pan-Arabic speech.

Introduction

MFCC features have been the standard front-end for Hidden Markov Model (HMM) based ASRs for over a number of years. Probabilistic MLP features were introduced as an alternative to MFCCs, with inherent phonetic discrimination [1]. With the introduction of large databases, MLP features have become popular recently [2-3]. They produce significantly reduced Word Error Rates (WER) in an end-to-end system, particularly when combined with their MFCC counterparts in some fashion, viz, feature fusion, multi-stream score combination or final hypothesis combination [4]. The standard techniques in building a state-of-the-art ASR including model (MLLR) and feature (FSA) adaptation, input decorrelation using LDA, semi-tied covariance (STC) are found effective with MLP features.

There are many variants to training a multi-layer neural network to obtain BN features. In general, a four layer network is used to map a group of windowed input frames to a set of pre-defined targets [5]. The input frames can be from various sources including, MFCCs, PLPs or outputs from a previous MLP. Different targets have also been tried in the literature including, phones, HMM states, phonetic/articulatory units [6]. The activation from the 3rd layer, also known as “bottle-neck” layer with relatively smaller number of units is extracted as front-end features. The reduced units in the bottle-neck layer is expected to perform a non-linear projection of the input feature space to a lower dimension, retaining only the information required to discriminate the target classes [7]. These features are then used alone or stacked with MFCCs, with necessary decorrelation and modelled with Gaussian Mixture Models (GMMs) as in standard HMM-based ASR training.

In our previous work [8], we analyzed the capability of MLP features with respect to gender normalization. In this paper, we focus on the behaviour of MLP features under the influence of dialect.

Experiment Design

We include questions representing each dialect, in addition to contextual questions while building the decision tree. We then calculate the ratio of leaf nodes under dialectal questions. This value is treated as a measure of dialect normalization - higher the ratio, more tied-contexts have the influence of dialect and vice-versa. An example of the decision tree with both dialectal and non-dialectal leaf nodes is shown below.

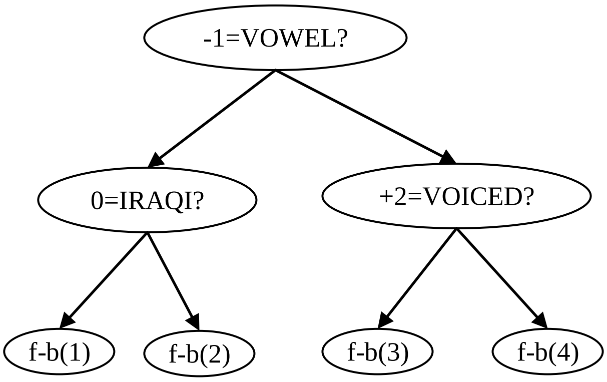

Figure 1: Contextual decision tree with dialectal and non-dialectal questions (+ Right context, - Left context)

In the above figure, the begin state of phoneme /f/ is clustered into 4 states. f-b(1) and f-b(2) are considered dialectal models, as they are derived by choosing a dialect question (Is current phone belong to IRAQI dialect?). f-b(3) and f-b(4) are non-dialectal models, because their derivation doesnt involve a dialect question in the decision tree.

Pan-Arabic database

We used Pan-Arabic database for our experiments. The database consists of spoken Arabic speech, transcriptions and lexicons for 5 different dialects: UAE (Gulf), Egyptian, Syrian, Palestinian and Iraqi. It has a total of 200 recordings sessions with 2 speakers per session, totalling 150 hours. The dictionary consists of 42 phones and is formed by combining the dialect-specific dictionaries. The audio sessions include both conversational and scripted recordings, although the latter is used in the experiments reported in this paper. The first five sessions in each dialect are used as test set, while the remaining data is used for training the ASR.

Speech recognizer setup

The speech recognizer consists of speaker independent acoustic models with 2000 codebooks and maximum of 32 gaussians per distribution. The language model (LM) is a trigram model trained on the audio transcriptions and broadcast news data, with an OOV rate of 2.4. We built 2 systems with different front-end processing namely, MFCC and MLP. MLP uses a 4-layer archiecture with the following configuration, 195x3000x42x111. We used ICSI QuickNet toolbox [9] to train the MLP. The training set is divided into *mlp-train* (90%) and *mlp-tune* (10%) sets. The parameters are estimated on the *mlp-train* set using back-propagation for each iteration. The training is stopped once the accuracy on the *mlp-tune* set saturates. The final MLP obtained a frame-level accuracy of 63.86% on *mlp-train* and 63.56% on *mlp-tune*. The outputs from the 3rd, BN layer is used as MLP features, and the HMM-GMM ASR is built similar to MFCC system. Note that the phonetic decisison trees used to build these baseline systems donot use any gender or dialect questions. The word error rates for both MFCC and MLP ASR systems is shown below.

Table 2: WER for MFCC and MLP systems

System	WER
MFCC	28.7%
MLP	28.1%

Decision Tree based analysis

Decision trees are trained to cluster the phonetic contexts in the training data. Each phoneme was tagged with the dialect, which allows the dialect questions to be used in the training process in addition to the phonetic questions. The number of dialect-dependent leaves is computed by tracing the tree in reverse, from each leaf node back to the root. If the trace for a leaf node encountered a dialect question, it is considered to be a dialectal context. The ratio of dialectal nodes to the total nodes is computed for both MFCC and MLP systems. The experiment is repeated by varying the size of decision tree, i.e number of clustered contexts. The following graph shows the number of dialectal and non-dialectal nodes for MFCC and MLP front-ends.

Table 3: Ratio of dialectal nodes in MFCC and MLP

Size	Dialect nodes	Non-Dialect nodes	Ratio	Dialect Nodes	Non-Dialect nodes	Ratio
	MFCC			MFCC (VTLN + FSA)		
1000	13	987	1.3%	9	991	0.9%
2000	82	1918	4.1%	72	1928	3.6%
3000	224	2776	7.5%	226	2774	7.5%
4000	483	3517	12.1%	465	3535	11.6%
	MLP			MLP (VTLN + FSA)		
1000	17	983	1.7%	14	986	1.4%
2000	99	1901	5.0%	73	1927	3.7%
3000	278	2722	9.3%	240	2760	8%
4000	589	3411	14.7%	524	3476	13.1%

It can be seen that the speaker adaptation reduces the number of dialectal contexts in the decision tree compared to unadapted models. With respect to MFCC and MLP systems, MFCC has lesser dialectal contexts than MLP. This is in contradiction with the experiments we conducted for gender, where MLP was less sensitive to gender variations than MFCC. Hence we conducted more rigorous experiments involving dialect normalization which include, using a single pronunciation dictionary to increase phonetic ambiguity, adding complex dialectal questions instead of just singleton questions, etc. In all the experiments, MFCC had fewer dialectal contexts than MLP.

Conclusion

We have analyzed the behaviour of MLP features under the influence of dialect and we confirm that while MLP front-end is robust against speaker-specific variations, aka gender, and improves the accuracy of ASR, it is more sensitive to linguistic variations, aka dialect. Future work will include testing this hypothesis on different language/dialect combinations, comparing the behaviour of these front-ends for other conditions like speaking styles, noise levels, etc.

Acknowledgements

This work is funded by the U.S ARFL/RIEC under contract FA8750-09-C-0067. Any opinions, findings, conclusions and/or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the sponsor.

References

[1] H. Hermansky, D. Ellis and S. Sharma, Tandem connectionist feature extraction for conventional HMM systems, Proc. ICASSP, Istanbul, June 2000

[2] Q. Zhu, B. Chen, N. Morgan and A. Stolcke, On Using MLP features in LVCSR, Proc. Interspeech, 2004

[3] J. Park, F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland, Training and adapting MLP features for Arabic speech recognition, Proc. ICASSP, 2009

[4] C. Ma, H. J. Kuo, H. Soltau, X. Cui, U. Chaudhari, L. Mangu and C.H. Lee, A comparative study on system combination schemes for LVCSR, Proc. Interspeech, 2010

[5] F. Grézl and P. Fousek, Optimizing bottle-neck features for LVCSR, Proc. ICASSP, 2008

[6] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and O. Cetin, Articulatory feature classifiers trained on 2000 hours of telephone speech, Proc. Interspeech, August 2007.

[7] F. Grézl, M. Karafiát and L. Burget, Investigation into bottleneck features for meeting speech recognition, Proc. Interspeech, 2009

[8] T. Schaaf and F. Metze, Analysis of gender normalization using MLP and VTLN features, Proc. Interspeech, 2010

[9] ICSI QuickNet toolbox - http://www.icsi.berkeley.edu/Speech/qn.html