A Weighted Discrete KNN Method for Mandarin Speech and Emotion Recognition

Tsang-Long Pao, Wen-Yuan Liao and Yu-Te Chen
Department of Computer Science and Engineering, Tatung University
Taiwan, R.O.C.

1. Introduction

Speech signal is a rich source of information and convey more than spoken words, and can be divided into two main groups: linguistic and nonlinguistic. The linguistic aspects of speech include the properties of the speech signal and word sequence and deal with what is being said. The nonlinguistic properties of speech have more to do with talker attributes such as age, gender, dialect, and emotion and deal with how it is said. Cues to nonlinguistic properties can also be provided in non-speech vocalizations, such as laught or cry.

The main investigated linguistic and nonlinguistic attributes in this article were those of audio-visual speech and emotion speech. In a conversation, the true meaning of the communication is transmitted not only by the linguistic content but also by how something is said, how words are emphasized and by the speaker's emotion and attitude toward what is said. The perception of emotion in the vocal expressions of others is vital for an accurate understanding of emotional messages (Banse & Scherer, 1996). In the following, we will introduce the audio-visual speech recognition and speech emotion recognition, which are the applications of our proposed weighted discrete K-nearest-neighbor (WD-KNN) method for linguistic and nonlinguistic speech, respectively.

The speech recognition consists of two main steps, the feature extraction and the recognition. In this chapter, we will introduce the methods for feature extraction in the recognition system. In the post-processing, the different classifiers and weighting schemes on KNN-based recognitions are discussed for the speech recognition. The overall structure of the proposed system for audio-visual and speech emotion recognition is depicted in Fig. 1. In the following, we will briefly introduce the previous researches on audio-visual and speech emotion recognition.

1.1 Audio-visual speech recognition

For past decades, automatic speech recognition (ASR) by machine has been an attractive research topic. However, in spite of extensive research, the performance of current ASR is far from the performance achieved by humans, especially in noisy condition. Most previous ASR systems make use of the acoustic speech signal only and ignore the visual speech cues. They all ignore the auditory-visual nature of speech.

Although acoustic-only-based ASR systems yield excellent results in the laboratory experiment, the error of the recognition can increase in the real world. Noise robust methods

have been proposed. To overcome this limitation, audio speech-reading system, through the use of visual information into audio information, has been considered (Faraj & Bigun, 2007; Farrell et al, 1994; Kaynak et al, 2004). In addition, there has been growing interested in introducing new modalities into the ASR and human–computer interface. With this motivation, enormous research on multi-model ASR has been carried out.

In recent years, there has been many automatic speech-reading systems proposed, that combine audio and visual speech features. For all such systems, the objective of these audio-visual speech recognizers is to improve recognition accuracy, particularly in difficult condition. They most concentrated on the two problems of visual feature extraction and audio-visual fusion. Thus, the audio-visual speech recognition is a work combining the disciplines of image processing, visual-speech recognition and multi-modal data integration. Recent reviews can be found in Chen (Chen & Rao, 1997; Chen, 2001), Mason (Chibelushi et al., 2002), Luettin (Dupont & Luettin, 2000) and Goldschen (Goldschen, 1993).

As above described, most ASR work on detecting speech states investigated speech data which were recorded in quiet environment. But humans are able to perceive emotions even in noisy background (Chen, 2001). In this article we will compare several classifiers for detecting speech from clean and noisy Mandarin speech.
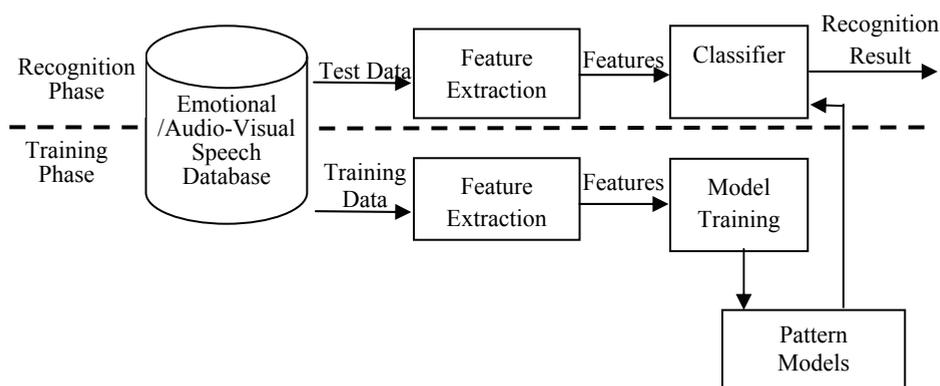

Fig. 1. Overall speech recognition system consisting of the speech extraction and recognition.

1.2 Speech emotion recognition

Besides being described as the audio-visual speech recognition, recognizing emotions from speech has gained increased attention in recent years. There are wide-ranging applications in real word (Huang & Ma, 2006), including health, public safety, education, slogan validation and call center. Taking advantage of the emotional information in speech allows more effective processing of the language information and yields a much more natural human-computer interaction.

Research on understanding and modeling human emotions, a topic that has been predominantly dealt within the fields of psychology and linguistics, is attracting increasing attention within the engineering community. A major motivation comes from the need to improve both the naturalness and efficiency of spoken language human-machine interfaces. Researching emotions, however, is extremely challenging for several reasons. One of the

main difficulties results from the fact that it is difficult to define what emotion means in a precise way. Various explanations of emotions given by scholars are summarized in (Kleinginna & Kleinginna, 1981). Research on the cognitive component focuses on understanding the environmental and attended situations that give rise to emotions; research on the physical components emphasizes the physiological response that co-occurs with an emotion or rapidly follows it. In short, emotions can be considered as communication with oneself and others (Kleinginna & Kleinginna, 1981).

Traditionally, emotions are classified into two main categories: primary (basic) and secondary (derived) emotions (Murray & Arnott, 1993). Primary or basic emotions generally can be experienced by all social mammals (e.g., humans, monkeys, dogs and whales) and have particular manifestations associated with them (e.g., vocal/facial expressions, behavioral tendencies and physiological patterns). Secondary or derived emotions are combinations of or derivations from primary emotions.

Emotional dimensionality is a simplified description of the basic properties of emotional states. According to the theory developed by Osgood, Suci and Tannenbaum (Osgood et al, 1957) and in subsequent psychological research (Mehrabian & Russel, 1974), the computing of emotions is conceptualized as three major dimensions of connotative meaning: arousal, valence and power. In general, the arousal and valence dimensions can be used to distinguish most basic emotions. The locations of emotions in the arousal-valence space are shown in Fig. 2, which provides a representation that is both simple and capable of conforming to a wide range of emotional applications.
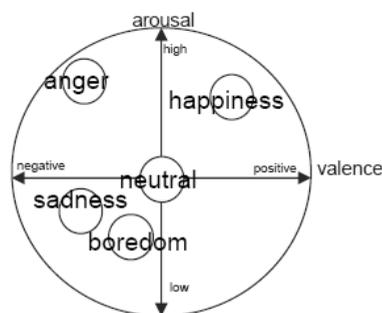

Fig. 2. Graphic representation of the arousal-valence dimension of emotions (Osgood et al. 1957)

1.3 Speech recognition methods

There are various techniques for classification such as K-nearest-neighborhood (KNN), weighted KNN (WKNN) (Dudani, 1976), KNN classification using Categorical Average Patterns (WCAP) (Takigawa *et al*, 2005), Gaussian mixture model (GMM) (Neiberg *et al*, 2006; Reynolds *et al*, 2001; Reynolds & Rose, 1995), Hidden Markov Model (HMM) (Brand et al,1997; Chibelushi *et al*, 2002), Neural Network (NN) and support vector machine (SVM). In our system, the classification was performed using our proposed weighted discrete K-nearest-neighbor (WD-KNN) classifier (Pao *et al*, 2007).

In this chapter we focus on recognition from speech signals and moreover on comparison of different weighting functions applied in various weighted discrete KNN-based classifiers.

The performance is verified by experiments with a Mandarin speech corpus. The baseline performance measure is based on the traditional KNN classifier.

This chapter is organized as follows. In section 2, we introduce the used classifiers and previous researches. In section 3, the feature selection policy and extraction methods for speech and emotion are described. In section 4, an speech emotion recognition system is reviewed and three common weighting functions and the used Fibonacci function are described. Experimental results are given in section 5. In section 6, some conclusions are outlined.

2. Classifiers

The problem of detecting the speech and emotion can be formulated as assignment a decision category to each utterance. Two main types of information can be used to identify the speaker's speech: the semantic content of the utterance and the acoustic features such as variance of the pitch. In the following, we will review various classification and other related literatures.

2.1 KNN

K-nearest neighbor (KNN) classification is a very simple, yet powerful classification method. The key idea behind KNN classification is that similar observations belong to similar classes. Thus, one simply has to look for the class designators of a certain number of the nearest neighbors and sum up their class numbers to assign a class number to the unknown.

In practice, given an instance y, KNN finds the k neighbors nearest to the unlabeled data from the training space based on the selected distance measure. The Euclidean distance is commonly used. Now let the k neighbors nearest to y be $N_k(\mathbf{y})$ and c(z) be the class label of z. The cardinality of $N_k(\mathbf{y})$ is equal to k and the number of classes is l. Then the subset of nearest neighbors within class $j \in \{1, ..., l\}$ is

$$N_k^j(\mathbf{y}) = \{\mathbf{z} \in N_k(\mathbf{y}) : c(\mathbf{z}) = j\} \tag{1}$$

The classification result $j^* \in \{1,...,l\}$ is defined as the majority vote:

$$j^* = \arg \max_j \left| N_k^j(\mathbf{y}) \right| \tag{2}$$

2.2 WKNN

Weighted KNN was proposed by Dudani (Dudani, 1976). In WKNN, the k nearest neighbors are assigned different weights. Let w_i be the weight of the ith nearest samples and $x_1, x_2, ..., x_k$ be the k nearest neighbors of test sample y arranging in increasing distance order. So x_1 is the first nearest neighbor of y. The classification result $j^* \in \{1,...,l\}$ is assigned to the class for which the weights of the representatives among k nearest neighbors sum to the largest value.

$$j^* = \arg \max_j \sum_{p=1}^{k} \begin{cases} w_p, & \text{if } c(\mathbf{x}_p) = j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

2.3 WCAP

The WCAP classification method was proposed by Takigawa for improving performance on handwritten digits recognition (Takigawa et al, 2005). Let w_i^j be the weight of the ith nearest samples of class j. After the k nearest samples of a test sample $\mathbf{y}$, denoted as $\mathbf{x}_i^j$, i=1,…,k, are extracted from class j by using Euclidean distance measure d_i^j. Then, the weight is calculated and normalized by equations which will be described in next section. Finally, the class of a test sample is determined by the following classification rule:

$$ j^* = \arg\min_j \left\{ \left\| \sum_{p=1}^{k} w_p^j \mathbf{x}_p^j - \mathbf{y} \right\|^2 \right\} \tag{4} $$

2.4 HMM

A hidden Markov model (HMM) is a statistical model for sequences of feature vectors that are representative of the input signal (Robert & Granat, 2003; Yamamoto *et al*, 1998). The observed data is assumed to have been generated by an unobservable statistical process of a particular form. This process is such that each observation is coincident with the system being in a particular state. Furthermore it is a first order Markov process: the next state is dependent only on the current state. The model is completely described by the initial state probabilities, the first order Markov chain state-to-state transition probabilities, and the probability distributions of observable outputs associated with each state. HMM has a long history in speech recognition. It has the important advantage that the temporal dynamics of speech features can be caught due to the presence of the state transition matrix. From the experimental results of (Kwon *et al*, 2003), HMM classifiers yielded classification accuracy significantly better than the linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).

2.5 GMM

Gaussian Mixture Models (GMMs) provide a good approximation of the originally observed feature probability density functions by a mixture of weighted Gaussians. The mixture coefficients were computed by use of an Expectation Maximization algorithm. Each emotion is modeled in one GMM. The decision is made for the maximum likelihood model. From the results of (Reynolds *et al*, 2001; Reynolds & Rose, 1995), the authors concluded that using GMMs on the frame level is a feasible technique for speech classification and the results of two models VQ and GMM are not worse than the performance of the HMM.

2.6 WDKNN

We proposed the WD-KNN method for classifying speech and emotion in previous research (Pao *et al*, 2007). Before presenting the proposed method, we describe unweighted-distance KNN classifier as it is the foundation of the method. Without loss of generality, the collected speech samples are split into data elements $x_1,…, x_t$, where t is the total number of training samples. The space of all possible data elements is defined as the input space X. The elements of the input space are mapped into points in a feature space F. In our work, a feature space is a real vector space of dimension d, $\Re^d$. Accordingly, each point f_i in F is

represented by a d-dimensional feature vector. A feature map is defined to be a function that takes an input element in the input space and maps it to a point in the feature space. We use ϕ to define a feature map

$$\phi : X \rightarrow F \tag{5}$$

Let $\boldsymbol{x}_i^j$, $i = 1,\dots,n_j$, be the i-th training sample of class j, where n_j is the number of samples belonging to a class j, $j \in \{1,\dots,l\}$ and l is the number of classes. The total number of training samples is

$$t = \sum_{j=1}^{l} n_j \tag{6}$$

When a test sample $\boldsymbol{y}$ and Euclidean distance measure are given, we obtain the k nearest neighbors belonging to class j, $N_{k,l}^{j}(\boldsymbol{y})$, which can be defined as

$$N_{k,l}^{j}(\mathbf{y}) = \left\{ \mathbf{z} \in N_{k,l}(\mathbf{y}) : c(\mathbf{z}) = j \right\} \tag{7}$$

where the cardinality of the set $N_{k,l}^{j}(\boldsymbol{y})$ is equal to k. Finally the class label of the test sample in unweighted-discrete MKNN classifier is determined by

$$j^* = \arg\min_{j=1,\dots,l} \sum_{i=1}^{k} dist_i^j \tag{8}$$

where $dist_i^j$ is the Euclidean distance between the ith nearest neighbor in $N_{k,l}^{j}(\mathbf{y})$ and the test sample $\boldsymbol{y}$.

Next, we will describe and formulate the weighted-distance KNN classifier (Dudani, 1976; Pao *et al*, 2007) as follows. Among the k nearest neighbors in class j, the following relationship is established:

$$dist_1^j \le dist_2^j \le \dots \le dist_k^j \tag{9}$$

Let w_i be the weight of the ith nearest samples. From above, we know that the one having the smallest distance value $dist_1^j$ is the most important. Consequently, we set the constraint $w_1 \ge w_2 \ge \dots \ge w_k$. Then, the classification result $j^* \in \{1, \dots, l\}$ is defined as

$$j^* = \arg\min_{j=1,\dots,l} \sum_{i=1}^{k} w_i \, dist_i^j \ . \tag{10}$$

In our proposed system, the selection of weights used in the WD-KNN is an important factor for the recognition rate. After extensive investigations and calculations, we found that the Fibonacci sequence weighting function yields the best result in the WD-KNN classifiers. The Fibonacci weighting function is defined as follows

$$w_i = w_{i+1} + w_{i+2}, \quad w_k = w_{k-1} = 1 \tag{11}$$

The definition is in the reverse order of the ordinary Fabonacci sequence. Why Fibonacci weighting function is used? The Fibonacci weighting function indicates that each weight is the sum of the two latter ones. This implies that when the weighted value of 1st nearest neighbor equals to the sum of the weighted values of the latter two neighbors nearest to test sample, the later two added up has the same importance as the first one. The second reason is that we compared different weighting schemes, including Fibonacci weighting, linear distance weighting, inverse distance weighting, and rank weighting, in KNN based classifiers to recognize speech and emotion sates in Mandarin speech (Pao *et al*, 2007). The experimental results show that the Fibonacci weighting function performs better than others.

3. Features extraction

3.1 Acoustic features for emotion recognition

For speech recognition system, a critical step is the extraction and selection of the feature set. Various features relating to pitch, energy, duration, tunes, spectral, and intensity, etc. have been studied in speech recognition and emotion recognition (Murray & Arnott, 1993; Kwon et al, 2003). Due to the redundant information, the forward feature selection (FFS) and the backward feature selection (BFS) are carried out to extract the most representative feature set based on KNN classifier among energy, pitch, formant (F1, F2 and F3), linear predictive coefficients (LPC), Mel-frequency cesptral coefficients (MFCC), first derivative of MFCC (dMFCC), second derivative of MFCC (ddMFCC), Log frequency power coefficients (LFPC), perceptual linear prediction (PLP).
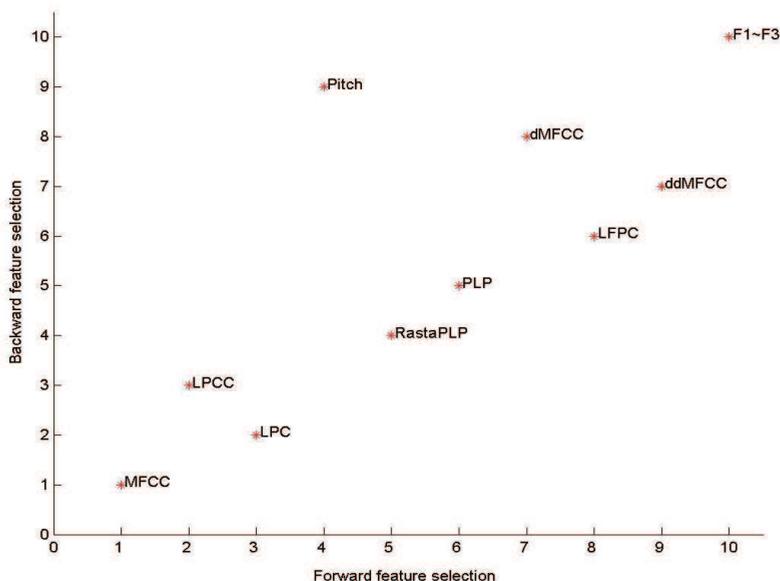

Fig. 3. Features ranking in speech emotion recognition using FFS and BFS by KNN classifier.

For each of these series, mean values are determined to build up the fixed-length feature vector. Besides, emotion was expressed mainly at the utterance level. It is crucial to normalize the feature vector. In this article, we use max-min normalization, to normalize feature vector to the range of [0, 1]. Fig. 3 shows the ranking of these features by KNN classifier. Features near origin are considered to be more important. Finally, we combine MFCC, LPCC and LPC as the best feature set used in the emotion recognition system. The zero-order coefficients of MFCC and LPCC are included as they provide energy information. When we obtain the features from the training and the test data, we can calculate the distance between them to classify the test data.

3.2 Acoustic and visual features for speech recognition

In the previous section, we introduce several features used in speech emotion recognition. For speech recognition, there have been many automatic speech-reading systems proposed recently that combine audio and visual speech features (Poamianos et al, 2003; Zeng et al, 2005; Tang & Li, 2001). For all such systems, the objective of these audio-visual speech recognizers is to improve recognition accuracy, particularly in difficult condition.

Generally speaking, the features for visual speech information extraction from image sequences can be grouped into the following classes: geometric features (Kaynak et al, 2004; Petajan, 1984; Petajan, 1988), model based features (Dupont & Luettin, 2000; Hazen, 2006; Poamianos et al, 2003; Aleksic et al, 2002), visual motion features (DeCarlo & Metaxas, 2000; Faraj & Bigun, 2006), and image based features(Matthews et al, 2002; Zhang et al, 2002). In our system, the visual motion feature is used.

The motion-based feature approach assumes that visual motion during speech production contains relevant speech information. Visual motion information is likely to be robust to different skin reflectance and speakers. Optical flow (DeCarlo & Metaxas, 2000) algorithms are usually used in the estimation of feature point movement. The algorithms usually do not calculate the actual flow field but visual flow field. A further problem consists of the extraction of related features from the flow field. However, recent research about motion-based segmentation got more performance than previous experiments. So the visual motion analysis can improve the performance of recognition.

To estimate the movement of feature point in the image, the estimation method similar to the motion compensation technique used in the video compression is used. The image is partitioned into a set of non-overlapped fixed sized small rectangular blocks. The translation motion within each block is assumed to be uniform. This model only considers translation motion. Other types of motion, such as rotation and zooming, may be approximated by the piecewise translation of these small blocks.

In the audio-visual speech recognition system, the region of interest is first extracted from the original image. The main features, corners or edges of the mouth, are then found as the cues for motion estimation. The motion vectors computation corresponding to the feature points are performed by the block matching based motion estimation algorithm. Finally, the motion vectors for selected feature points are carried out for the feature vectors as the input of the recognition. Fig. 4 shows the example of segmented mouth images and selected feature points, marked by the white dots, and their corresponding motion vectors.

Since speech basically is a non-stationary random process, the characteristics of the acoustic features for audio-visual speech recognition is stochastic and its statistics are time-varying. According to the study of speech production, differences of human speech are generated by

the variation of the mouth and vocal tract. In the audio-visual recognition, we extract the MFCC features, including basic and derived features, for the audio-visual speech recognition.

Fig. 4. Examples of segmented mouth image series for pronunciation "Yi" of Mandarin word "1" and selected feature points in the audio-visual speech recognition system

4. Speech recognition system

4.1 The emotional and audio-visual speech corpus

For the initial stage, an emotional corpus needs to be built up in order to form a base for eliciting emotions from speech signals. In this chapter, five primary emotions, anger, happiness, sadness, boredom, and neutral, are investigated. The emotion corpus database was recorded by 18 males and 16 females who portray 20 prompting Mandarin sentences with the above mentioned emotional states. These sentences are one to six words and are purposely neutral and meaningful so the participants can easily express them with these emotions. Human subjective judgment was conducted to filter out ambiguous emotional utterances for further recognition analysis. Corpora with 80% or higher agreement was kept. As for the audio-visual database, there are some databases exist for the audio-visual research area. But almost all of them are in English or other language, such as Tulips1, AVLetters, M2VTS (Messer et al, 1999), CUAVE (Patterson et al, 2002), etc. The Mandarin database is rare in comparison with other languages. In our experiment, we recorded and created an audio-visual database of Mandarin speech, including Mandarin digits 0 ~ 9.

Our audio-visual database consists of two major parts, one in English and one in Mandarin. The video in English was recorded from 35 speakers while the video in Mandarin was recorded from 40 speakers. The importance of the database is to allow the comparison of recognition of English speech and Mandarin speech. The video is in color with no visual aids given for lip or facial feature extraction. In both parts of database, each individual

speaker was asked to speak 40 isolated English and Mandarin digits, respectively, facing a DV camera.

The video was recorded at a resolution of 320×240 with the NTSC standard of 29.97 fps, using a 1-mega-pixel DV camera. The on-camera microphone was used to record the speeches. Lighting was controlled and a blue background was used to allow change of different backgrounds for further applications. In order to split the video into the visual part and the audio part, we developed a system to decompose the video format (*.avi) into the visual image files (*.bmp) and speech wave files (*.wav) automatically.

4.2 Weighting schemes

As described in Section 1, speech perception by human is a bimodal process characterized by high recognition accuracy and attractive performance degradation in the presence of distortion. There are many classifiers available for decision making such as support vector machines, Bayesian networks, decision trees, artificial neural networks, and fuzzy neural networks, HMM, GMM and KNN-based classifiers. This section focuses on weighting schemes for KNN-based classifiers, which include traditional KNN, WKNN, and WD-KNN. The details about these classifiers are presented in Section 2. The key idea of WKNN, and WD-KNN classifiers is to assign more weights to closer samples by a weighting function to improve the recognition rate. The common weighting functions are as follow. Dudani has proposed a simple linear distance weighting function (Dudani, 1976)

$$w_i = \begin{cases} 1, & \text{if } d_k = d_1 \\ \dfrac{d_k - d_i}{d_k - d_1}, & \text{if } d_k \neq d_1 \end{cases}, \qquad (12)$$

where d_i is the distance to the test sample of the ith nearest neighbor, and d_1 and d_k indicate the distance of the nearest neighbor and the farthest neighbor respectively. Dudani has further proposed an inverse distance weighting function

$$w_i = \frac{1}{d_i} \quad if \quad d_i \neq 0 \qquad (13)$$

and a rank weighting function

$$w_i = k - i + 1 . \qquad (14)$$

From the experimental results done by Dudani, it is well known that weighted version of KNN can improve error rates by using above weighting functions. In this chapter, we propose to use Fibonacci sequence as the weighting function in these classifiers. The Fibonacci weighting function is defined in Eq. (11).

5. Experimental results

5.1 Experimental results for different weighting schemes

First, the value of k in KNN, WCAP and WD-KNN classifiers must be determined. In our previous experiments (Chang, 2005), the distribution of recognition accuracy from clean speech on different k indicates that k sets to 10 can make an acceptable performance with

relatively simple computation in KNN. Therefore, the value of k in KNN, WCAP and WD-KNN classifiers is set to 10.

Table 1 summarizes the experimental results of different weighting functions in speech emotion recognition using various classifiers. The accuracy ranges from 73.8%~76.1%, 73.1%~74.5%, and 78.7%~81.4% in WKN, WCAP, and WD-KNN, respectively. One important finding is that Fibonacci weighting function outperform others in all classifiers. Compared to the baseline attained from KNN method, the largest accuracy improvement of 4.9%, 2.8% and 12.3% can be achieved in these classifiers. The highest recognition rate is 81.4% with WD-KNN classifier weighted by Fibonacci sequence.

		Weighting functions			
		Linear distance	Inverse distance	Rank	Fibonacci
Classifiers	WKNN	75.6%	74.2%	73.8%	76.1%
	WCAP	74.2%	73.6%	73.1%	74.5%
	WD-KNN	78.7%	79.5%	81.2%	81.4%

Table 1. Experimental results of using different weighting functions in speech emotion recognition

In the audio-visual speech recognition, we use our Mandarin speech database as the input data. The database used here contains the Mandarin digits 0 to 9 by 40 speakers. There are a total of 1600 utterances. In the training phase, the 400 utterances of the database containing Mandarin digits 0-9 from all speakers are used as the training set. After we train the model, the other 1200 utterances are used as the testing set in testing phase.

The video stream is a sequence of 17 to 25 images with resolution of 200×120 pixel from the database. Before we compute the visual parameter, some image processing techniques are applied to image in order to make the computation convenient and increase the precision of the visual parameter. In our system, all of image sequences for Mandarin utterance was used in GMM and HMM recognition. In KNN and WD-KNN classifiers, since the distance between the feature vectors is computed, the size of each feature vector must be the same. The images of each utterance used for recognition is selected for a fixed number of images as the fixed-length feature vectors.

Table 2 summarizes the experimental results of different weighting functions in audio-visual speech recognition using various classifiers. The accuracy ranges from 72.8%~79.2%, and 84.6%~98.0% in WKNN and WD-KNN, respectively. The important finding is that

		Weighting functions			
		Linear distance	Inverse distance	Rank	Fibonacci
Classifiers	KNN	74.6%*			
	WKNN	76.4%	74.1%	72.8%	79.2%
	WD-KNN	84.6%	86.3%	88.5%	98.0%

*Weighting function is not used in KNN.

Table 2. Experimental results of using different weighting functions in audio-visual speech recognition

Fibonacci weighting function outperform others in all classifiers. Compared to the baseline attained from KNN method, the largest accuracy improvement of 6.4% and 13.4% can be achieved in these classifiers. The highest recognition rate is 98.0% with WD-KNN classifier weighted by Fibonacci sequence.

5.2 Experimental results using different classifiers on clean and noisy speech

Table 3 demonstrates the emotion recognition accuracy of clean speech and speech interfered by white Gaussian noise from the used classifiers. Accuracy in the table is the average recognition ratio of five emotions. From the results, the proposed WD-KNN is observed outstanding performance at all SNR among the three KNN-based classifiers. Compared with all other methods, the accuracy of WD-KNN is the highest on clean speech and noisy speech from 40dB to 20dB. GMM outperformed others on the 5dB noisy speech. The accuracy of HMM is decreased least among all classifiers from clean speech to 5dB noisy speech.

SNR	KNN	GMM	HMM	WCAP	WD-KNN
Clean	72.2%	70.3%	62.5%	74.5%	81.4%
40dB	65.7%	61.3%	50.2%	51.6%	71.5%
35dB	62.3%	59.7%	49.1%	55.4%	70.3%
30dB	60.8%	60.6%	51.6%	48.0%	67.3%
25dB	61.3%	60.0%	50.5%	44.4%	65.1%
20dB	51.8%	55.7%	39.1%	32.3%	57.1%
15dB	38.7%	47.1%	46.2%	43.6%	45.3%
10dB	32.7%	41.9%	39.6%	36.6%	37.4%
5dB	26.5%	38.7%	35.5%	30.1%	30.6%

Table 3. Comparison of the speech emotion recognition accuracy using five classifiers on clean and noisy speech

Table 4 summarizes the experimental results of audio-visual speech recognition for different classifiers on noise at various SNR values. The accuracy ranges from 78.7%~34.1%, 80.2%~34.3%, 75.1%~42.2%, and 81.8%~45.3% for GMM, HMM, KNN and WD-KNN, respectively. The results show that the WD-KNN with Fibonacci weighting function

SNR	*GMM*	*HMM*	*KNN*	*WD-KNN*
clean	92.5%	99.5%	95.8%	98.0%
30dB	78.7%	80.2%	75.1%	81.8%
25dB	70.4%	75.4%	76.3%	80.3%
20dB	68.5%	70.2%	70.4%	73.4%
15dB	55.2%	62.2%	63.7%	69.2%
10dB	49.2%	57.4%	57.5%	61.5%
5dB	46.5%	52.8%	54.2%	58.4%
0dB	34.1%	34.3%	42.2%	45.3%

Table 4. Comparison of the audio-visual speech recognition rate using different classifiers on clean and noisy speech

outperforms others in most of the cases. Compared to the baseline attained from KNN method, the recognition accuracy improvement of 2.2% to 5.5% at various SNR values can be achieved. In clean condition, the performance of the HMM recognizer seems better than the WD-KNN one. But in the noisy condition, the performance of WD-KNN classifier weighted by Fibonacci sequence is better than other classifiers.

6. Conclusions

In this chapter, we present a speech emotion recognition system to compare several classifiers on the clean speech and noisy speech. Our proposed WD-KNN classifier outperforms the other three KNN-based classifiers at every SNR level and achieves highest accuracy from clean speech to 20dB noisy speech when compared with all other classifiers. Similar to (Neiberg et al, 2006), GMM is a feasible technique for emotion classification on the frame level and the results of GMM are better than performances of the HMM. Although the performance of HMM is the lowest on clean speech, it is robust when the noise increase. The accuracy of KNN dropped rapidly when noise increases from 20dB to 15dB. WCAP performed the same from clean speech to 40dB noisy speech. The accuracy of 10dB noisy speech exceeds 20dB noisy speech in HMM and WCAP classifiers, which are unusual phenomena. In the future, more efforts will be made to investigate these strange results.

Automatic recognition of audio-visual speech aims at building classifiers for classifying audio-visual speech in test audio-visual speech. Until now, several classifiers were adopted independently. Among them, KNN is a very simple but elegant approach to classify various audio-visual speech. Later, some extensions of KNN, such as WKNN and WD-KNN, were proposed to improve the recognition rate.

Moreover, our focus is also to discuss weighting schemes used in different KNN-based classifier, including traditional KNN, weighted KNN and our proposed weighted discrete KNN. The key idea in these classifiers is to find a vector of real-valued weights that would optimize classification accuracy of the classification or recognition system by assigning lower weights to farther neighbors that provide less relevant information for classification and higher weights to closer neighbors that provide more reliable information. Several weighting functions were studied, such as linear distance weighting, inverse distance weighting and rank weighting. In this chapter, we propose to use the Fibonacci sequence as the weighting function. The overall results of the proposed classifier have proved that Fibonacci weighting function in three extended versions of KNN outperform others.

From the experimental results, we can observe that each classifier has their own advantages and disadvantages. How to combine these advantages of each classifier to achieve higher recognition rate requires further study. Besides, how to get an optimal weighting sequence is also deserved to be investigated.

7. References

P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos (2002), "Audio-Visual Speech Recognition Using MPEG-4 Compliant Visual Features", *EURASIP Journal on Applied Signal Processing*, No. 11, pp. 1213-1227, 2002

R. Banse & K. R. Scherer (1996), "Acoustic Profiles in Vocal Emotion Expression," *Journal of Personality and Social Psychology* 70, pp. 614-636. 1996.

M. Brand, N. Oliver & A. Pentland (1997), "Coupled hidden Markov models for complex action recognition," *Proc. IEEE CCVPR*, pp. 994–999, 1997.

Y. H. Chang (2005), "Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification", *Conference on Computational Linguistics and Speech Processing XVII* (ROCLING XVII), pp. 96-105, 2005.

T. Chen & R. Rao(1997), "Audiovisual interaction in multimedia communication," *ICASSP*, vol. 1. Munich, pp. 179-182, Apr. 1997.

T. Chen (2001), "Audio-visual speech processing," *IEEE Signal Processing Magazine*, Jan. 2001.

C. C. Chibelushi, F. Deravi & J. S. D. Mason (2002), "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, pp. 23-37, Feb. 2002.

D. DeCarlo & D. Metaxas (2000), "Optical Flow Constraints on Deformable Models with Applications to Face Tracking," *Int'l J. Computer Vision*, vol. 38, no. 2, pp. 99-127, 2000.

S. Dupont & J. Luettin (2000), "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, Sept. 2000.

S. A. Dudani (1976), "The Distance-Weighted K-Nearest-Neighbor Rule," *IEEE Trans Syst. Man Cyber* (1976) 325-327, 1976.

M.I. Faraj & J. Bigun (2006), "Person Verification by Lip-Motion," *Proc. Conf. Computer Vision and Pattern Recognition Workshop* (CVPRW '06), pp. 37-45, 2006.

M. I. Faraj & J. Bigun (2007), "Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition", *Computers, IEEE Transactions,* Vol. 56, No. 9 pp.1169 – 1175, Sept. 2007.

K. Farrell, R. Mammone & K. Assaleh (1994), "Speaker Recognition Using Neural Networks and Conventional Classifiers," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, pp. 194-205, 1994.

T. J. Hazen (2006), "Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 1082-1089, 2006.

R. Huang & C. Ma (2006), "Toward a Speaker-Independent Real-Time Affect Detection System", Vol. 1, *the 18th International Conference on Pattern Recognition*, pp. 1204-1207, 2006.

M. N. Kaynak, Q. Zhi, etc (2004), "Analysis of Lip Geometric Features for Audio-Visual Speech Recognition," *IEEE Transaction on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 34, pp. 564-570, July 2004.

P. R. Kleinginna & A.M. Kleinginna (1981), "A Categorized List of Emotion Definitions with Suggestions for a Consensual Definition," *Motivation and Emotion*, 5(4), pp.345-379, 1981.

O. W. Kwon, K. Chan, J. Hao & T. W. Lee (2003), "Emotion Recognition by Speech Signals", *Proceedings of EUROSPEECH*, pp. 125-128, 2003.

I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox & R. Harvey (2002), "Extraction of visual features for lipreading," *IEEE Trans. pattern analysis and machine intelligence*, vol. 24, pp. 198-213, 2002.

K. Messer, J. Matas, J. Kittler & J. Luettin (1999), "Xm2vtsdb: The Extended M2VTS Database," *Proc. Second Int'l Conf. Audio- and Video-Based Biometric Person Authentication* (AVBPA '99), pp. 72-77, 1999.

I. Murray & J. L.Arnott (1993), "Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion," *Journal of the Acoustic So-ciety of America* 93(2), pp. 1097-1108, 1993.

D. Neiberg, K. Elenius & K. Laskowski (2006), "Emotion Recognition in Spontaneous Speech Using GMMs", *In Proc. of Interspeech 2006*. Pittsburg, pp.809-812, 2006.

C. E.Osgood, J.G. Suci & P. H. Tannenbaum (1957)., The Measurement of Meaning. The University of Illinois Press, Urbana, 1957.

T. L. Pao, Y. M. Cheng, Y. T. Chen & J. H. Yeh (2007), "Performance Evaluation of Different Weighting Schemes on KNN-Based Emotion Recognition in Mandarin Speech," *International Journal of Information Acquisition*, Vol. 4, No. 4, pp. 339-346, Dec. 2007.

E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy (2002), "Moving-Talker, Speaker-Independent Feature Study, and Baseline Results Using the CUAVE Multimodal Speech Corpus," *EURASIP Journal on Applied Signal Processing*, No. 11, pp. 1189-1201, 2002

E. D. Petajan, N. M. Brooke, B. J. Bischoff & D. A. Boddoff (1988), "Experiments in automatic visual speech recognition," in *Proc. 7th FASE Symp.*, Book 4, pp. 1163–1170, 1988.

E. D. Petajan (1984), "Automatic lipreading to enhance speech recognition," in *Proc. IEEE Global Telecommunications Conf.*, Atlanta, GA, pp. 265-272, Nov.1984.

G. Poamianos, etc (2003),"Recent Advances in the Automatic Recognition of Audiovisual Speech" *Proceeding of the IEEE*, Vol. 91, No. 9, September 2003.

D. Reynolds, T. Quatieri & R. B. Dunn (2001), "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, nos. 1-3, pp. 9-21, 2001.

D. Reynolds & R. Rose (1995), "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

E. Robert & A. Granat (2003), "A Method of Hidden Markov Model Optimization for Use with Geophysical Data Sets", *International Conference on Computational Science*, pp. 892-901, 2003.

Y. Takigawa, S. Hotta, S. Kiyasu & S. Miyahara, (2005), "Pattern Classification Using Weighted Average Patterns of Categorical k-Nearest Neighbors" *Proc. of the 1th International Workshop on Camera-Based Document Analysis and Recognition* 111-118, 2005

X. Tang & X. Li (2001), "Fusion of Audio-Visual Information Integrated Speech Processing," *Proc. Third Int'l Conf. Audio- and Video-Based Biometric Person Authentication* (AVBPA '01), pp. 127-143, 2001.

E. Yamamoto, S. Nakamura & K. Shikano (1998), "Lip Movement Synthesis from Speech Based on Hidden Markov Models," *J. Speech Comm.*, vol. 26, no. 1, pp. 105-115, 1998.

X. Zhang, C. C. Broun, R. M. Mersereau & M. A. Clements (2002), "Automatic Speechreading with Applications to Human-Computer Interface", *EURASIP Journal on Applied Signal Processing*, No. 11, pp. 1228-1247, 2002

Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu & T. S. Huang (2005), "Audio-visual affect recognition in activation-evaluation space", *Proc. IEEE International Conference on Multimedia and Expo*, pp. 828–831, July 2005.

Speech Recognition

Edited by France Mihelic and Janez Zibert

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Tsang-Long Pao, Wen-Yuan Liao and Yu-Te Chen (2008). A Weighted Discrete KNN Method for Mandarin Speech and Emotion Recognition, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from:
http://www.intechopen.com/books/speech_recognition/a_weighted_discrete_knn_method_for_mandarin_speech_and_emotion_recognition

INTECH
open science | open minds