mood model. $M$ log-likelihood results are then submitted to the emotion decision process.

4.1.2. Bimodal Emotion Recognition from Facial Expression and Speech. Facial expression and speech are representative indicators that directly convey human emotional information. Because those indicators provide emotional information that is supplementary and/or complementary to each other, they have been successfully combined in terms of bimodal indicators. The bimodal emotion recognition approach integrates the recognition results respectively obtained from face and speech.

In facial expression recognition, accurate detection of the face has an important influence on the recognition performance. A bottom-up, feature-based approach is widely used for the robust face detection. This approach searches an image through a set of facial features indicating color and shape, and then groups them into face candidates based on the geometric relationship of the facial features. Finally, a candidate region is decided as a face by locating eyes in the eye region of a candidate's face. The detected facial image is submitted to the module for facial expression recognition.

The first step of facial expression recognition is to normalize the captured image. Two kinds of features are then extracted on the basis of Ekman's facial expression features [28]. The first feature is a facial image consisting of three facial regions: the lips, eyebrows, and forehead. By applying histogram equalization and the threshold of the standard distribution of the brightness of the normalized facial image, each of the facial regions is extracted from the entire image. The second feature is an edge image of those three facial regions. The edges around the regions are extracted by using histogram equalization.

Next, the facial features are trained according to a specific classifier in order to determine explicitly distinctive boundaries between emotions. The boundary is used as a criterion to decide an emotional state for a given facial image. Various techniques already in use for conventional pattern classification problems are likewise used for such emotion classifiers. Among them, neural network-based approaches have been widely adopted for facial emotion recognition, and have provided reliable performance [29–31]. A recent work reported the efficiency of the back propagation (BP) neural network algorithm proposed by Rumelhart and McClelland in 1985 [32]. In this study, we follow a training procedure introduced in [31] that uses an advanced BP algorithm called error back propagation.

Each of the extracted features is trained using two neural networks for each type of emotion. Each neural network is composed of 1610 input nodes, six hidden nodes, and $M$ output nodes. The 1610 input nodes receive 1610 pixels from the input image, and the output nodes respectively correspond to each of $M$ emotions. The number of hidden nodes was determined by an experiment. Finally, the decision logic determines the final emotion from the two neural network results. The face-emotion decision logic utilizes the weighted sum of the two results and a voting method of the result transitions over the time domain. The overall process of emotion recognition through facial expression is shown in Figure 3.

Once audio signals transmitted through a robot microphone are determined to be human voice signals, the speech emotion recognition module is activated. In the first step, several acoustic features representing emotional characteristics are estimated from the voice signals. Two types of acoustic features are extracted: a phonetic feature and a prosodic feature. MFCC and LPC pertaining to musical mood recognition are also employed for speech emotion recognition in terms of phonetic features, while spectral energy and pitch are used as prosodic features. As in musical mood recognition, the first and second derivatives of all features are added to the feature set.

Next, the acoustic features are recognized through a pattern classifier. Even though various classifiers such as HMM and SVM have been fed into speech emotion recognition tasks, we employ the neural network-based classifier used in the facial expression recognition module in order to efficiently handle the fusion process in which the recognition results of two indicators are integrated. We organize a sub-neural network for each emotion. The construction of each sub-network has basically the same architecture. A sub-network comprises input nodes corresponding to the dimension of the acoustic features, hidden nodes, and an output node. The number of hidden nodes varies according to the distinctness of respective emotions. When there are $M$ emotions, acoustic features extracted from the voice signals are simultaneously fed into $M$