Consonant recognition and articulatory feature classification from signal analysis at multiple time scales

Meng SUN, Hugo VAN HAMME
Department of Electrical Engineering-ESAT, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium

ABSTRACT

Consonants exhibit acoustic features that are manifested at different time scales: bursts in plosives live on a millisecond scale, while semivowels are best analyzed at a centisecond or decisecond scale. Yet, it is not easy to integrate analysis at different time scales in the prevailing Hidden Markov Model (HMM). In this paper, we consider the Non-negative Matrix Factorization (NMF) model for consonant recognition and classification of Articulatory Features (AF) on a Vowel-Consonant-Vowel corpus. Different lengths and shifts of the analysis windows and their combinations were tested on these tasks. Classifiers that combine features with different time scales tend to perform better than their components. Consonant recognition rate was improved to 78.0% from the 70.0% with 25ms window and 10ms shift. For plosives, the best performance was 87.7% accuracy with short contextual dependence. The NMF model outperformed a system based on HMMs with Support Vector Machines in the classification of some AF values, but underperformed in other tasks.

Keywords: consonant recognition, articulatory feature classification, non-negative matrix factorization, frame rate, window length

1 INTRODUCTION

Listeners outperform automatic speech recognition (ASR) systems at every level of speech recognition, including the very basic level of consonant recognition. However, what is not clear is where this human advantage originates [1]. The transient parts in speech need to be given more attention because human perceptual mechanisms are more sensitive to those parts [2]. So finer time scales than the conventional analysis window with 25ms length and 10ms shift could be better for the recognition of transient-like sounds in speech. Plosives for instance require an analysis at a finer time scale to separate the closure from the burst phase. However, short analysis windows are suboptimal for other consonants such as nasals, for these are characterized by (anti-)resonances in their spectrum, which are best analyzed with a window of a few pitch periods. If treating each phone class with its most suitable time scale, the recognition framework needs to deal with features computed at multiple time scales, possibly asynchronously.

Variable Frame Rate (VFR) processing has been proposed to emphasize the dynamic aspects of speech [3]. An optimal threshold has to be found in the training and tuning procedure. The proposed model of this paper attempts to integrate analysis at different fixed frame rates instead.

Multiple features with different temporal scales can be integrated by modifying the structure of the conventional Hidden Markov Model (HMM) [4]. In the K-stream recognizer, anchor points are needed between speech units to synchronize the streams. During training it requires prior knowledge about segmentation, which makes the recognition system more complex.

The Non-negative Matrix Factorization (NMF) model proposed in [5] is used for discovering recurring word-sized units in utterances without relying on additional segmentation information. Each utterance is characterized by a representation called Histograms of Acoustic Co-occurrence (HAC), a high-dimensional vector for which the HAC of an utterance is (approximately) the sum of the HAC of individual words. By stacking HACs of different utterances into a data matrix, the HAC of the words can be estimated by factorizing the data matrix. The model can integrate information at different time scales conveniently. The topological structure can be easily modified for different temporal resolution, and their combination.

In this paper, the approach was evaluated on the recognition of smaller units than words: consonants (esp. plosive), by using analysis windows with mul-

tiple time scales. A Vowel-Consonant-Vowel (VCV) corpus was used [7]. Based on the NMF model, we investigated the recognition performance of input representations with different frame lengths and shifts and their combination.

This paper is organized as follows: in section 2, the NMF model on the VCV corpus is described. The results and discussion are in section 3 and section 4. The conclusion is in section 5.

2 NMF Model on VCV Corpus

Non-negative Matrix Factorization decomposes non-negative data into additive parts [6]. In [5] the HAC representation of speech was proposed such that NMF can find word-sized parts from a large amount of utterances. In this paper, we apply this method to recognize and classify consonants in the VCV utterances. Below, we therefore describe the corpus as well as the modifications made w.r.t. the model in [5].

2.1 VCV Corpus and Articulatory Features

The experiments were made on the VCV Corpus [7]. There are 24 consonants, *b,ch,d,dh,dj,f,g,h,k,l,m,n,ng, p,r,s,sh,t,th,v,w,y,z,zh,* in nine vowel contexts consisting of all possible combinations of the three vowels, *aa,iy,uw,* from 24 (male & female) speakers. The training set consists of 6,664 available clean VCV utterances, while the clean test set has 384 available utterances. For more information about the corpus, see [7].

The consonant recognition task is to find out which consonant is in each utterance. Specifically, the recognition of plosives was also made to test the performance of the model with high temporal resolution. The consonant in each utterance was also classified in terms of its articulatory features (AF). The AF considered are *manner*, *place* and *voicing*. The *manner* has six values: *plosive, fricative, affricate, nasal, glide, liquid*, the *place* has six values as well: *labial, alveolar, velar, palatal, dental, glottal*, and the *voicing* has two values: voiced (*+voice*) and unvoiced (*-voice*).

2.2 Grounding Matrix

The NMF model can be used for blind discovery of patterns. However, in this context, the supervisory information is disposed, i.e. one has knowledge about the identity of consonants and AF that are present in a given training utterance. This information will be cast into a matrix G, as outlined below for the example of *consonant* recognition. The NMF models for the consonant/plosive recognition and articulatory feature classification share a similar structure.

For the training set, if the n-th utterance is known to contain the m-th consonant, construct the $M \times N$ grounding matrix G with 1 in its m-th row and n-th column and 0 elsewhere, where M is the total number of different consonants (or plosives, or AF values) and N is the number of training utterances. G' denotes the grounding matrix of the test set. The oracle information described in G helps to extract patterns that are associated with the ground truth [5].

One might suggest to simplify the learning task by adding supervisory information about e.g. the vowels while training the consonant models. From our experiments, there was actually no significant difference with or without these rows of vowel when recognizing consonants.

2.3 Acoustic Feature Representation

Like in G, each column of the acoustic feature matrix A represents an utterance. The elements in each column are the frequencies of the acoustic co-occurrences found in the Vector Quantization (VQ) label sequences [5].

- Codebooks and VQ
  Firstly, the utterance is chopped up in overlapping signal analysis frames. For each frame, MEL-Frequency Cepstral Coefficients (MFCC) are computed to characterize its spectrum. Then codebooks are trained using k-means for each of three feature streams separately. These are the MFCC coefficient vectors themselves (*Static(S)* features), their first order difference (*Velocity(V)*) and their second order difference (*Acceleration(A)*). Then the frames of the training or test set are quantized with the respective codebooks, resulting in a label sequence per stream and per utterance.

- Co-occurrences Statistics
  The label sequence of a stream of an utterance is now transformed into a label co-occurrence sequence by considering label pairs separated by a fixed number of frames, called *lag*. A label co-occurrence is *flattened* to a number between 1 and $|S|^2$, where $|S|$ is the codebook size of the stream. The result is a histogram with $|S|^2$ bins containing the number of times a label pair occurred in the utterance for that stream. Repeating the process for all the streams and with multiple *lag*s and also repeating the process with multiple analysis window sizes and frame shifts, one can stack all the histograms in one column of the acoustic

feature matrix A. Each utterance results in one column of A.

The integration of information at different time scales in the NMF approach is done by simply stacking the histograms. In HMMs, different time scales often require a different topology, which is more complex and it is fairly complex to combine information form multiple, possibly asynchronous, streams in one HMM model [4].

2.4 Training Method

The task is learning W_g and W_a by computing the NMF.

$$\begin{bmatrix} G \\ A \end{bmatrix} = \begin{bmatrix} W_g \\ W_a \end{bmatrix} H \tag{1}$$

We minimize Kullback-Leibler divergence to find the optimal decomposition,

$$D(V||WH) = \sum_{i,j} \left(V_{ij} log \frac{V_{ij}}{[WH]_{ij}} - V_{ij} + [WH]_{ij} \right) \tag{2}$$

where, $V = \begin{bmatrix} G \\ A \end{bmatrix}, W = \begin{bmatrix} W_g \\ W_a \end{bmatrix}$.
The updating formula is therefore [6],

$$H_{kl} \leftarrow H_{kl} \frac{\sum_i \frac{W_{ik} V_{il}}{[WH]_{il}}}{\sum_j W_{jk}}, \quad W_{ik} \leftarrow W_{ik} \frac{\sum_l \frac{H_{kl} V_{il}}{[WH]_{il}}}{\sum_m H_{km}} \tag{3}$$

2.5 Recognition Method

In the recognition process, we compute the activation of each learned part by estimating H' using,

$$A' = W_a H' \tag{4}$$

where A' is the acoustic co-occurrence matrix of the utterances from the test set. Only H' needs to be updated,

$$H'_{kl} \leftarrow H'_{kl} \frac{\frac{\sum_i W_{a_{ik}} V'_{a_{il}}}{[W_a H']_{il}}}{\sum_j W_{a_{jk}}} \tag{5}$$

Then we can get the activation of the consonants for the utterances.

$$Q = W_g H' \tag{6}$$

By selecting the largest activation in each column of Q, the recognition results are obtained. We denote them in a logical matrix P, containing one activation in each column. The evaluation is made by comparing P with the grounding matrix G'. The error rate(or substitution rate) is,

$$\frac{\sum_{i,j} (G'_{ij} - P_{ij})_+}{\sum_j G'_{ij}} \tag{7}$$

where $x_+ = x$ if $x > 0$, and $x_+ = 0$ otherwise.

3 Consonants Recognition and AF Classification Results

Three different values were used for the signal analysis window length (10ms, 20ms and 25ms) and three different values for the frame shift (2ms, 5ms and 10ms). The experiment indexes and the parameters are listed in Table 1, where *S,M,L,** denotes *Short,Medium,Long,all* respectively. So *SM* means *Short* window length (10ms) and *Medium* frame shift (5ms). The **S* means the combination of *SS,MS,LS*.

To make the results comparable, we used the same codebook for the experiments with the same window length. Hence, the co-occurrences at 10ms shift were a subset of that at 2ms shift. The *lags* were $[10, 25, 45]$ for 2ms shift, $[4, 10, 18]$ for 5ms shift and $[2, 5, 9]$ for 10ms shift, so the actual elapsed time was the same in all cases: 20ms, 50ms and 90ms. Because plosives are short lived, using only the shorter time spacing of 20ms and 50ms (2 lags) was also tested in Table 2. The codebook size for the *Static ,Velocity and Acceleration* is 150, 100 and 50 respectively. The feature was extracted with 12-dimensional MFCC plus 1-dimensional energy computed from 30 MEL-filter banks. The factorization dimension in Equation 1 was 45 for consonants, plosives and vowels, 10 for *manner*, 10 for *place*, and 3 for *voice*. The initial values of W was generated randomly from the uniform distribution $U[0.02, 1]$. The upper part of H was initialized by G, while the remainder was from $U[0.1, 1]$. The number of the NMF iteration is 100 for training and 30 for testing.

The diagonal property of W_g obtained in the NMF factorization shows that the NMF model has *found* the expected parts, that is every column only belongs to one consonant or AF value. The results of consonant recognition and AF classification on the testset are in Table 2 where each value is the mean of 5 NMF attempts using different random initializations.

Due to space limitation, for the AF value classification, only the results with 5ms shift and the combined features are shown in Table 3, where the performance of HMM+SVM is from [1].

A 9-fold cross-validation was made to check whether the performance of different features is consistent with that of the original training and testing set. We took the original partition as *Fold 1*, then divided the training set into 8 subsets with disjoint speakers. Each subset contains a male and a female speaker. In folds 2 through 9, the test set of fold 1 plus 7 subsets were used for training and the remaining subset was used for testing. The mean values and the error bars on the 9 Folds are in Figure 1 and Figure 2. The uncertainties in the two figures and in Table 5 are the

Table 1: *Experiment Index*

Index	SS	MS	LS	SM	MM	LM	SL	ML	LL
window	10ms	20ms	25ms	10ms	20ms	25ms	10ms	20ms	25ms
shift	2ms	2ms	2ms	5ms	5ms	5ms	10ms	10ms	10ms
lags	[10,25,45]			[4,10,18]			[2,5,9]		

estimations of the standard deviations of the mean accuracies on the 9 Folds.

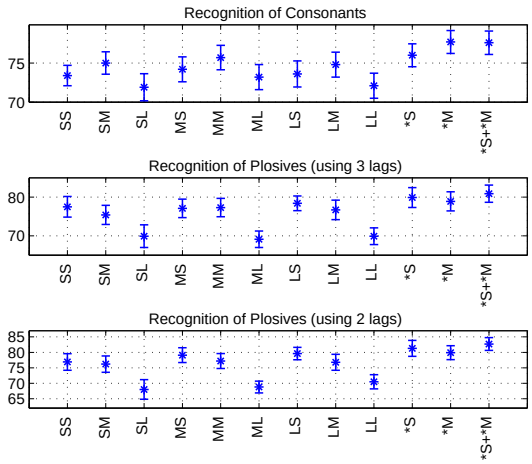

Figure 1: *Error bars of recognition accuracy of consonants and plosives. The mean and uncertainty values were computed on the 9 Folds.*

4 Discussion

The best performance on consonant recognition of the proposed model is 78%, while the human performance is 93.8% for native listeners [8]. The performance on the three AFs is, *voice*>*manner*>*place* (see Table 2 and Figure 2), which is the same as in [1]. But our results are 5% worse for *manner*, 10% worse for *place* and 6% worse for *voice* than those in [1], which is probably because we are not exploiting the segmentation information separating the consonant part from the vowel as in [1]. It is an advantage of the NMF model to achieve such recognition rates without prior segmentation information. Another reason for our lower performance could be the fact that the modeling framework requires VQ, which leads to loss of accuracy when compressing the 13 dimensional MFCC vector into a (1 dimensional) label. However, for some AF values, *affricate, nasal, glide, liquid, velar, palatal, dental, glottal*, the model's performance is better than that in [1]. Since the performance figures of [1] are only available on fold 1, the statistical significance of

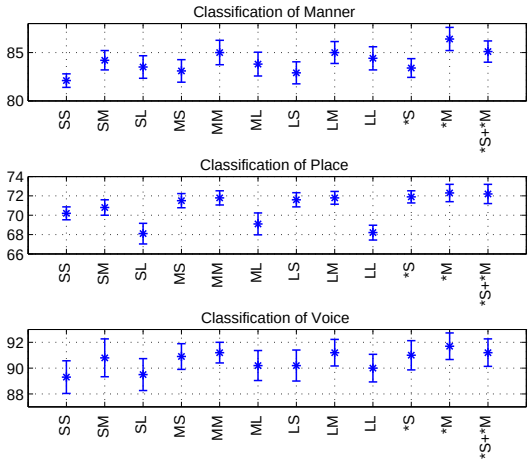

Figure 2: *Error bars of classification accuracy of the three articulatory features. The mean and uncertainty values were computed on the 9 Folds.*

such a comparison is limited. Some further comments are given below.

4.1 Frame Shift

As can be seen from the tables, 10ms shift is always worse than 2ms and 5ms shift. 5ms shift seems to make a good compromise. There may be some *overlearning* at 2ms and some *insufficient training* at 10ms. For a given frame shift at recognition time, one could argue that training with shorter frame shifts should yield better results, since it generates more training data: the co-occurrence data at the large shift plus the interleaved data. From Table 4, for *fold1*, we can see matched shifts are in fact better. The "matched" here means that the training and testing used the same frame shift. Hence, the NMF model also captures the relations between the co-occurrence events, which is broken in unmatched conditions. Another interesting observation is that the top right results in the table are worse than the bottom left ones, i.e. making the test data more sparse (by decimation) is worse than making the training data more sparse.

4.2 Window Length

The best results on consonant recognition are obtained with a 10ms or 20ms window in Table 2 and Figure 1. We expected to get the best performance on plosive recognition using short window lengths, but it is not always true as the medium window length also performs well with short and medium frame shifts in Table 2 and Figure 1. Shorter lags do work better as

expected. Notice also that the comparison is a difficult one since different window lengths also lead to different codebooks. Thus we suggest to use 2ms or 5ms frame shifts in combination with the optimal windows (10ms,20ms).

4.3 Combining information

From Table 2, Figure 1 and Figure 2, it was observed that combining all window sizes improves the accuracy over single time scale analysis. Moreover, combining short and medium frame shifts does not hurt accuracy. We hence advise to use all available information sources.

4.4 Does the gain of combinatorial features come from *multi-codebooks*?

One thing one should notice is that by using the **S *M* and **S+*M* features we actually used multi-codebooks with different time scales implicitly because at each time scale the model uses one codebook. The technique of multi-codebooks is making an acoustic feature matrix A_q as in Section 2.3 by each codebook and concatenating them as the new acoustic feature matrix $A = [A_1; \ldots; A_Q]$. Even with the same MFCCs, the codebooks can be different due to the training process of k-means. In our previous work, we found that multi-codebooks partially overcome the accuracy loss in VQ when making HAC. It can then improve the performance of the model.

Experiments were made on the 9 Folds to check if the improvement of combinatorial features comes from *multiple time scales* or *multi-codebooks*. The results are in Table 5 where $3 \times SS$ means 3 codebooks with *Short* window length and *Short* frame shift. So the acoustic feature matrix of 3 (6) codebooks has the same size as the ones of **S* and **M* (**S+*M*). There are indeed improvements from one codebook to 3 or 6 codebooks for all the features, but not as large as by using features with multiple time scales. Thus we conclude that the gain comes from the combined information at different time scales.

5 Conclusion

Using the NMF model, different window lengths and shifts were compared on the recognition of consonants and plosives and the classification of articulatory features on the VCV corpus. There was still a lot of room for improvement of the recognition results. The AF classification results were not as good as what is presented in [1], which can be attributed to the vector quantization process in our approach or to the fact that we do not rely on segmentation of the consonant parts in the utterances. But the performance on some AF values outperformed that of the HMM+SVM model. The 10ms frame shift that has settled as a compromise for the HMM recognizer [9] is better replaced by 5ms in the NMF method. The multiple window length and frame shift idea seemed to work indeed: its performance is always better than its components because it incorporates the complementary features of the different analysis. Furthermore, the NMF model is capable of exploiting asynchronous multi-scale information. For plosives, the best performance is achieved with short contextual dependence.

While this analysis focused on the representation of speech, we should be aware that a lot of parameters can affect the actual recognition results, e.g. codebook sizes, parameters in MFCC extraction, *lags*, factorization dimensions and the NMF initial values. We controlled these parameters in the experiments as ex-

Table 2: *Accuracy of Consonant Recognition and AF classification on Fold 1*

Features	Consonant	Plosive (3 lags)	Plosive (2 lags)	Manner	Place	Voice
SS	72.3	85.0	83.5	83.1	69.0	87.9
SM	77.7	80.4	83.5	86.6	72.0	**90.1**
SL	75.0	81.9	81.0	85.8	71.6	89.4
MS	76.5	82.5	83.3	82.9	71.4	88.8
MM	76.3	84.6	85.2	86.5	**72.8**	90.0
ML	73.1	73.3	69.6	84.4	71.4	89.3
LS	75.7	82.7	84.0	83.7	72.5	88.8
LM	75.1	80.0	79.8	85.8	71.6	89.4
LL	70.0	69.2	74.8	84.4	69.4	88.7
*S	76.0	**85.6**	86.9	82.6	71.5	89.6
*M	**78.0**	84.4	84.8	**86.7**	72.1	89.9
*S+*M	77.7	85.0	**87.7**	85.5	72.1	90.0
HMM+SVM [1]	-	-	-	91.7	82.1	95.8
Human [8]	93.8	-	-	-	-	-

Table 3: *Average Accuracy of AF Value Classification on Testset of Fold1*

AF value	Accuracy(%)							# Test Utt.
	SS	MM	LS	*S	*M	*S+*M	HMM +SVM	
manner								384
plosive	81.9	84.0	83.5	82.3	88.5	**91.7**	**96.9**	96
fricative	83.8	90.4	80.6	77.8	**94.4**	77.8	**96.5**	144
affricate	86.3	88.8	**91.3**	90.6	90.6	90.6	87.5	32
nasal	83.8	75.0	83.3	83.3	**87.5**	85.4	85.4	48
glide	81.3	85.6	**92.5**	93.8	78.1	84.4	87.5	32
liquid	77.5	**90.0**	76.3	78.1	81.3	78.1	81.3	32
place								384
labial	71.3	**83.8**	77.9	68.8	75.0	80.2	**94.8**	96
alveolar	47.3	51.3	59.4	**76.0**	45.8	56.3	**83.3**	96
velar	78.3	72.9	80.0	68.8	79.2	**85.4**	81.3	48
palatal	81.9	86.5	84.8	76.0	**94.8**	84.4	90.6	96
dental	**56.9**	45.6	46.3	56.3	53.1	53.1	53.1	32
glottal	**91.3**	76.3	83.8	87.5	81.3	87.5	68.8	16
voice								384
+voice	87.7	**93.5**	91.3	86.8	86.1	86.8	**96.7**	144
-voice	87.5	85.8	86.0	89.6	**92.5**	90.4	**95.8**	240

Table 4: *Training and Recognition With Different Shift on Testset of Fold1*

Train \ Test		MS		MM		ML	
		Max.	Mean	Max.	Mean	Max.	Mean
MS	manner	84.4	83.0	73.2	70.7	45.3	43.0
	place	72.9	71.7	68.2	66.2	56.5	54.5
	voice	89.8	88.7	89.8	84.2	79.4	69.3
MM	manner	83.1	80.3	88.0	86.3	77.3	75.2
	place	73.7	68.9	74.7	71.5	66.9	64.6
	voice	88.3	87.8	91.9	90.6	90.1	88.8
ML	manner	77.6	77.0	83.6	82.2	85.2	84.1
	place	71.1	70.0	74.2	72.4	74.0	71.9
	voice	90.1	86.7	90.6	89.5	90.1	89.3

Table 5: *Consonant Recognition rates (%) on the 9 Folds using multi-codebooks*

	1×SS	1×MS	1×LS	1×SM	1×MM	1×LM	1×ML	1×LL
Mean	73.4	74.2	73.6	75.0	75.7	74.8	73.2	72.1
Unc.	1.3	1.6	1.7	1.1	1.6	1.6	1.6	1.6
	3×SS	3×MS	3×LS	3×SM	3×MM	3×LM	*S	*M
Mean	74.6	74.7	74.4	75.8	76.2	75.6	76.0	77.7
Unc.	1.4	1.4	1.6	1.3	1.7	1.6	1.1	1.1
	6×SS	6×MS	6×LS	6×SM	6×MM	6×LM	*S+*M	
Mean	74.3	74.3	74.1	75.5	75.8	75.6	77.6	
Unc.	1.3	1.5	1.7	1.2	1.7	1.6	1.5	

ploring the parameters would require more data.

6 Acknowledgment

This research was funded by the European Commission under contract FP6-034362 (ACORNS) and K.U.Leuven research grant OT/09/028 (VASI).

7 REFERENCES

[1] Scharenborg, O.E., Cooke, M.P.,"Comparing Human and Machine Recognition Performance on a VCV Corpus", **Proceedings of the workshop on Speech Analysis and Processing for Knowledge Discovery**, Aalborg, Denmark, 2008, pp.CD-ROM.

[2] Hant, J.,Alwan, A., "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise", **Speech Communication**, vol.40(3), pp.291-313, 2001.

[3] Bridle, J.S., Brown, M.D., " A date-adaptive frame rate technique and its use in automatic speech recognition", **Proceedings of Institute of Acoustics Autumn Conference**, pp.C2.1-C2.6, 1982.

[4] Bourlard, H., Dupont, S., Ris, C., "Multi-Stream Speech Recognition", **CC AI The Journal for the Integrated Study of Artificial Intelligence, Cognitive Science and Applied Epistemiology**, vol.15(3), pp.215-234, 1998.

[5] Van hamme, H.,"Integration of asynchronous knowledge sources in a novel speech recognition framework", **Proceedings of the workshop on Speech Analysis and Processing for Knowledge Discovery**, Aalborg, Denmark, paper 038, 2008.

[6] Lee, D.D., Seung, H.S., "Learning the parts of objects by non-negative matrix factorization", **Nature** 401, pp.788-791, 1999.

[7] http://www.odettes.dds.nl/challenge_IS08/

[8] Cooke, M.P., Scharenborg, O.E., "The Interspeech 2008 Consonant Challenge," **Proceedings of Interspeech**, Brisbane, Austrilia, 2008.

[9] K.F.Lee, H.W.Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", **International Conference on Acoustics, Speech, and Signal Processing**, Vol.1, pp.123-126, April, 1988.