2018 17th IEEE International Conference on Machine Learning and Applications

Modifying LSTM Posteriors with Manner of Articulation Knowledge to Improve Speech Recognition Performance

Pradeep R[1], K Sreenivasa Rao[2]
[1]TCS Research Scholar, [2]Professor
Dept. of Computer Science and Engineering
Indian Institute of Technology Kharagpur, Kharagpur - 721302, India
Email: {pradeep_raj31, ksrao}@iitkgp.ac.in

***Abstract*—The variant of recurrent neural networks (RNN) such as long short-term memory (LSTM) is successful in sequence modelling such as automatic speech recognition (ASR) framework. However the decoded sequence is prune to have false substitutions, insertions and deletions. We exploit the spectral flatness measure (SFM) computed on the magnitude linear prediction (LP) spectrum to detect two broad manners of articulation namely sonorants and obstruents. In this paper, we modify the posteriors generated at the output layer of LSTM according to the manner of articulation detection. The modified posteriors are given to the conventional decoding graph to minimize the false substitutions and insertions. The proposed method decreased the phone error rate (PER) by nearly 0.7 % and 0.3 % when evaluated on core TIMIT test corpus as compared to the conventional decoding involved in the deep neural networks (DNN) and the state of the art LSTM respectively.**

***Keywords*—phoneme posteriors, spectral flatness measure (SFM), sonorant detection rate (SDR), phone recognition**

I. Introduction

In recent years, deep neural networks (DNNs) combined with hidden Markov models (HMM) have become the dominant approach in acoustic modeling [1]. Based on increased computation power and quantity of data, substantial error rate reduction has been achieved for speech recognition tasks [2]. Recurrent neural networks (RNNs) as well as Long Short-Term Memory RNNs (LSTM RNNs) [3] are more suitable for sequence tasks such as sequence modeling and prediction, and have been helpful to improve robustness in ASR [4] .

The use of signal processing techniques in identifying discriminative information in speech signal has provided an insight to improve recognition performance [5]. In order to automatically segment speech into broad manner of articulation, it is essential to derive discriminative information particular to different manner of articulation [6]. Sonorants are the class of speech sounds that is produced with continuous, non-turbulent airflow in the vocal tract. Sonorants include vowels, semivowels and nasals. We exploit the spectral flatness measure (SFM) computed on the magnitude linear prediction (LP) spectrum for sonorant detection and use this information in ASR framework.The speech frames that are detected as sonorants are given to a sparse matrix generator where the values are high for sonorant IDs and low for obstruent phoneme IDs. The posteriors generated from the DNN/LSTM acoustic models (AM) are multiplied by the sparse matrix output and normalized to obtain modified posteriors. The modified posteriors are given to the conventional decoding graph to minimize the false substitutions and insertions.

We would like to thank Tata Consultancy Services (TCS) for sponsoring the research under TCS-Research Scholar Programme.

II. Background

The conventional method of decoding a test speech utterance is as illustrated in Figure 1 (a). The speech utterance is represented using standard features such as mel frequency cepstral co-efficients (MFCC) or feature space maximum likelihood linear regression (fMLLR) transformation [7].

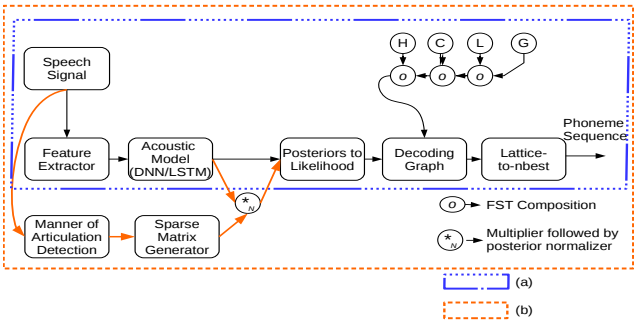

Fig. 1: Block diagram of (a) Baseline Decoding Mechanism in ASR (b) Proposed posterior modification framework

In order to arrive at phone likelihoods [8], we will first sum the pdf-id posteriors $p(i|x)$ and priors $p(i)$, where i denotes a pdf-id, over all pdf-ids that contribute to the same base phone, independent of HMM state, phone context, etc:

$$p(f|x) = \sum_{i \in f} (p(i)|x); p(f) = \sum_{i \in f} p(i) \quad (1)$$

where $i \in f$ indicates the mapping of pdf-id i to base phone f. The decoder computes the most likely word sequence $w = w_1 ... w_M$ for the observed speech signal, represented as sequence of acoustic feature vectors $x = x_1 ... x_T$:

$$\hat{w} = \arg\max_{w} (p(\mathbf{x}|w)p(w)) \quad (2)$$

The finite-state transducer (FST) framework provides well studied graph operations [9] which can be effectively used

978-1-5386-6805-4/18/$31.00 ©2018 IEEE
DOI 10.1109/ICMLA.2018.00122

for speech decoding. The decoding graph is represented as a weighted finite-state transducer (WFST) [10] that can be constructed using Eq (3).

$$HCLG = min(det(H \circ (C \circ (L \circ G)))) \quad (3)$$

where for the case of phoneme recognition, G is a phoneme grammar, L is a lexicon, C is the context dependency specification and H transforms sequences of senones (tied triphone states) to triphones. $\circ$ represents the composition operator.

A. Motivation

We believe that most of the sonorant sounds are produced using relatively less constricted vocal-tract shape and glottal vibration. This results in regions of regular structure having high energy and high degree of periodicity. Hence it is necessary to capture the manner of articulation transitions in the decoder to minimize falsely decoded phoneme sequences. The current research is focussed purely on acoustic model (AM) or lattice modification or embedding discriminative information to the features and retraining the AM. It is also important to observe the manner of articulation in the test utterance and adopt to modify the state of the art LSTM posteriors. We believe that the focus on explicitly modifying the DNN/LSTM posteriors by embedding manner of articulation knowledge is limited. Hence in our work we attempt to impose some restrictions in generating the posteriors according to the manner of articulation knowledge and focus in that direction.

III. Proposed Work

The fundamental blocks of the proposed posterior modification mechanism is illustrated in Figure 1 (b).

A. Manner of Articulation Detection

The block diagram for the proposed sonorant detection framework is shown in Figure 2 . Here the speech signal is analysed at short time intervals and for each frame the p^{th} order LPC is obtained. The LP spectrum is obtained by calculating the frequency response of the all pole digital filter where the denominator coefficients involve the LPCs. SFM

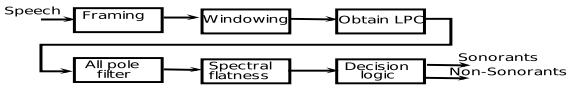

Fig. 2: Block diagram of the proposed sonorant detection framework

[11] is calculated on the magnitude of the LP spectrum for each frame using the Eq. (4) where *X(k)*

$$SFM = \frac{exp(\frac{1}{N}\sum_{k=0}^{k=N-1}|X(k)|)}{\frac{1}{N}\sum_{k=0}^{k=N-1}|X(k)|} \quad (4)$$

is the magnitude of LP spectrum, *k* varies from 0 to *N*-1 and *N* being the number of FFT points. Sonorants and non-sonorants are identified by the appropriate choice of SFM threshold.

B. Sparse Matrix Generator

Let us consider ten context independent acoustic models obtained for the phonemes /aa/$_S$, /ax/$_S$, /b/$_O$, /el/$_S$, /iy/$_S$, /ih/$_S$, /p/$_O$, /s/$_O$, /sh/$_O$ and /z/$_O$ for illustration. Here the phoneme /X/$_S$ and /Y/$_O$ indicate that the model /X/ belongs to sonorant manner and /Y/ belong to obstruent respectively. Let the decoded sequence for some speech utterance be "p iy ax el" which was supposed to be decoded as "p iy p el". This resulted in one substitution error where /p/ is substituted as /ax/. The

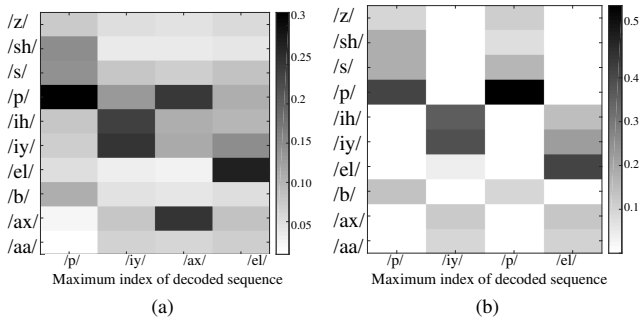

Fig. 3: (a) Baseline Posterior (b) Modified Posterior generated for the decoded sequence

posteriors for four segments of speech is shown as a gray scale image in Figure 3 (a). Higher the posterior for a speech segment, the darker (black) is the plot. From the figure, the posterior probability of /ax/ phone is more than that of /p/. Hence the final decoded utterance decodes as /ax/.

Sparse matrix is generated as per the manner of articulation knowledge embedded in speech segment. The size of sparse matrix is $m \times n$ where m is the number of segments in the decoded sequence, n is the number of phoneme models or the number of nodes at the output layer of DNN/LSTM.

$$M = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

In this example, the size of sparse matrix is 4×10. If the segment of speech belongs to sonorant manner then the index of sonorant IDs are made '1' else they are highlighted by '0'. Since the first and third segments of speech are assumed to be of obstruent manner, they should have same sparse values. The raw posteriors generated at the output layer are multiplied by the sparse matrix and normalized to obtain the modified posteriors. Figure 3 (b) shows the representation of the modified posteriors. We can observe that the third segment in the figure has only the posteriors related to obstruent manner. Hence the sonorant posteriors are forced to zero in the third segment as indicated by white regions in Figure 3 (b) . Now the system decodes the third segment from only the obstruent manner. In such a way, the sonorant identities being falsely decoded as non-sonorants are minimized using proposed method.

IV. Experiments

All phone recognition experiments are performed on the training set and evaluated on the test set of the TIMIT [1] corpus.

A. Variation of SFM on sonorants and obstruents

Figure 4 shows the plot of magnitude LP spectrum obtained at different frequency and its corresponding SFM for the class of sonorants (a) and non-sonorants (b). SFM variation on the

[1] https://catalog.ldc.upenn.edu/ldc93s1

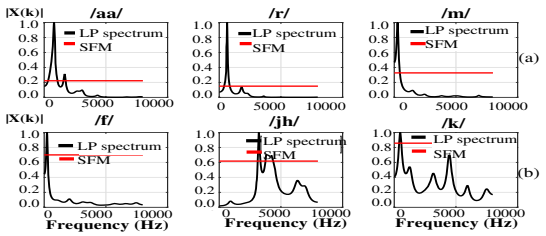

Fig. 4: Illustration of the SFM value on LP spectrum for (a) sonorants and (b) non-sonorants

magnitude spectrum varies for sonorants /aa/, /r/ and /m/ and non-sonorants /f/, /jh/ and /k/ is illustrated. It is observed that the SFM values per frame of sonorants phonemes /aa/, /r/ and /m/ is found to be 0.25, 0.18 and 0.35 and that for non-sonorants phonemes /f/, /jh/ and /k/ is found to be 0.7, 0.62 and 0.85 respectively. Since the class of sonorants doesn't undergo any constrictions in the vocal tract, the magnitude LP spectrum is relatively smoother and hence the SFM on these segments produces low values. In contrast, high constriction (for plosive /k/ and affricate /jh/) and noise information (for fricatives /f/) play dominant role for the obstruent sound production. As a result the SFM for non-sonorants has high values for noise-dominated regions which have a relatively flat spectrum.

The sonorant detection rate (SDR) is calculated using $\%SDR = \frac{T_p+T_n}{T_p+F_p+F_n+T_n}$ where T_p, T_n, F_p and F_n are the true-positive, true-negative, false-positive and false-negative rates respectively. T_p for sonorants correctly identified as sonorants, T_n for obstruents incorrectly identified as sonorants, F_p for obstruents correctly identified as obstruents and F_n for sonorants incorrectly identified as obstruents. The detection rate is obtained for different SFM threshold varied from 0.1 to 1.0 in steps of 0.01. Since the SFM value of 0.5 showed maximum detection rate of 0.943 on test set of TIMIT, we fix this as the sonorant-obstruent discriminating index. This is to say that a speech frame whose SFM value is lesser than 0.50 is treated as sonorant else its an obstruent.

B. Sonorant detection framework in ASR

We used Pytorch-kaldi [2] for running the experiments on LSTM. The DNN part is managed by pytorch, while feature extraction, label computation, and decoding are performed with the kaldi toolkit [12].

1) System Details: The architecture adopted for the experiments consisted of multiple bi-LSTM recurrent layers [3], which were stacked together prior to the final softmax context-dependent classifier. We studied the impact of sonorant detection framework in ASR under different cases.

- Case-1: We trained the ASR system using HMM-GMM, DNN and LSTMs on MFCC+Δ +$\Delta\Delta$. The performance is evaluated on the baseline models.
- Case-2: The raw posteriors generated at the output layer of DNN/LSTM are multiplied by the sparse matrix and normalized to obtain the modified posteriors. The performance of the system (%PER) is measured on the modified posteriors.

[2]https://github.com/mravanelli/pytorch-kaldi/

V. Results and Discussion

In this section the performance of the sonorant detector and its impact in ASR performance is discussed.

A. Performance of the sonorant detector

The phone level TIMIT alignments are used to find the SDR for a particular phoneme. Higher the value of SDR indicates that all the frames in the expected alignment belongs to the class of sonorants; lower values indicate that SFM values of those frames are greater than threshold and hence they belong to the class of non-sonorants.

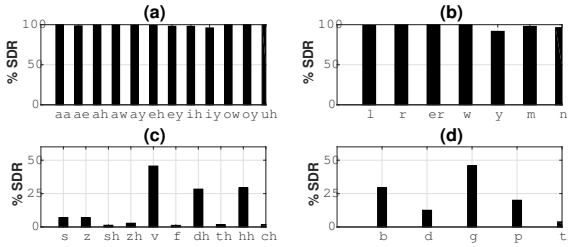

Fig. 5: Sonorant detection rate on different phonemes (a) Vowels (b) Semi-vowels and Nasals (c) Fricatives and Affricates (d) Stops

Figure 5 shows the SDR for (a) vowels, (b) semi-vowels and nasals, (c) fricatives, affricates and (d) stop consonants. It is observed that the SFM works well in detection of sonorants except for the fact that some of the non-sonorants show mixed characteristics. The voiced weaker fricatives such as /v/ and /dh/ and the voiced weaker stops such as /b/ and /g/ show the % SDR more than 20%. This is to say that 20% of weaker fricatives and stop consonants are falsely detected as sonorants. We believe that this false indication is due to the influence of vowels in their neighbouring context, low level voicing and also the duration of these phonemes is quite less. The overall SDR obtained on the core test set of TIMIT is 0.95. The further reduction and analysis of falsely detected phonemes is one of the future scopes of the paper.

B. Phoneme Recognition after sonorant detection

The DNN is initialized with stacked restricted Boltzmann machines (RBMs) that are pretrained in a greedy layerwise fashion [13]. The baseline LSTM models had 1945 pdf-ids, hence the fully connected layer dimension is $1945 \times (512 \times 2)$ (Bi-LSTM). Table I shows the performance of the context independent HMM-GMM , DNN and the state-of-the-art LSTM systems studied under different case studies indicating the number of deletions (D), substitutions (S) and insertions (I) errors. %Corr and %Acc are calculated using *(N-S-I)/N* and *(N-D-S-I)/N* respectively where *N* is the number of phonemes in the expected test transcriptions.

In order to visualize the 2D embeddings, a test utterance (dr1/faks0/sa1.wav - "She had your dark suit in greasy wash water all year") from the test set of TIMIT is considered. The test utterance is divided into short segment of 20 ms with 10 ms overlap and relevant MFCC+$\Delta + \Delta\Delta$ feature is extracted. There are 14 unique labels from the test utterance (/sh/, /iy/, /hv/, /ae/, /cl/, /d/, /y/, /er/, /k/, /s/, /aa/, r/, /w/ and /l/) and is

TABLE I: Overall PER obtained using HMM-GMM and DNN with different case studies

System	Cases	D	S	I	% Corr	% Acc	% PER
HMM-GMM	Case-1	686	1422	239	77	67.5	32.5
	Case-2	686	1400	217	77.6	68.1	**31.9**
DNN	Case-1	780	1241	188	80.2	69.4	30.6 (1-H)
		693	1140	217	81.2	71.6	28.4 (2-H)
		643	1097	195	82.1	73.2	26.8 (3-H)
	Case-2	758	1191	181	81	70.5	29.5 (1-H)
		686	1097	217	81.8	72.3	27.7 (2-H)
		635	1054	188	82.8	74	**26.1 (3-H)**
Bi-LSTM	Case-1	218	707	166	87.2	84.9	15.1
	Case-2	218	690	159	88.2	85.2	**14.8**

used for reference. t-Stochastic Neighbourhood Embedding (t-SNE) receives the high dimensional features and converts into 2 dimension [14]. Figure 6 (a) shows the 2D Visualization of DNN output layer activations obtained for MFCC+Δ + $\Delta\Delta$ feature at individual phoneme level. The features that belong to sonorants (yellow) or obstruents (blue) are highlighted in Figure 6 (b). Figure 7 (a) shows the 2D Visualization of

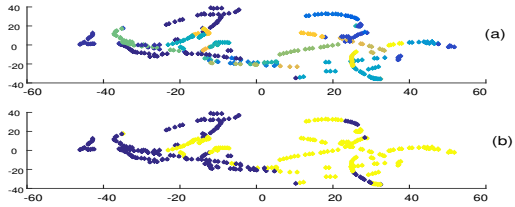

Fig. 6: 2D Visualization of DNN output layer activations for MFCC+Δ + $\Delta\Delta$ feature (a) Distribution of phonemes (b) sonorant-obstruent scatter

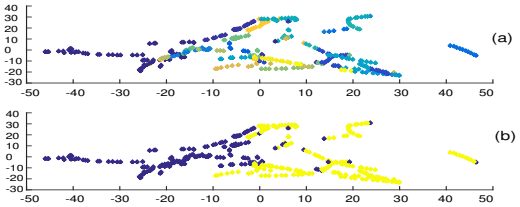

Fig. 7: 2D Visualization of DNN output layer activations modified using proposed method (a) Distribution of phonemes (b) sonorant-obstruent scatter

DNN output layer activations modified using proposed method for MFCC+Δ + $\Delta\Delta$ feature at individual phoneme level. The features that belong to sonorants (yellow) or obstruents (blue) are highlighted in Figure 7 (b). DNNs tries to cluster phonemes of similar category and will discriminate them from other. From the figure it can be observed that the degree of overlap among sonorants and obstruents is reduced as compared to that of Figure 6 (b).

DNNs can provide only limited temporal modeling and can only model the data within the window and are unsuited to handle different speaking rates and longer term dependencies. By contrast, LSTM contain cycles that feed the network activations from a previous time step as inputs to the network to influence predictions at the current time step. Hence LSTMs outperform over DNNs. However, the context information captured in LSTM doesn't explicitly consider the manner of articulation knowledge present in the test frames. Hence when we embed the manner of articulation detection knowledge in modifying the LSTM posteriors, false substitutions and insertions are further reduced.

VI. Conclusion

In this paper, automatic detection of sonorants based on spectral flatness measure is discussed. Sonorant detection scheme is applied to modify the LSTM posteriors at the ASR decoder. The modified posteriors are given to the conventional decoding graph to minimize the false substitutions and insertions. The proposed method decreased the phone error rate (PER) by nearly 0.7 % and 0.3 % when evaluated on core TIMIT test corpus as compared to the conventional decoding involved in the DNN and the state of the art LSTM respectively.

In future, we wish to gather the different manners of articulation knowledge in re-training LSTM.

References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[2] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

[3] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5755–5759, 2016.

[4] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[5] A. Juneja and C. Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition a," *Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1154–1168, 2008.

[6] P. Schwarz, P. Matvjka, and J. Cernocky, "Towards lower error rates in phoneme recognition," in *Proc. International Conference on Text, Speech and Dialogue*, Springer, pp. 465–472, 2004.

[7] M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[8] D. A. van Leeuwen and J. van Doremalen, "Calibration of phone likelihoods in automatic speech recognition," *arXiv preprint arXiv:1606.04317*, 2016.

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

[10] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiát, S. Kombrink, P. Motlivcek, Y. Qian *et al.*, "Generating exact lattices in the WFST framework," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4213–4216, 2012.

[11] N. Madhu, "Note on measures for spectral flatness," *Electronics letters*, vol. 45, no. 23, pp. 1195–1196, 2009.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, no. EPFL-CONF-192584, IEEE Signal Processing Society, 2011.

[13] A. R. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4273–4276, 2012.

[14] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.