Sparse Modeling of Neural Network Posterior Probabilities for Exemplar-Based Speech Recognition

Pranay Dighe*†, Afsaneh Asaei*, and Hervé Bourlard*†
*Idiap Research Institute, Martigny, Switzerland
†École Polytechnique Fédérale de Lausanne, Switzerland
Emails: {pranay.dighe, afsaneh.asaei, herve.bourlard}@idiap.ch

We study automatic speech recognition by direct use of acoustic features (exemplars) without any assumption on the underlying stochastic process. The prior studies exploit spectral exemplars. In this work, we present the use of neural network sub-word posterior probabilities as exemplars. The space of sub-word observations is low-dimensional (e.g. $\mathbb{R}^{K\times T}$) whereas the word transcription requires reconstructing a high-dimensional representation (e.g. $\mathbb{R}^{L\times T}$, $L \gg K$). Given the prior knowledge that for any given utterance, the word representation is highly sparse, we cast the speech recognition problem as sparse reconstruction of word posteriors given the compressed (low-dimensional) acoustic observation.

The sub-word units (phones) are denoted by $\{q_k\}_{k=1}^{K}$. Given an input spectral feature x_t at time t, a (deep) neural network [1] is used to estimate the posterior probabilities $\{p(q_k|x_t)\}_{k=1}^{K}$. The phone posterior probabilities is related to the word posterior probabilities $p(w_l|x_t)$ through

$$p(q_k|x_t) = \sum_{l=1}^{L} p(q_k, w_l|x_t) = \sum_{l=1}^{L} p(q_k|w_l)p(w_l|x_t); \tag{1}$$

the last equality holds due to conditional independence of the acoustic observation and input speech given a super-phone lexical unit such as word. Defining an over-complete dictionary $\mathbf{D}$ such that the atoms are exemplars obtained by conditioning the phone posteriors on a different linguistic unit w_l, we have

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{z_t} = \underbrace{\begin{bmatrix} p(q_1|w_1) & \cdots & p(q_1|w_l) & \cdots & p(q_1|w_L) \\ p(q_2|w_1) & \cdots & p(q_2|w_l) & \cdots & p(q_2|w_L) \\ \vdots & & \vdots & & \\ p(q_K|w_1) & \cdots & p(q_K|w_l) & \cdots & p(q_K|w_L) \end{bmatrix}}_{\text{Dictionary: } \mathbf{D}=[d_1 \ldots d_l \ldots d_L]} \times \underbrace{\begin{bmatrix} p(w_1|x_t) \\ \vdots \\ p(w_l|x_t) \\ \vdots \\ p(w_L|x_t) \end{bmatrix}}_{\alpha_t} \tag{2}$$

Construction of the dictionary as described in (2) requires modeling the subspaces of each word using the acoustic features in terms of phone posterior probabilities. To that end, we learn word-specific dictionaries such that each column of the dictionary in (2), d_l has a sparse representation stated as

$$\underbrace{\begin{bmatrix} p(q_1|w_l) \\ p(q_2|w_l) \\ \vdots \\ p(q_K|w_l) \end{bmatrix}}_{d_l} = \underbrace{\begin{bmatrix} p(q_1|sw_1^{w_l}) & \cdots & p(q_1|sw_s^{w_l}) & \cdots & p(q_1|sw_{S_{w_l}}^{w_l}) \\ p(q_2|sw_1^{w_l}) & \cdots & p(q_2|sw_s^{w_l}) & \cdots & p(q_2|sw_{S_{w_l}}^{w_l}) \\ \vdots & & \vdots & & \\ p(q_K|sw_1^{w_l}) & \cdots & p(q_K|sw_s^{w_l}) & \cdots & p(q_K|sw_{S_{w_l}}^{w_l}) \end{bmatrix}}_{\text{Word manifold modeling dictionary: } \mathbf{D}_{w_l}} \times \begin{bmatrix} p(sw_1^{w_l}|w_l) \\ \vdots \\ p(sw_s^{w_l}|w_l) \\ \vdots \\ p(sw_{S_{w_l}}^{w_l}|w_l) \end{bmatrix} \tag{3}$$

where $sw_s^{w_l}$ denotes the s^{th} sub-word unit of the word w_l, S_{w_l} represents the total number of (over-complete) "bases" to model the sub-space of word w_l.

Equations (2) and (3) lead us to an intuitive and natural representation for continuous speech in terms of posterior features and word-to-subword hierarchical dictionaries. Thereby, the posterior-based sparse modeling dictionary is obtained as $\mathbf{D} = [\mathbf{D}_{w_1} \cdots \mathbf{D}_{w_l} \cdots \mathbf{D}_{w_L}]$. The dictionary $\mathbf{D}$, has an internal partitioning defined by the boundaries of individual sub-dictionaries $\mathbf{D}_{w_l}$. Ideally, an input posterior feature z_t belonging to a realization of word w_l, when sparse decoded using the dictionary above will have a sparse representation α_t such that only the atoms corresponding to the subdictionary $\mathbf{D}_{w_l}$, denoted as $\alpha_t^{w_l}$, will have non-zero values. $\alpha_t^{w_l}$ is expressed as

$$\begin{aligned} \alpha_t^{w_l} &= p(w_l|x_t)\left[p(sw_1^{w_l}|w_l)\ldots p(sw_s^{w_l}|w_l)\ldots p(sw_{S_{w_l}}^{w_l}|w_l)\right]^\top \\ \alpha_t &= \left[\alpha_t^{w_1}\ldots\alpha_t^{w_l}\ldots\alpha_t^{w_L}\right]^\top \end{aligned} \tag{4}$$

A sequence of posterior features $\mathbf{Z} = [z_1, \ldots z_t]$, extracted from an utterance of word w_l, will have a hierarchical group structure underlying the sparse representation $\mathbf{A} = [\alpha_1, .., \alpha_t]$ where all the coefficients tend to collaborate in time to activate a higher level group corresponding to w_l. This collaborative hierarchical structure (depicted in Figure 1) is leveraged using the C-HiLasso algorithm [2] to obtain the sparse representation α_t as

$$\alpha_t = \min_{\alpha} \frac{1}{2}\|\mathbf{Z} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda_2\psi_{\mathcal{G}}(\mathbf{A}) + \lambda_1 \sum_{t=1}^{T}\|\alpha\|_1 \tag{5}$$

where $\psi_{\mathcal{G}}$ is group Lasso regularizer defined as $\psi_{\mathcal{G}}(\alpha_t) = \sum_{G\in\mathcal{G}}\|\alpha_{[G]}\|_2$. The posterior probability $p(w_l|x_t)$ for a word w_l is estimated as $p(w_l|x_t) := \|\alpha_t^{w_l}\|_1/\|\alpha_t\|_1$.

For speech recognition, frame-level word-posterior probabilities $p(w_l|x_t)$'s are used to obtain the maximum-a-posteriori word recognition through

$$w_{\text{recognized}} := \arg\max_{w_l} p(w_l|\mathbf{X}) = \arg\max_{w_l} \prod_{t=1}^{T} p(w_l|x_t) \tag{6}$$

where $\mathbf{X} = [x_1 \ldots x_T]$. The potential of the proposed approach is demonstrated on isolated word (Phonebook corpus [3]) and continuous speech (Numbers corpus [4]) recognition tasks. The results are listed in Table I. For dictionary learning and sparse decoding, online dictionary algorithm [5] and lasso solver [6] were used respectively.

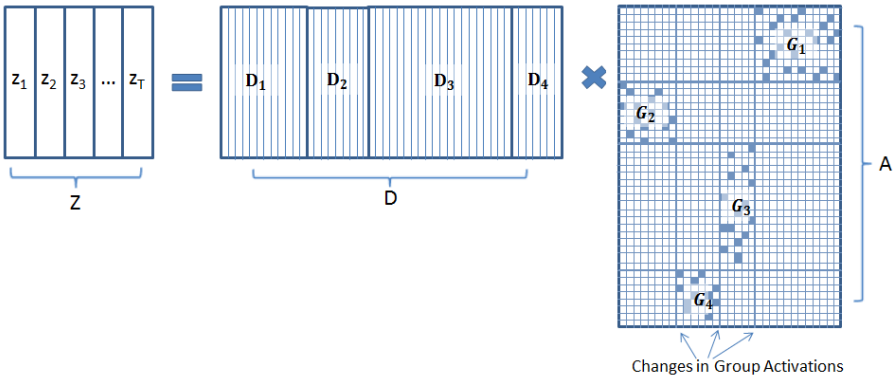

Fig. 1: Given a sequence of acoustic features in $\mathbf{Z}$, the sparse representation matrix $\mathbf{A}$ will have a block structure associated to the word-specific dictionaries where the inner block coefficients are sparse. This collaborative hierarchical sparsity structure is exploited in [2] to devise an efficient C-HiLasso algorithm for sparse recovery.

#	Task	Accuracy
1	Isolated Word (Phonebook-75 words)	97.8%
2	Isolated Word (Phonebook-600 words)	93.2%
3	Connected Digit (Numbers)	87.5%

TABLE I: Results for Isolated Word and Connected Digit Recognition using sparse modeling. Accuracies in case of Connected Digit are given by (100 - WER), where WER is word error rate obtained by Levenshtein distance. The conventional spectral exemplars yield less than 50% accuracy in Isolated Word recognition and around 70% accuracy in Connected Digit recognition tasks.

ACKNOWLEDGMENT

The research leading to these results has received funding from by SNSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" grant agreement number 200021-153507. The authors would like to acknowledge Dr. David Imseng for his assistance with speech recognition experiments.

REFERENCES

[1] G. Aradilla, H. Bourlard *et al.*, "Posterior features applied to speech recognition tasks with user-defined vocabulary," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3809–3812.

[2] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4183–4198, 2011.

[3] J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, "Phonebook: a phonetically-rich isolated-word telephone-speech database," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, May 1995, pp. 101–104 vol.1.

[4] R. A. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at cslu," 1995.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.

[6] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.