A FAST COMPRESSIVE SENSING APPROACH FOR PHONEME CLASSIFICATION

Armin Saeb[1], *Farbod Razzazi*[2]

[1]Electrical Engineering Department, Islamic Azad University, Shahr-e-Rey Branch, Tehran, Iran.
[2]Electrical and Computer Engineering Department, Islamic Azad University, Science and Research Branch, Tehran, Iran.

ABSTRACT

In this paper, a new fast compressive sensing (CS) algorithm for phoneme classification is introduced. In this approach, unlike common CS classification approaches that use CS as a classifier, we use CS as an N-best class selector to limit the secondary classifier input into certain classes. In addition, we use a tree search strategy to select most similar training set for the specific test sample. This makes the system adapted to each test utterance and reduces the empirical risk. By this approach, we obtain promising results comparing with other well known classifiers. In addition, the employed CS approach is a fast l^0 norm algorithm which dramatically reduced the computational complexity in the recognition phase.

Index Terms—Compressive sensing, Phoneme classification

1. INTRODUCTION

Phoneme classification is the procedure of labeling isolated segments of speech by the most likely phonetic labels. Nowadays, phoneme classification plays a key role in most automatic speech recognition (ASR) algorithms and constructs the core of most ASR algorithms.

In most phoneme classification algorithms, it is assumed that utterance ensembles that are used for training, describes the test unseen data well. Therefore, model parameters are determined by training samples and are employed to classify the test samples. As a result, the model parameters are not adapted to test examples and the empirical risk of classification is high. In addition, in noisy environments, the model is usually trained with both noisy and noiseless data to make the system robust against different signal to noise ratios conditions. Indeed, it makes the model parameters match to the average of noisy samples, while it is better to adapt the system to the specific noisy test input example. Although some discriminative Exemplar-based classifiers have tried to decrease the empirical risk of classification in speech recognition (e.g. Support Vector Machines (SVM) and Relevance Vector Machines (RVM) classifiers), however as they have not used the test input utterance to adapt the model parameters, their success have been limited in ASR applications.

Compressive sensing (CS) as a technique to represent a signal by small number of basic signals (atoms) [1], has been shown to be successful in many signal processing applications (e.g. face recognition [2], phoneme classification [3], data compression, channel coding and data acquisition applications [4]). In this approach, an $n \times 1$ signal vector $\mathbf{y}$ is represented by a linear combination of basic $n \times 1$ signal vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, so that the $m \times 1$ coefficient vector $\boldsymbol{\lambda}$ in the equation $\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\lambda}$ is sparse where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$ is the $m \times n$ matrix and the linear equations system is underdetermined ($n < m$).

CS seems to be an appropriate approach in phoneme classification. Its reason is the availability of large number (or even unlimited number) of training examples in speech standard databases that increases the chance of similarity between the test samples with a sparse set of training examples. Sainath et al have used this approach for phoneme classification [3] and have extended their method to large vocabulary continuous speech recognition (LVCSR) [5]. They have used ABCS algorithm as sparse representation method that is reported by IBM research group [6]. Gemmeke et al have employed CS for noise robust ASR [7]. LASSO algorithm [8] has been the sparse representation method in their study. Both of mentioned two algorithms are based on l^1 norm minimization which is too complex and time consuming. Therefore, as stated in [5], implementing exemplar based method in LVCSR applications has been reported as a computationally hard approach.

In this paper, we use a fast l^0 norm CS algorithm for phoneme classification. This algorithm has been introduced by Mohimani et al [9] as smoothed l^0 norm CS algorithm (SL0). Although previous approaches in using CS in phoneme classification have employed CS as the classification engine, we show that SL0 CS approach may be regarded as a training set selector for a classic pattern recognition system. This tunes the model to the test sample with a limited computational cost. The evaluation of the idea shows that this method gives good results in a fair

complexity for phoneme classification, outperforming benchmark classifiers.

The rest of the paper is organized as follows. Section 2 formulates CS approach for pattern classification and points SL0 and its properties. Section 3 explains the proposed secondary training set selection approach. The results of evaluation of the idea on a phoneme classification benchmark are presented in section 4. Finally, section 5 concludes the paper and discusses future works.

2. CS APPROACH FOR PATTERN CLASSIFICATION

CS as a classifier tries to find out similar training samples to the test example and assigns the most similar class to the test sample using a distance measure. Suppose that $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_m$ and $\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_p$ are training and test $n \times 1$ signal vectors respectively ($n,p << m$ which is valid in most speech recognition problems). Because of large number of training samples, it is expected that each test sample can be determined by a few training examples. Therefore, by solving the following problem, the classification may be done:

$$Minimize \ \| \boldsymbol{\lambda} \|_0 \quad subject\ to \quad \mathbf{y} = \mathbf{X} \cdot \boldsymbol{\lambda} \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$ is the $m \times n$ matrix. The classification rule is:

$$c^* = arg\ max \ \| \boldsymbol{\delta}_c(\boldsymbol{\lambda}) \|_1 \tag{2}$$

where $\| \boldsymbol{\delta}_c(\boldsymbol{\lambda}) \|_1$ implies l^1 norm of the coefficients of the training samples that belong to class c. Although, other criteria may be used for classification [5], simulations showed that (2) has the best result in our approach.

Unfortunately, as m and n increases, solving (1) is a NP-hard problem. Therefore, some researchers have examined other approaches like using l^1 or l^2 norm instead of l^0 norm in (1). Some successful algorithms are FOCUSS [10], LASSO [8] and ABCS [6]. Although these algorithms are tractable, they are still slow, especially in LVCSR case.

Smoothed l^0 norm CS algorithm (SL0) [9] is an approach that the problem (1) is solved without substitution l^0 norm with l^1 or l^2 norm. Instead, the l^0 norm term $\| \boldsymbol{\lambda} \|_0$ of (1) is substituted by a suitable continuous function of $\boldsymbol{\lambda}$. In this approach, the above equation is substituted by the following equation:

$$Minimize \ F_\sigma(\boldsymbol{\lambda}) \quad subject\ to \quad \mathbf{y} = \mathbf{X} \cdot \boldsymbol{\lambda} \tag{3}$$

where $F_\sigma(\boldsymbol{\lambda})$ is a smooth differentiable function of $\boldsymbol{\lambda}$ and its minimization is both fast and robust to noise [9]. σ is a parameter that controls the smoothness and accuracy of the approximation. For large σ, the function is very smooth and its minimization will not result in local minima, however, it is not accurate enough. In contrast, small σ makes the function accurate and sharp. However, there are many local minima in the cost function. To overcome this deficiency, as it has been proposed in [9], σ has been decreased gradually from large values to small values. SL0 is a fast algorithm that its complexity is $O(m^2)$ and may be reduced to $O(m^{1.376})$ by using MSL0 [11]. The evaluation of this representation has been shown to be comparable (and even better in some problems) comparing to LASSO with $O(mn^2)$ and its extracted algorithms like Relaxed LASSO with $O(mn^3)$ [12] and ABCS with $O(mn^2)$ complexity or reduced complexity ABCS with $O(mn)$ complexity [13]. Therefore, because of reasonable complexity of SL0, It is expected that SL0 would be a good candidate for LVCSR speech recognition applications in future works.

3. PROPOSED PHONEME CLASSIFICATION ALGORITHM

The main idea of the proposed phoneme classification algorithm in this paper is that by solving (3) and applying the classification rule that is shown in (2), either the final classification may be decided or the correct class would be located at top ranked list (the classes with high $\| \boldsymbol{\delta}_c(\boldsymbol{\lambda}) \|_1$ in (2)). Therefore, by using this idea, the number of classes can be reduced and classification may be performed in a few most probable classes. On the other hand, discriminative exemplar-based classifiers like SVM or RVM classifiers usually have better accuracy when the number of classes is low. Especially SVM classifier has originally designed for large margin binary classification and its accuracy is optimum in this case. Also other margin based online learning algorithms like Passive Aggressive (PA) algorithms [14] have been originally presented as online learning binary classifiers. Therefore it seems that CS, instead of using as a classifier, may be used as an N-best class selector to limit the classifier into certain classes. In addition, a tree search strategy may be used to select the test utterance most similar training set to adapt the training data to each test sample. By using this approach, the secondary classifier may be trained by a limited number of training data that are adapted to the current test example. On the other hand, the number of labels in the classification problem will be limited. As a result, test samples can be classified with better accuracy and with an acceptable complexity.

The architecture of the proposed phoneme classifier algorithm is depicted in Fig. 1. First, the training selected set $\mathbf{X}$ in (3) should be constructed. This is performed by choosing m neighbors of the test vector from the training set, in a simple KD-tree search algorithm [3]. The number of train vectors that constructs $\mathbf{X}$ should be limited as the solution of (1) (or equivalently (3)) to be sparse [1]. After

constructing the set **X**, (3) is solved using SL0 algorithm and *N*-best classes are chosen. Finally a discriminative exemplar-based classifier is applied to *N*-best classes to determine the final decision on the label of the test example which was trained by X set. In this paper PA algorithm is used as a large margin discriminative exemplar-based classifier. Using this online learning algorithm as the secondary classifier, makes it flexible in future works for more gradual adaptation of the model to test examples.

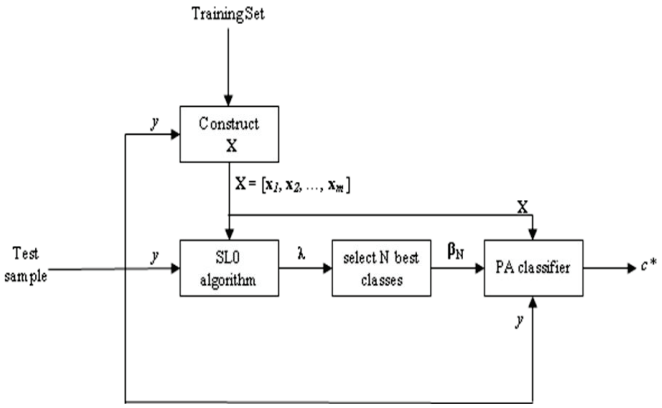

Fig. 1. Block diagram of proposed phoneme classifier

4. EXPERIMENTS

The experiments were conducted on extracted features from TIMIT database. TIMIT contains phonetically balanced 6300 sentences where 10 sentences are uttered by each of 630 speakers from 8 major dialect regions of the United States. In this study, 3096 utterances from standard NIST training set and 100 utterances from standard NIST test set were used as training and evaluation sets respectively. The acoustic model was trained with phonemes with 60 phoneme labels (standard TIMIT phones labels except h#) and was evaluated by smaller set of 39 labels [15]. The segmental features were extracted as in [3]. At first, 13 Mel frequency cepstral coefficients (MFCC) of each frame were extracted and by averaging the MFCC vectors of beginning, middle and ending frames of each phonetic segment and merging these three vectors, a 39 dimension vector was obtained. Then, a 117 dimension vector per each three consecutive segments was generated. Finally, the dimension of this vector was reduced to 40 by Linear Discriminative Analysis (LDA) transform [16]. By this approach, 3096 and 100 training and test utterances were converted to 113349 and 3652 vectors respectively.

The experiments were evaluated based on the architecture of Fig. 1. In this architecture, the KD-tree algorithm [17] was used for matrix **X** construction and *m* was chosen as 200 [3].

In the first experiment, the best classes that are selected by SL0 algorithm [18] were determined. Therefore, the probability that the test sample is located at the N-best class list was investigated. As shown in Fig. 2, the test sample is located at 3 to 5 best classes with probability of 0.9 to 0.95 respectively. Therefore, it seems that only 5 best classes may be selected and be used at the secondary classifier.

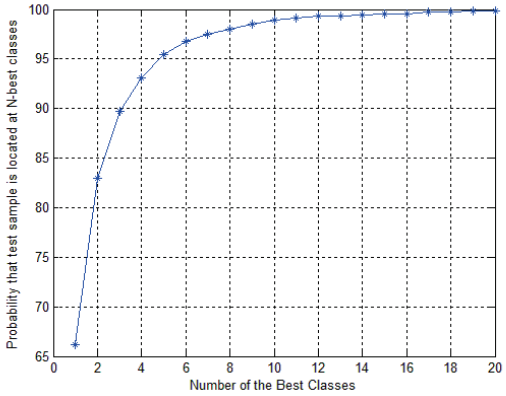

Fig. 2. Probability of being test sample in the N-best list

In the second experiment, the test set accuracy of an SVM classifier [19] by applying the KD-tree selected examples of SL0 selected classes as the train set was evaluated. As shown in Fig. 3, the best accuracy was achieved when the number of selected classes was two. It means that although the correct label of %83 of test samples were located at 2-best classes in compare of nearly %96 in 5 best classes case, however, SVM classifier classified %86 of them correctly and the accuracy of %71.33 was obtained. (in 5-best classes case, only %70 of test samples were classified correctly). Therefore, it is better to use 2-best class candidates for the final classifier training.

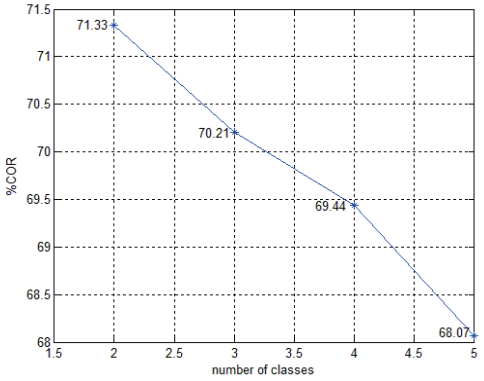

Fig. 3. Percentage of correct classification of proposed phoneme classifier algorithm

In the final experiment, the accuracy of proposed phoneme classifier was compared with some well known

classifiers. In this experiment, the online learning PA classifier [20] was used as the secondary classifier and the result of 2-best class candidates was used for training this final classifier. Results are shown in table 1. As indicated in this table, proposed algorithm outperforms to fast well known classifiers like PA and CS-SL0. Also, SVM classifier has better accuracy in compare with proposed algorithm. But this classifier has high complexity and may not be good candidates for LVCSR problems (or even common ASR). For example, in our experiment, almost 96000 support vectors was extracted while training SVM classifier (almost %85 training samples). This makes classification process complex and time consuming.

Table 1. accuracy for different classifiers on TIMIT

classifier	%accuracy
KNN	68.7
SVM	75.3
PA	68.4
CS-SL0	66.2
proposed classifier	72.2

Finally, we would compare the proposed classifier with one that was introduced by IBM group as the Bayesian CS phoneme classifier [3]. In [3] it was reported the accuracy of %76.44 for phoneme classification on TIMIT with MFCC features. We mention that this reported accuracy cannot be exactly compared with proposed classifier's accuracy in this paper. First, our test condition is a little restrictive than [3]. For example we did not use from silence phonemes (h#) for training and test and used other 60 phonemes for training the algorithm. Second, the most important advantage of the proposed classifier is its high speed and its capability for adapting to test examples that makes it attractive in noisy condition and LVCSR applications.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we introduced a new phoneme classifier based on searching the whole training data set by KD-tree search method and then select N-best classes by CS SL0 algorithm. Then, these reduced training set and class set were used by a well known secondary classifier for phoneme classification on TIMIT corpus. The results showed good accuracy with reasonable complexity that makes this approach attractive for future works on ASR and LVCSR applications.

6. REFERENCES

[1] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289-1306, 2006.

[2] J. Wright, A. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, 2009.

[3] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *proc. ICASSP*, pp. 4370-4373, 2009.

[4] E. J. Candes, M. B. Wakin, "an introduction to compressive sampling," *IEEE Signal Processing Magazine*, pp. 21-30, March 2008.

[5] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exampler-based sparse representation features: from TIMIT to LVCSR," *IEEE Transactions on Audio, Speech and Language Processing* , vol. 19, 2011.

[6] A. Carmi, P. Gurfil, D. Kanevsky, and B. Ramabhadran, "ABCS: Approximate Bayesian Compressed Sensing," *IBM Technical Report, Human Language Technologies*, 2009.

[7] J. F. Gemmeke, H. V. Hammen, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of selected topics in Signal Processing* , vol. 4, no. 2, pp. 272-287, 2010.

[8] R. Tibshirani, "Regression shrinkage and selection via the LASSO," J*ournal of Royal Statistical Society Series B (Methodological)*, vol. 58, no. 1, pp. 267-288, 1996.

[9] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed L0 norm," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 289-301, January 2009.

[10] I. F. Gorodnitsky, B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Transactions on Signal Processing,* vol. 45, no. 3, 1997.

[11] H. Mohimani, M. Babaie-Zadeh, M. Gorodnitsky, and C. Jutten, "Sparse recovery using smoothed L0 norm (SL0): convergence analysis," arXiv:cs.IT/1001.5073, 2010.

[12] N. Meinshausen, "Relaxed LASSO," *Computer and Statistical Data Analysis,* vol. 52, pp.347-393.

[13] T. N. Sainath, B. Ramabhadran , D. Nahamoo, and D. Kanevsky, "Reduced computational complexities of exampler-based sparse representation with applications to large vocabulary speech recognition," in *proc. Interspeech2011*.

[14] K. Crammer, O. Dekel, J. Keshet, S. Shalev-/shwartz, and Y. Singer, "Online passive aggressive algorithms," *Journal of Machine Learning Research,* vol. 7, March 2006.

[15] K. F. Lee, and H. W. Hon, "speaker independent phone recognition using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 37, pp. 1641-1648, 1989.

[16] L. V. D. Maaten, drttoolbox: a Dimensionality Reduction toolbox, http://homepage.tudelft.nl/19j49, 2010.

[17] G. Shechter, KD Tree program, http://guy.shechter.org, 2004.

[18] M. Babaeie-Zadeh, and H. Mohimani, Smoothed L0 (SL0) algorithm for sparse decomposition, http://ee.sharif.ir/~SLzero, 2010.

[19] J. Ma, Y. Zhao, and P. Pavlidis, OSU SVM classifier Matlab toolbox (ver 3.0), http://sourceforge.net/projects/svm, 2002.

[20] F. Orabona, DOGMA: a Matlab toolbox for online learning, http://dogma.sourceforge.net, 2009.