

Various NMF Analyses for Emotion Recognition

JeongMok Ha, Jaiyoun Song, Hong Jeong

Department of Electronic & Electrical Engineering, POSTECH (Pohang University of Science and Technology), Korea
 jmokha@postech.ac.kr, tichiel@postech.ac.kr, hjeong@postech.ac.kr

Abstract—To extract emotion from speech signals, we must specify the representation and grammatical model, which are still challenging issues. We proposed a new feature called Non-negative Matrix Factorization (NMF) feature. The proposed algorithm has been tested in several different ways by varying NMF and the speech database. We compared the recognition rate only Euclidian distance with enhancing ways (SVM, Partial multiplication of vectors, SFM) of the NMF classification. Observing all these together, we found that total recognition rate is improved. We concluded that the NMF feature indicates both spectral information and temporal information, which is an efficient tool over other spectrum-based features.

Keywords— Emotion recognition, Non-negative matrix factorization, Feature extraction, Spectral analysis

I. INTRODUCTION

Human emotion is important because it influences and interacts with mental processes and behaviour processes. One method for emotion extraction is the use of speech signals. Unfortunately, assessing emotion and analysing speech signals, is a very difficult task. Regardless, there have been continuing efforts to develop better algorithms.

Unlike speech recognition, emotion is more or less independent from the temporal variations. Instead, the prominent characteristics are more related with frequency distributions and timbre, as we will observe in the next section. In this light, we suggest an algorithm based on NMF feature for spectral information.

This paper is presented as follows. In Section II, we define the problems in terms of estimation. In Section III, we suggest an NMF method to extract features. The classification is also mentioned here. In addition, in Section IV, we adapt the NMF method for each frames consisting of a sentence to extract temporal features. Various methods using NMF are suggested in Section V. In Section VI, we explained the database and test method and showed the results.

II. PROBLEM FORMULATION

We define variables and parameters. We define the speech signals in a frame n as the Markov chain $\mathbf{X}(n) = \{X_1(n), X_2(n), \dots, X_T(n)\}$, where n denotes the frame number and T denotes the number of frames. We assume that the emotion process $\mathbf{E}(n)$ is quasi-stationary and thus constant within a given frame. Then, the dependency relationship can be represented as in Fig. 1.

We assume the model for speech and emotion processes.



Figure 1. Dependency between speech signals and emotion processes.

- 1) The emotion is quasi-stationary so that $\mathbf{E}(n)$ is constant in the frame n .
- 2) The emotion is defined as a 1st-order Markov process. $\mathbf{E}(n)$ depends upon the emotion $\mathbf{E}(n-1)$ in the previous frame only. That is, $p(\mathbf{E}(n) | \mathbf{E}(n-1)) = p(\mathbf{E}(n) | \mathbf{E}(n-k), k=1, 2, \dots)$. Also, this quantity is considered to be time-invariant.
- 3) The speech signal $\mathbf{X}(n)$ is context-free, so that it depends only upon $\mathbf{E}(n)$. That is, $p(\mathbf{X}(n) | \mathbf{X}(n-k), \mathbf{E}(n-k)) = p(\mathbf{X}(n) | \mathbf{E}(n)), k=1, 2, \dots$

Although a more general model can be considered, this simple model is considered in this paper.

Our task is to estimate $\mathbf{E}$, given $\mathbf{X}$. The simplest possible approach is the ML approach.

$$\mathbf{E}^*(n) = \arg \max_{\mathbf{E}} p(\mathbf{X}(n) | \mathbf{E}(n)). \quad (1)$$

If the prior knowledge about $\mathbf{E}(n)$ is known, we may instead use the MAP:

$$\begin{aligned} \mathbf{E}^* &= \arg \max_{\mathbf{E}} p(\mathbf{E} | \mathbf{X}) \\ &= \arg \max_{\mathbf{E}} \frac{p(\mathbf{X} | \mathbf{E})p(\mathbf{E})}{p(\mathbf{X})} \\ &= \arg \max_{\mathbf{E}} p(\mathbf{X} | \mathbf{E})p(\mathbf{E}). \end{aligned} \quad (2)$$

Hear, $p(\mathbf{E})$ represents the distribution of the emotion.

III. EMOTION AND DISTANCE MODEL

The next step is to specify $p(\mathbf{X} | \mathbf{E})$, $p(\mathbf{E})$, and $p(\mathbf{E}(n) | \mathbf{E}(n-1))$ in (2).

A. Prior distribution

$p(\mathbf{E})$, as already noted, denotes a priori probability about the emotion $\mathbf{E}$. Fig. 2 is a statistical database compiled by Morrison et al. [1]. It shows a typical example for a call center database. The data consisted mainly of neutral speech, with the second largest category being angry speech. This means

that real world application reveals neutral and angry situations better than others situations.

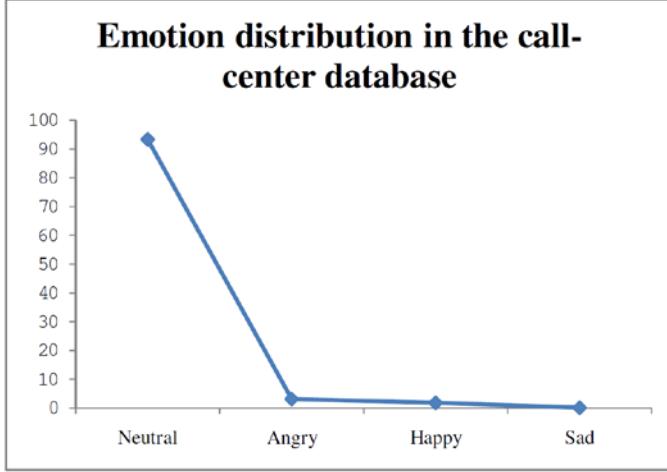


Figure 2. Emotion Distribution in the call-center DB

In this graph the data has been arranged in decreasing order. It is reasonable that neutral emotions occur most of the time. The other emotions occur far less frequently. Their relative distribution may depend upon the type of database such as drama, movie and news. However, it is clear that the neutral distributions are more probable and must be considered in estimations. Let us number the class of emotion $k = 0, 1, \dots, K-1$, representing neutral, angry, happy, and sad in this graph. Then, we can define the likelihood of a priori as:

$$\ln p(E = k) \triangleq c_k, \quad \sum_{k=0}^{K-1} c_k = 1, \quad (3)$$

where the determination of parameter c is dependent upon the speech database. For the shown example, $c_0 = 0.93$, $c_1 = 0.031$, $c_2 = 0.018$, and $c_3 = 0.001$.

This quantity is 1-gram that can be calculated by observing the given corpus already labelled by emotion classes. For the given corpus consisting of N events of emotion:

$$p(k) = p(E = k) = \frac{f(k)}{N}, \quad k \in [0, K-1], \quad (4)$$

where $f(k)$ denotes the frequency of the emotion class k .

B. NMF feature

Next we represent the emotion $E(n)$ with the non-negative matrix factorization (NMF) [2]. For this purpose, we first obtain a spectrogram $X(n)$ which is an $n \times m$ matrix. Since the $n \times m$ matrix X , which is an amplitude spectrogram X , nonnegative, it can be decomposed to two non-negative factors W and H :

$$X \approx WH, \quad (5)$$

where W and H are $n \times r$ and $r \times m$ matrix, respectively for some integer n and r . One reasonable choice of r is the sample size of FFT. This factorization is characterized with the fact that all elements are all nonnegative.

We consider the squared Euclidean distance as a

discrepancy measure between the data X and the model WH , leading to the following least squares error function

$$J = \frac{1}{2} \|X - WH\|^2. \quad (6)$$

NMF involves the following optimization:

$$\arg \min_{A \geq 0, S \geq 0} J = \frac{1}{2} \|X - WH\|^2. \quad (7)$$

Gradient descent learning (which is additive update) can be applied to determine a solution to (7), however, nonnegativity for W and H is not preserved without further operations at iterations.

On the other hand, a multiplicative method developed in [2] provides a simple iterative algorithm to solve (7). We give a slightly different approach from [2] to derive the same multiplicative algorithm. Suppose that the gradient of an error function has a decomposition that is of the form

$$\nabla J = [\nabla J]^+ - [\nabla J]^-, \quad (8)$$

where $[\nabla J]^+ > 0$ and $[\nabla J]^- > 0$. Then multiplicative update for parameters Θ has the form

$$\Theta \leftarrow \Theta \odot \left(\frac{[\nabla J]^-}{[\nabla J]^+} \right)^\eta, \quad (9)$$

where $\odot$ represents Hadamard product (elementwise product) and $(\cdot)^\eta$ denotes the elementwise power and η is a learning rate ($0 < \eta \leq 1$). It can be easily seen that the multiplicative update (9) preserves the nonnegativity of the parameter Θ , while $\nabla J = 0$ when the convergence is achieved.

Derivatives of the error function (6) with respect to W with H fixed and with respect to H with W fixed, are given by

$$\begin{aligned} \nabla_A J &= [\nabla_A J]^+ - [\nabla_A J]^- \\ &= WHH^T - XH^T, \end{aligned} \quad (10)$$

$$\begin{aligned} \nabla_S J &= [\nabla_S J]^+ - [\nabla_S J]^- \\ &= W^T WH - W^T X. \end{aligned} \quad (11)$$

With these gradient calculations, the rule (9) with $\eta = 1$ yields the well-known Lee and Seung's multiplicative updates [2].

$$W \leftarrow W \odot \frac{XH^T}{WHH^T}, \quad (12)$$

$$H \leftarrow H \odot \frac{W^T X}{W^T WH}, \quad (13)$$

where $\odot$ is element-wise product (Hadamard product) and all divisions are element-wise division.

The dimension r must be determined heuristically depending upon the nature of the original matrix X . In our case, other than $r = 1$ it does not give better performance. Fig. 3b shows typical examples with $r = 1$.

Fig. 3a shows the voiced parts in the spectrogram and Fig. 3b is its NMF. One can observe that the distribution in W represents the average energy in each frequency band. This quantity is relative because the signals are normalized by the total sum. On the other hand, H represents temporal variations of the total energy. This quantity is also relatively

normalized. Actually, the spectrogram is decomposed into two orthogonal axes that are frequency and time. Among these axes, the frequency gives the vital information on emotion. We will use W as a feature vector for emotion and call it *NMF feature*.

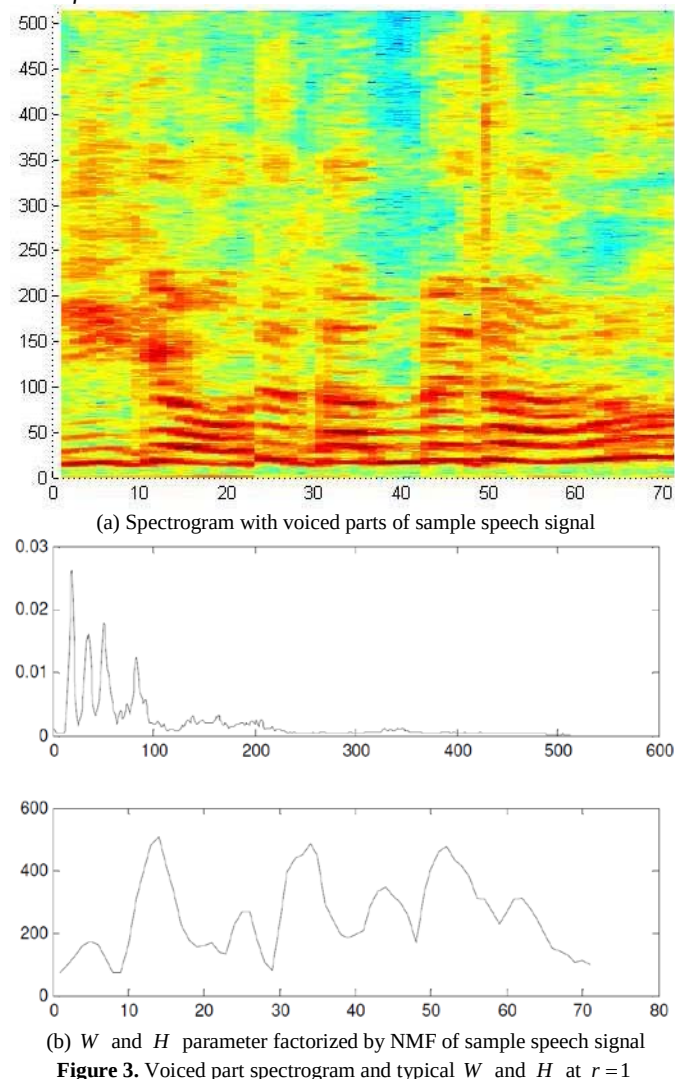


Figure 3. Voiced part spectrogram and typical W and H at $r=1$

Fig. 4 is a typical example of the spectrogram and its voiced parts. Fig. 4a is the original and Fig. 4b is the voiced part only. In order to detect the voiced region, we used the ordinary zero-crossing detection method, which is simple and fast, for both training and recognition phases. This is shown in Fig. 5. The top path is the normal process for the spectrogram. The bottom path is added to find the voiced parts using the normal zero-crossing method. Once the voiced portion is found, only those parts are concatenated to be a new spectrogram, which is again the input to NMF part.

IV. FRAME NMF FOR TEMPORAL FEATURES

From the previous matrix X in (5), we can extract the NMF basis. As it was mentioned before, NMF for whole sentence fails to catch temporal features because W represents only

frequency information. Not only the frequency features, but also the temporal features are important to extract emotion features. For instance, in [3], they extract both statistical features and temporal features for HMM based emotion recognition.

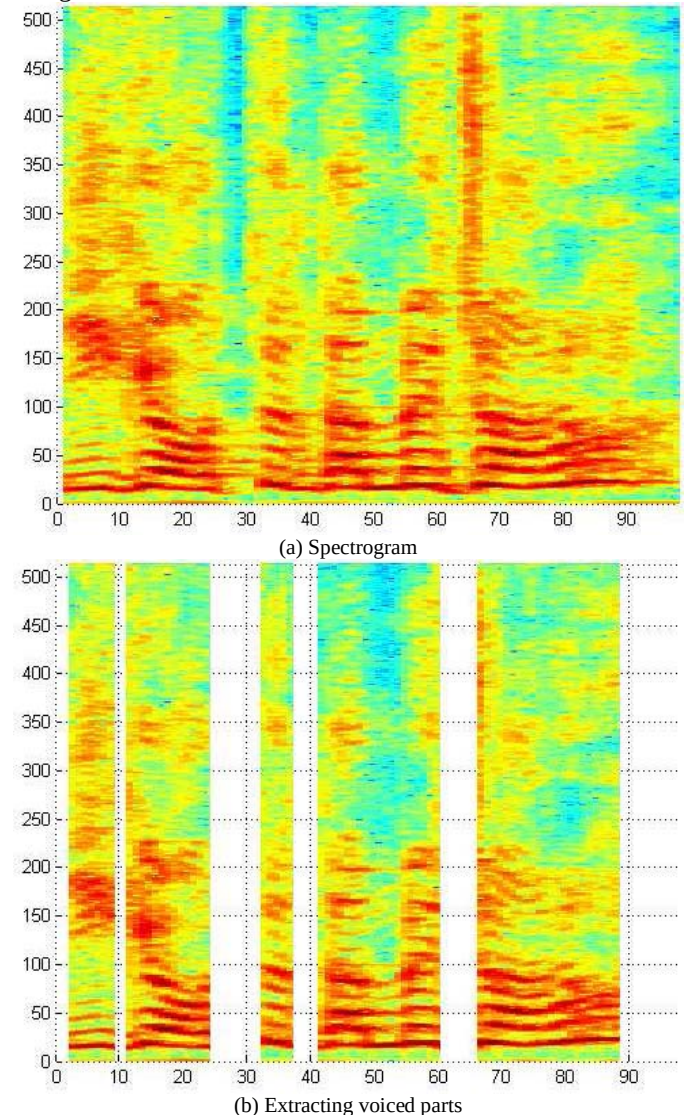


Figure 4. Example of an angry voiced region of a spectrogram

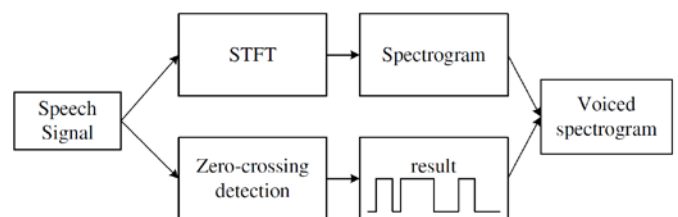


Figure 5. Block diagram of the zero crossing detection processes

In the point of view, we tried to separate the matrix X by t frames. Fig. 6 shows how we separated the spectrogram. After separating the spectrogram X each fragment has $n \times t$ size of the preprocessed spectrogram. Then, apply NMF to each pieces of the spectrogram such as whole sentence NMF, so

that the basis W can be derived. Frame NMF basis which has been obtained from a sentence can be used to find temporal features while H in (5) cannot find any emotion feature.

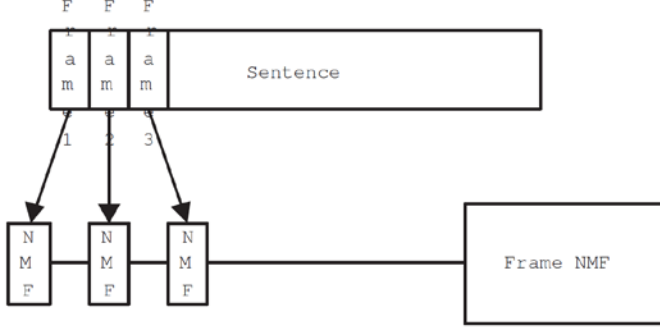


Figure 6. Frame NMF system

V. ENHANCING WAYS OF THE NMF CLASSIFICATION

In this section, we suggest some ways for the better performance of the emotion classification.

A. Support vector machine

The classifying performance is improved by the SVM (Support Vector Machine) [4], which is widely used in pattern recognition. Fig. 7 describes how the SVM works.

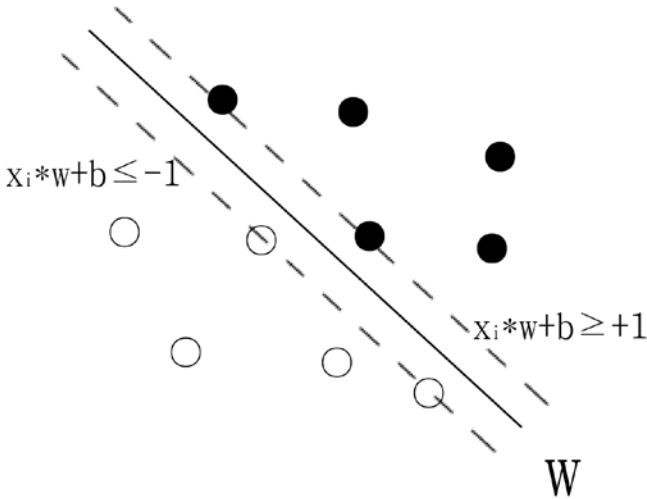


Figure 7. Linear separating hyperplanes for the separable case

The dots x_i are each values. Two classes are divided by a hyperplane w with margin $\frac{2}{\|w\|}$. Suppose that the value $y_i \in \{-1, 1\}$ points each class. To derive the hyperplane and margin, Lagrange multiplier is used satisfying the condition:

$$x_i \cdot w + b \geq +1 \quad \text{for } y_i = +1, \quad (14)$$

$$x_i \cdot w + b \leq -1 \quad \text{for } y_i = -1. \quad (15)$$

These can be combined into one set of inequalities:

$$y_i (x_i \cdot w + b) - 1 \geq 0 \quad \forall i. \quad (16)$$

From the conditions, the Lagrange multiplier can be constructed:

$$L_p \equiv \frac{1}{2} \|w\|^2 - \sum_{i=1}^l \alpha_i y_i \{x_i \cdot w + b\} + \sum_{i=1}^l \alpha_i. \quad (17)$$

We consider x_i as the basis of NMF, w so as to classify w toward each emotions. In normal way, neutral and sad emotions are separated well. Happy and angry emotions are differentiated from the two emotions. However, they are blended up. To solve the problem, we need another approach.

B. Partial multiplication of vectors

As we said earlier, using whole vectors depresses the recognition rate, especially happy and angry. Fig. 8a shows how they are mingling. 36 happy sentences and 42 angry sentences were investigated for the result of the multiplication between the hyperplane w for “angry” and the basis x_i . The equation can be written in (18):

$$\zeta_i = \sum_{k=1}^n x_{ik} \cdot w_k \quad \forall i. \quad (18)$$

Observing the hyperplanes for emotion “angry” and “happy”, we can find the difference. In Fig. 9, the difference between two emotions is seen in the intervals $k = [51, 100]$ and

$k = [351, 513]$. Multiplicating the intervals of the w to the x_i , Fig. 8b is derived. The modified equation can be written in (19):

$$\zeta_i = \sum_{k=\text{intervals}} x_{ik} \cdot w_k \quad \forall i. \quad (19)$$

C. Spectral flatness measure of the frame NMF

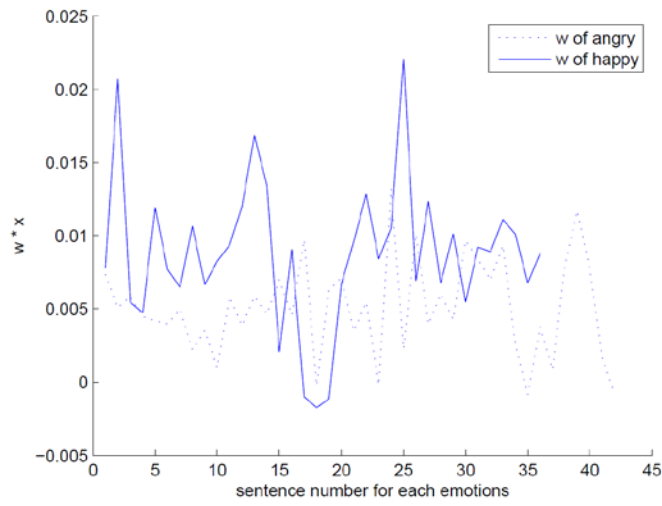
Spectral flatness measure (SFM) is used to investigate the flatness of a signal. In timbre classification [5], it is used to distinguish noise from the original speech since a white noise is relatively flatter than the meaningful speech. SFM can be demonstrated as in (20):

$$SFM = \frac{\sqrt[n]{\prod_{n=0}^{N-1} x(n)}}{\left(\frac{\sum_{n=0}^{N-1} x(n)}{N} \right)}. \quad (20)$$

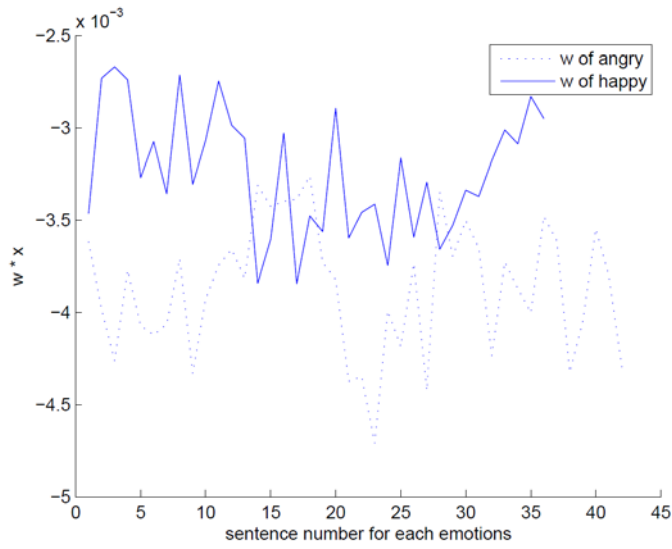
$x(n)$ means the magnitude of bin number n . The geometric mean is located to the denominator and the arithmetic mean is located the numerator. Therefore, a high spectral flatness means that the spectrum has a slight difference of power in total interval.

We analyzed the frame NMF to find temporal features. Frame NMF varies according to the time preserving fundamental formation of sentence NMF basis. Changing of the maximum argument frame NMF can be seen easily. Fig. 10 illustrates the maximum argument frame NMF according to the time.

Fig. 10a is the maximum argument frame NMF of neutral sentences, and Fig. 10b is the maximum argument frame NMF of angry sentences. Maximum argument fluctuation of the angry sentences is more apparent than that of the neutral sentences. The fluctuation can be translated to the SFM. Coarse graph leads lower value of the SFM; otherwise, flat graph leads higher value of the SFM.



(a) Multiplication whole parts of w and x_i



(b) Multiplication some intervals of w and x_i

Figure 8. The difference of $w \cdot x_i$ between original and partial calculation

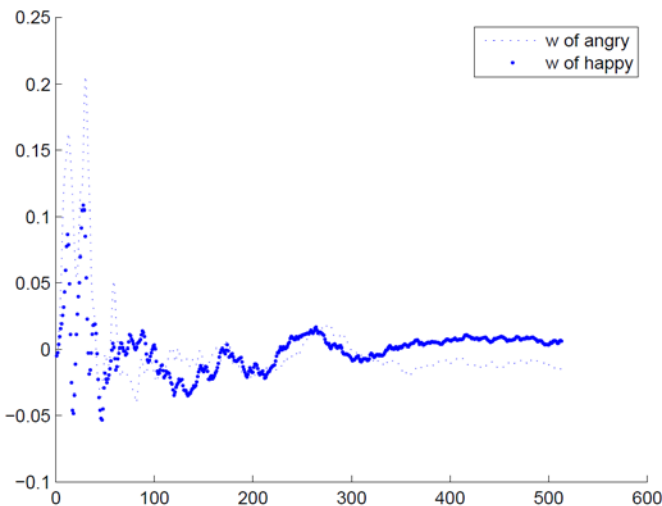
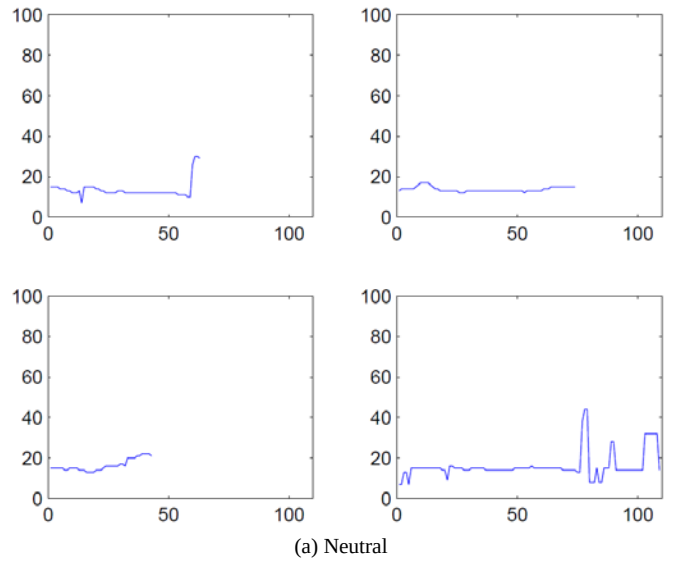
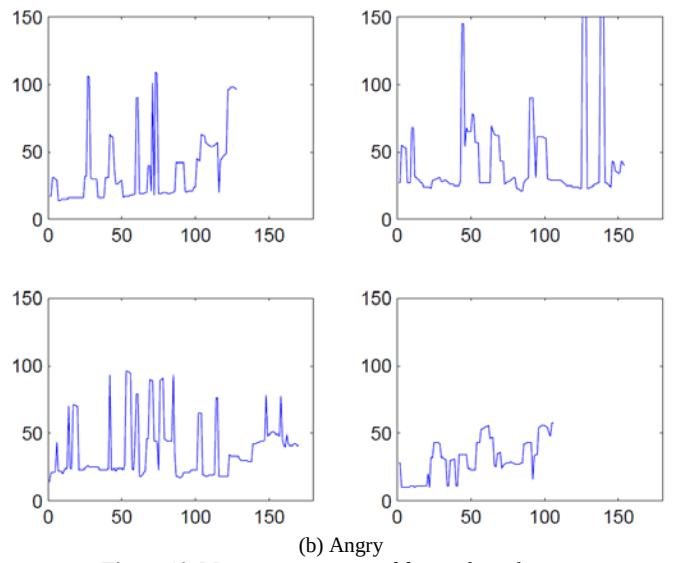


Figure 9. The trained hyperplane w for emotion “angry” and “happy”



(a) Neutral



(b) Angry

Figure 10. Maximum argument of frame along the time

VI. EXPERIMENTAL RESULTS

In this section, we explain the actual realization and the results in detail.

A. Speech data

Generally, there are three methods of data acquisition in emotion research [6]. The first method is natural expression, where data are collected from a real-world situation where users are not under obvious observation and are free to express emotions naturally, as they would in an everyday situation. The second method is induced emotional expression, where naive users are presented with scenarios that induce the required emotional response. The last one is speech acquisition using simulated or portrayed emotional expressions using professional or non-professional actors and actresses.

In this study, data are collected from a drama in public broadcasting. First, utterances were manually segmented by a human at every sentence interval. Second, each emotion is

classified and labeled by a human. Third, the labeled data is once more classified by a machine classifier to obtain cluster centers for each emotion class. The databases contain 250 sentences produced by 20 professional actors and actresses, 10 male and 10 female in four emotions (neutral, angry, happy and sad). The segmentation process results in around 1,000 utterances. Each utterance was stored in 16 bit PCM format sampled at 16 kHz.

B. Spectrogram

In order to achieve feasible estimations, and avoid the difficulties caused by the non-stationary nature of speech, it is assumed that the properties of the signal change relatively slowly with time. This allows examination of a short-time window of speech to extract relevant parameters that are presumed to be fixed within the duration of the window. For each frame, spectrum values are computed. Short-time processes are estimated on a frame basis:

$$f_s(n:m) = s(n)w(m-n), \quad (21)$$

where $s(n)$ is the speech signal and $w(m-n)$ is a window of length N_w . The signal samples are segmented into frames of 40ms each with 30ms overlap between consecutive frames.

Voiced frames differentiate themselves from unvoiced frames by means of a high amplitude value, a relative low zero crossing rate and high energy values. Here, a zero-crossing rate is the number of zero-crossings per time unit. We used zero crossing rates and normalized energy to extract voiced regions from the original spectrogram [7].

$$\text{Zero cross} = \text{num_cross} \cdot \frac{f_s}{N}, \quad (22)$$

$$\text{Norm energy} = \frac{1/N \cdot \sum s_n^2}{\text{Range} \cdot \text{MaxRange}}, \quad (23)$$

where, f_s denotes sampling frequency in Hz, i.e., “16000Hz”, N is the frame length in samples, and s_n the n -sample value of the signal. num_cross represents the amount of zero crossings during a frame, Range is the difference between maximum and minimum values in the signal and MaxRange is the maximum feasible range, dependent on the quantification, i.e., “16 bit $\rightarrow$ MaxRange = 65536”.

C. Recognition rate

We obtained the results in Fig. 11.

We compared the recognition rate only Euclidian distance with enhancing ways (SVM, Partial multiplication of vectors, SFM) of the NMF classification. Observing all these together, we can conclude that total recognition rate is improved. The worst case is “Happy” which can be determined from the rate of 62%. However, the best case is “Neutral” emotion with a recognition rate of 95%.

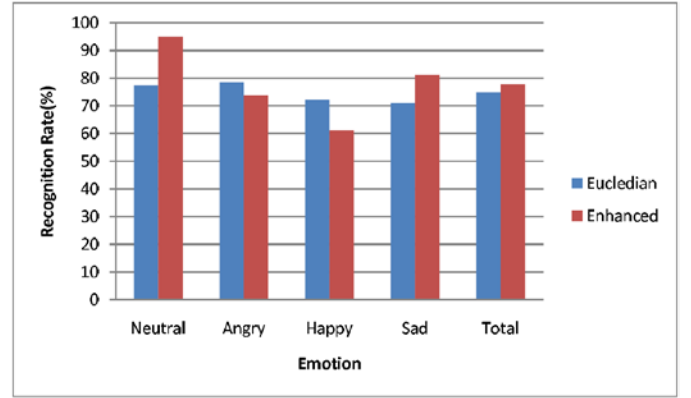


Figure 11. Recognition rate of Euclidian distance and enhanced way

VII. CONCLUSION

In this paper, we have presented a feature representation based upon NMF. We suggested an NMF feature for representing speech emotions. Various clustering methods are suggested for this feature and some of them are rigorously tested. The overall performance is over 75% with the worst case for “happy”. We conclude that the NMF feature indicates both spectral information and temporal information, which is an efficient tool over other spectrum-based features. Sooner or later, frame NMF will show possibilities of real-time emotion recognition that enables to respond immediately.

ACKNOWLEDGMENT

This work was supported by the following funds: the Brain Korea 21 project and the Ministry of Knowledge Economy under Human Resources Development Program for Convergence Robot Specialists in 2010.

REFERENCES

- [1] D. Morrison, R. Wang, and L. C. De Silva, “Ensemble methods for spoken emotion recognition in call-centres,” *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.
- [2] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, vol. 13. MIT Press, 2001, pp. 556–562.
- [3] D. N. Jiang, L. H. Cai, “Speech emotion classification with the combination of statistic features and temporal features,” in *International Conference on Multimedia and Expo*, June 2004, vol. 3, pp. 1967–1970.
- [4] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” in *Data Mining and Knowledge Discovery*, 1998, vol. 2, pp. 121–167.
- [5] T. H. Park, “Towards Automatic Musical Instrument Timbre Recognition,” in *Princeton University*, Nov 2004.
- [6] K. R. Scherer, “Vocal communication of emotion: A review of research paradigms,” *Speech Communication*, vol. 40, pp. 227–256, 2003.
- [7] R. Tato, R. Santos, R. Kompe, and J. M. Pardo, “Emotion space improvement emotion recognition,” in *In Proc. ICSLP 2002*, 2002, pp. 2029–2032.