MULTI-VIEW AUDIO-ARTICULATORY FEATURES FOR PHONETIC RECOGNITION ON RTMRI-TIMIT DATABASE

Ioannis K. Douros, Athanasios Katsamanis, Petros Maragos

School of Electrical and Computer Engineering
National Technical University of Athens, Athens 15773, Greece
ioandouros@gmail.com, nkatsam@cs.ntua.gr, maragos@cs.ntua.gr

ABSTRACT

In this paper, we investigate the use of articulatory information, and more specifically real time Magnetic Resonance Imaging (rtMRI) data of the vocal tract, to improve speech recognition performance. For the purpose of our experiments, we use data from the rtMRI-TIMIT database. Firstly, Scale Invariant Feature Transform (SIFT) features are extracted for each video frame. Afterwards, the SIFT descriptors of each frame are transformed to a single histogram per picture, by using the Bag of Visual Words methodology. Since this kind of articulatory information is difficult to acquire in typical speech recognition setups we only consider it to be available in the training phase. Thus, we use a multi-view setup approach by applying Canonical Correlation Analysis (CCA) to visual and audio data. By using the transformation matrix, acquired during the training stage, we transform both train and test audio data to produce MFCC-articulatory features, which form the input for the recognition system. Experimental results demonstrate improvements in phone recognition in comparison with the audio-based baseline.

Index Terms— SIFT features, Canonical Correlation Analysis, Bag of Visual Words, multi-view, rtMRI-TIMIT

1. INTRODUCTION

Speech recognition systems, by harnessing the power of deep neural networks, have achieved significant performance gains in recent years. However, there is still room for improvement, especially when the acoustic conditions are not ideal, as for example, when there is background noise or reverberation. To overcome these problems, various approaches have been proposed, quite a few of which are based on the successful exploitation of another modality, e.g., facial information, that may be available in parallel with audio during speech production. For example, visual features from the face, like Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Active Appearance Model coefficients combined with audio features have been used in audiovisual recognition setups to lower recognition error [1, 2]. There is also great interest around articulatory information, in the form of, e.g., Electromagnetic Articulography (EMA), X-ray Microbeam (XRMB), and real-time MRI data of the vocal tract, and how it could benefit speech technologies [3]. In this direction, we particularly focus on rt-MRI data of speech production and use them to improve speech recognition performance.

Our proposed scheme is based on the multi-view approach. The main idea is about employing different kinds of measurements (views) gathered at the same time and for the same task, with the goal to use one of the views to train effective transformations of the other view. Usually, two views are used but this is not mandatory. For speech recognition, popular views are audio with visual or articulatory features. Another option is to use the labels themselves but in practice this is not very common. In contrast to multi-modal setups, multi-view can handle data with two views, one of which is possibly available only at the training phase. Usually, CCA is used for the transformation to be learned. Such a setup was firstly used in [4] for speaker recognition. Similar setups have been used with success for speech recognition like in [5] which uses the XRMB database. In this paper we adapt this technique to the rtMRI-TIMIT [6] dataset. Although the MRI image quality is not very good, we expect to improve audio-based speech recognition results as the view of the entire vocal tract which is available in this dataset is expected to provide (to some degree) complementary information to the audio stream. The rtMRI-TIMIT database has also been used for phone classification in [7] but the classification in that case is only broad and requires human interaction for placing a masks on each speaker's midsagittal view by manually locating the nose of the speaker at the start of each utterance. To the best of our knowledge, we are unaware of any previous work on the MRI-TIMIT database for phone recognition that requires no human involvement.

In our study, the SIFT features are used for describing each video frame. By applying the Bag of Visual Words technique we transform those descriptors into one histogram per image. We extract MFCCs which are, together with the visual-articulatory histograms, the two views of our experiment. Finally we employ the multi-view setup using CCA. Experimental results demonstrate improvements in phone recognition in comparison with the audio-based baseline.

2. METHODS

The proposed feature extraction scheme essentially comprises two main components: a) the visual feature extraction module, generating a Bag-of-Words representation of SIFT features, and b) the fusion module to properly combine audio and articulatory/visual features. The acoustic features for a single frame are enhanced with their CCA-transformed variant and together they form the speech recognition input at the frame-level, see Fig. 1.

2.1. SIFT features

In our approach, we employ the SIFT features [8] which are robust, scale invariant, and are optimal for matching under different types of invariances [9]. Hence, we expect that head movements or blurring due to poor image quality will not influence recognition results. It has also been discovered that SIFT-like features give the best results on image classification [10]. Moreover, SIFT will catch the movement of different parts of the vocal tract like the tongue. Another interesting choice of features to use, which are also robust and scale invariant, are the Speed Up Robust Features (SURF) features [11]. However good the SURF features may be at object tracking, taking into account the time it takes to be computed, SIFT appear to be a better choice [12]. Even if SIFT are computationally costly in general, this is not really a problem for the rtMRI-TIMIT database which has relatively low resolution.

Generally, to compute the detectors, SIFT algorithm creates different versions of a given image by applying Gaussian filtering with various values of σ. The Difference of Gaussians (DoG) is computed, by subtracting each image from the one with the closest higher value of σ. Every point of the images produced this way, is compared with its neighboring points in a 3×3 grid and with the corresponding ones of the grid on the new images for one values of σ up and down, meaning that each point is compared with $8 + 9 + 9 = 26$ points and labeled as extrema keypoint if its value is minimum or maximum among its neighborhood. Some of them are discarded for better matching results. For the remaining keypoints, orientation θ and gradient magnitude m are computed in the neighboring region and combined to form the final orientation of the keypoint. In Figure 2, the center of each circle correspond to the final keypoints, while the size of the circle and the radius printed represent the value of σ for which the keypoint detected and its orientation respectively (multiple orientations may be assigned to each keypoint).

2.2. Bag of visual words

After the extraction of SIFT we compute the final representation of the visual information which is based on visual bags of words. This method was originally invented and applied for text classification [13, 14]. The main idea is that a document is represented as a histogram of frequencies of the words included in the text. A number of variations have been invented, like term frequencyinverse document frequency (tf-idf) [15] which are based on the same idea and may produce better results. The main difference is that tf-idfs produce weighted histograms [16, 17], while in the classic Bag of Words approach the histograms are binary [18], depending on the presence or the absence of a vocabulary word in the text. Although some studies suggest that using such methods for image classification will improve the results [19], others propose that it does not make much of a difference, as in most cases binary histograms have better results [20]. We choose to follow the original approach and use binary weights for each feature.

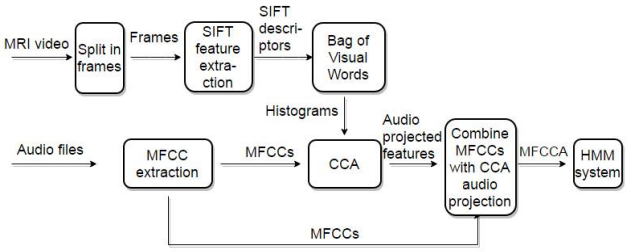

Fig. 1: The multi-view setup for the rt-MRI-TIMIT dataset: Visual feature extraction from the MRI videos and CCA-based combination with the corresponding acoustic features.

2.3. Canonical Correlation Analysis

To combine the visual with the acoustic information, one can use various methods, such as [21,22]. However, there is a major disadvantage in these methods since they require the availability of articulatory information both at the training and at the testing phase. This is not the case for the data we study since it is practically impossible to collect MRI data in a typical speech recognition setup. Our intention is to use the articulatory information only during the training phase and find a relevant transformation for the acoustic features that would also be available in a testing phase. A relatively new approach to achieve this is the multi-view [5], which is a method based on CCA [23].

According to the multi-view approach, CCA is applied to audio and articulatory features at train time in order to find maximum correlated pairs of linear projections of the data in two spaces (views). The idea behind this is that noise in the articulatory domain and noise in the audio domain are highly uncorrelated, therefore by doing such a projection we mostly keep the informative part of the signal without noise. The audio transformation matrix acquired via the CCA procedure is used to transform the audio part of both training and testing sets. To get improved results, we also combine the audio projection with the original audio features which, in our case, are Mel Frequency Cepstral Coefficients (MFCCs) to produce the MFCC-articulatory features called MFCCA [24], for optimal results, since not all uncorrelated information is noise [5].

3. EXPERIMENTS

For the purpose of our experiments we use the data of the MRI-TIMIT database [6]. The database consists of simultaneous audio and MRI recordings of ten speakers, five male and five female, uttering 460 sentences from the TIMIT corpus. For each of them, it includes audio, visual and audio-visual files of the 460 sentences in 92 sets of five. A variety of tools are used to implement the final system including SailAlign, Sox, libsvm, Matlab, vlfeat and Kaldi [25]. The recipe we created is based on the Kaldi s5 recipe for TIMIT.

First, we preprocess the audio files to manually remove any files where the speaker mispronounces a word, which reduces the total number of utterance sets from 920 (92 files $\times$ 10 speakers) to 771. Then we employ SailAlign [26], a tool for robust speech-text alignment, to acquire word as well as phonetic alignments for our dataset. The frequency of the audio files is 20 KHz and their duration is twenty-five seconds approximately. There is also a "beep" sound just before the speaker starts to talk. We used sox to create three additional audio sets: The first one where we remove the beep from the beginning of the files, the second where we downsample the audio files to 16 KHz and the last one where we apply downsampling as well as removing the "beep". We trained two HMM-based and a Deep Neural Network (DNN) system using Kaldi and each of the four audio sets (the original one and the three we created) using 13 MFCCs, their derivatives and accelerations. The best results were derived from the audio with 20 KHz frequency without the "beep" sound. Therefore we continue our experimentation only with this dataset. We stick to the same systems used above throughout our experiment We will provide more details on them in the following. The phone error rate results are summarized in Table 1.

We also cut from each video a small part at the beginning so that each video file has the same length with the corresponding audio file. We use sox again to split the video into frames. The video frequency is about $23.31Hz$ and the produced frames per utterance are around five hundred fifty. At this point we label each frame with its corresponding phone.

To obtain visual information from the frames we use vlfeat to acquire SIFT descriptors. Each frame has about fifty to seventy such descriptors. One may expect more descriptors to be found, however the dimensions of the frames are 68×68 pixels, which explains the low number. Some of the keypoints for different phones are shown in Figure 2.

Table 1: Phone Error Rate (PER) of audio-based speech recognition results.

	$mono$	tri	DNN
original	73%	65%	64.1%
no beep	72%	65%	**63.0%**
downsample	74%	67%	64.5%
combine	72%	65%	63.4%

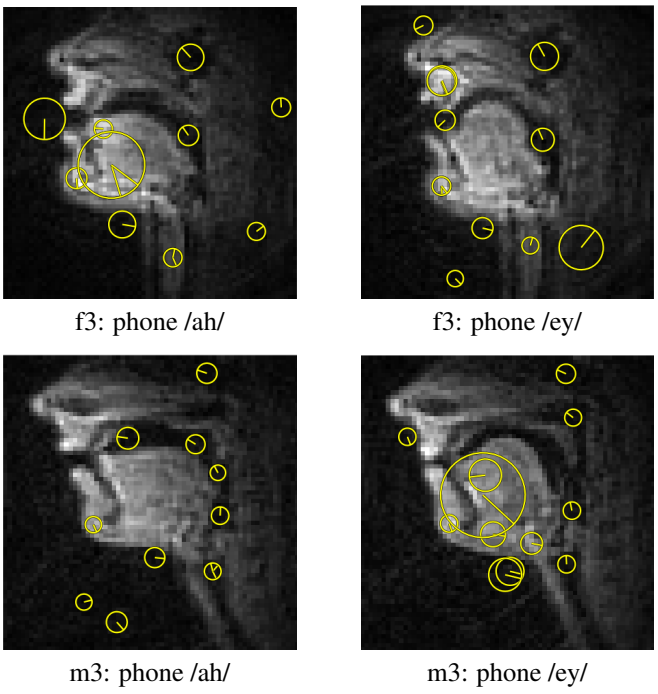

Fig. 2: SIFT detectors. The center of the circles gives the location of each keypoint, the size shows the corresponding value of σ and the printed radius shows the orientation

We then use Bag of Visual Words with the SIFT descriptors as input. The main problem that this method solves is the variability in the total number of SIFT descriptors in every image. Selecting the optimal number k of classes for $k-means$ can be done in a number of ways, e.g., [27]. We estimate the sum of squared errors (SSE) and we pick the optimal k using the elbow method. The SSE shows how concrete a cluster is, meaning how close each descriptor is to the centroid it was assigned. If k equals the number of training point, the SSE will be zero as only one point will be assigned to each centroid and the centroid will have the same value as the training point. Hence we are not looking for the value of k that gives the minimum SSE, but for the value of k above which no significant decrease in SSE value is observed.

To tackle the computational challenges arising due to the large size of all generated SIFT descriptors we proceed with a two-step histogram estimation process. In the first step, we run the $k-means$ algorithm for each speaker separately. We examined the values of k from 30 to 120 with step 15. A good value of k seemed to be between seventy-five and ninety so we chose $k = 85$ for all speakers (Figure 3).

Moreover, we create two additional sets of histograms for every speaker. In the first we used soft normalization across each dimension per utterance by subtracting the mean value and then dividing by two times the standard deviation (plus a small number e in case standard deviation was zero). In the second, apart from the previous normalization we also nor-

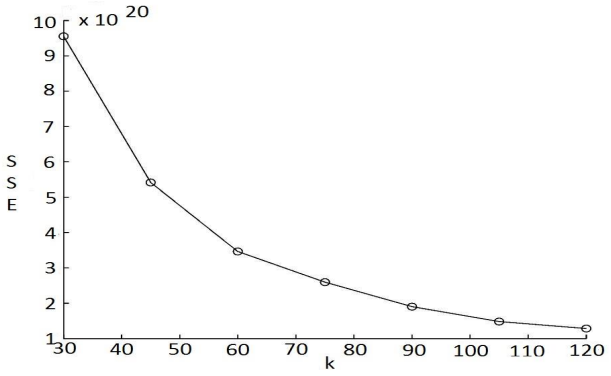

Fig. 3: $k - SSE$ graph for choosing the k for Bag of Visual Words $k-$means clustering of speaker $f3$. For $k > 85$, increasing the value of k does not result in a significant decrease in SSE

malized per sample so that the sum of the squared values of each row equals one. Only for this step we removed phones sil and sp and we trained three SVMs classifiers to determine which type of histogram worked better. To tackle the problem of unbalanced data set, we applied the Synthetic Minority Over-sampling Technique (SMOTE) method [28] in order to create extra samples for minority classes. We used RBF kernel with $\gamma = 0.02$, cost parameter $C = 1$ for every class. 90% of data was used for training and 10% for testing. For the implementation of multi-label SVM training we used the One-Against-All approach using libsvm in Matlab. As expected the best results were achieved by double normalized histograms therefore we continued our experiments only with this type of histograms. The results are shown in Table 2.

In the second step of the histogram estimation phase, we used the centers of the classes of each speaker from the first step and we run again the $k - means$ algorithm only with the centers (10 speakers $\times$ 85 classes $=$ 850 points) this time. Again, we used the elbow method to determine which k to choose. This time k was chosen to be 90.

The idea behind the two-steps histogram creation is that the same phones in the articulatory domain, will correspond approximatively to the same visual words and that the same visual words will be represented in the same region in the $128-$dimensional space of the SIFT descriptors. Therefore we represent every visual word with the center of the class it was assigned to and then we apply $k - means$ only to the centers of every speaker. This idea can be further supported by the fact that the SSE in the second step is significantly lower than the SSE in the first step.

At this point, we compute 13 MFCCs using $25ms$ window with $10ms$ sift. CCA is applied only to the training data between 13-dimensional MFCCs and 85 dimensional double normalized histograms. We acquire the audio transformation matrix with 13×13 dimensions and we transform the audio features of both the training and test sets. We append MFCCs with the CCA projections to create MFCCA features.

The data is divided as follows: 60% for the training set, 20% for the dev set and 20% for the test set. Kaldi toolkit is now used to train the DNN [29]. We briefly remark that out of the three systems trained, the first one is a monophone trained system, the next one is a triphone system which uses MFCCA plus the first and second derivatives as input feature vector, and the last one is a 6-layer DNN system. The results of the phone error rate (PER) can be found in Table 3. We use $5-$fold cross-validation to validate the results.

4. CONCLUSIONS

Our results show that even with the low quality of MRI images the recognition result can be improved if we use articulatory data. This result comes as an additional confirmation that articulatory information can indeed lead to improved speech recognition as it has been shown in other studies in the past [30, 31]. However, our PER is significantly higher compared to the TIMIT corpus due to the fact that the audio and the MRI recordings were made simultaneously therefore they are noisy, as authors in [7] have also argued.

In our case, as we increase the complexity of the system used we notice an improvement with both MFCC and MFCCA features. The best results in every case are coming from the DNN system. The improvement between $monophone$ and DNN system is 11.16% using MFCC and 16.33% using MFCCA. The improvement in the DNN system between MFCC and MFCCA is 5.96% We can see that of the three systems only the DNN system has improved performance comparing to the standard MFCC approach (the improvement of the $monophone$ system is insignificant). A possible reason may be that since we used a complex architecture in order to create speaker independent features, a complex system is required to handle the various correlations.

However more work needs to be done towards the direction of generalizing these results to the entire rt-MRI-database. For example, an interesting idea for future work is the study of other techniques to acquire speaker independent histograms in the Bag of Visual Words step. Finally, one can try to use different kinds of feature descriptors to describe the articulatory information.

Table 2: Choosing the best histogram normalization. Average result of 10 speakers.

	$original$	$normalize_1$	$normalize_2$
Accuracy	5.8%	14.1%	**17.3%**

Table 3: Cross-validated speech recognition results (PER)

	$mono$	tri	DNN
MFCC	72.22%	65.18%	64.16%
MFCCA	72.12%	69.38%	**60.34%**

5. REFERENCES

[1] N.Ahmad, S.Datta, D.Mulvaney, and O.Farooq, "A comparison of visual features for audiovisual automatic speech recognition," *JASA*, 2008.

[2] G.Papandreou, A.Katsamanis, V.Pitsikalis, and P.Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.

[3] S.King, J.Frankel, K.Livescu, E.McDermott, K.Richmond, and M.Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[4] K.Livescu and M.Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 82–86.

[5] S.Bharadwaj, R.Arora, K.Livescu, and M.Hasegawa-Johnson, "Multiview acoustic feature learning using articulatory measurements," in *Intl. Workshop on Stat. Machine Learning for Speech Recognition*. Citeseer, 2012.

[6] S.Narayanan, E.Bresch, P. K.Ghosh, L.Goldstein, A.Katsamanis, Y.Kim, A. C.Lammert, M. I.Proctor, V.Ramanarayanan, Y.Zhu, et al., "A multimodal real-time mri articulatory corpus for speech research.," in *Interspeech*, 2011, pp. 837–840.

[7] A.Prasad and P. K.Ghosh, "Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition," *Computer Speech & Language*, vol. 39, pp. 108–128, 2016.

[8] D. G.Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] K.Mikolajczyk and C.Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[10] K.Mikolajczyk, B.Leibe, and B.Schiele, "Local features for object class recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. IEEE, 2005, vol. 2, pp. 1792–1799.

[11] H.Bay, T.Tuytelaars, and L.Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[12] L.Juan and O.Gwun, "A comparison of sift, pca-sift and surf," *International Journal of Image Processing (IJIP)*, vol. 3, no. 4, pp. 143–152, 2009.

[13] S.Scott and S.Matwin, "Feature engineering for text classification," in *ICML*, 1999, vol. 99, pp. 379–388.

[14] C.Boulis and M.Ostendorf, "Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams," in *Proc. of the International Workshop in Feature Selection in Data Mining*. Citeseer, 2005, pp. 9–16.

[15] J.Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003.

[16] T.Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[17] Y.Yang and X.Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 42–49.

[18] S.Dumais, J.Platt, D.Heckerman, and M.Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*. ACM, 1998, pp. 148–155.

[19] S.O'Hara and B. A.Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *Computing Research Repository*, 2011.

[20] J.Yang, Y.-G.Jiang, A. G.Hauptmann, and C.-W.Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, 2007, pp. 197–206.

[21] A.Wrench and K.Richmond, "Continuous speech recognition using articulatory data," in *ICSLP*. 2000, International Speech Communication Association.

[22] J.Frankel and S.King, "Asr-articulatory speech recognition," in *Proceedings Eurospeech, vol. 1, pp. 599-602*. 2001, International Speech Communication Association.

[23] K. V.Mardia, J. T.Kent, and J. M.Bibby, *Multivariate analysis*, Academic press, 1980.

[24] R.Arora and K.Livescu, "Multi-view cca-based acoustic features for phonetic recognition across speakers and domains," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7135–7139.

[25] 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 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[26] A.Katsamanis, M.Black, P. G.Georgiou, L.Goldstein, and S.Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[27] G.Hamerly and C.Elkan, "Learning the k in k-means," *Advances in neural information processing systems*, vol. 16, pp. 281, 2004.

[28] N. V.Chawla, K. W.Bowyer, L. O.Hall, and W. P.Kegelmeyer, "Smote: Synthetic and minority over-sampling and technique," in *Journal of Artificial Intelligence Research*, 2002.

[29] M.Gales and S.Young, "The application of hidden markov models in speech recognition," *Foundations and trends in signal processing*, vol. 1, no. 3, pp. 195–304, 2008.

[30] K.Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *5th ICSLP*, 1998.

[31] K.Kirchhoff, G. A.Fink, and G.Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, 2002.