A Method Using Acoustic Features to Detect Inadequate Utterances in Medical Communication

Michihisa Kurisu Kazuya Mera Ryunosuke Wada Yoshiaki Kurosawa Toshiyuki Takezawa

Graduate School of Information Sciences
Hiroshima City University
Hiroshima, Japan
{kurisu_m, mera, wada, kuro, takezawa}@ls.info.hiroshima-cu.ac.jp

Abstract— **We previously proposed a method that uses grammatical features to detect inadequate utterances of doctors. However, nonverbal information such as that conveyed by gestures, facial expression, and tone of voice are also important. In this paper, we propose a method that uses eight acoustic features to detect three types of mental states (sincerity, confidence, and doubtfulness/acceptance). A Support Vector Machine (SVM) is used to learn these features. Experiments showed that the system's accuracy and recall rates respectively ranged from 0.79-0.91 and 0.80-0.94.**

Keywords: Mental State; Acoustic Features; Support Vector Machine

I. Introduction

Recently, not only western countries but also Japan are paying increasing attention to medical communication between doctors and patients. Although specific medical communication techniques are taught and learned in the course of medical education, communication gaps between doctors and patients still remain. One reason for this is that patients often cannot correctly understand the real intention behind doctors' utterances.

To address this problem, we previously proposed a method to detect inadequate utterances by doctors through the use of a Support Vector Machine (SVM) which uses grammatical features obtained by analyzing a dialogue corpus [1]. However it is very difficult to deal with "mental states" such as "confidence" and "sincerity" through the use of text information alone. To deal with these, it is important to consider nonverbal factors such as gestures, facial expression, and acoustic features.

There are many researches to classify mental states such as emotion [2, 3], intention [4], and attitude [5] by using acoustic features. In order to detect inadequate utterances, we try to classify three types of mental states (sincerity, confidence, and doubtfulness/acceptance). Furthermore, various methods are utilized to classify mental states. Schuller experimented and compared the classifying results of linear classifiers, Gaussian Mixture Models, Neural Nets, and SVM. As a result, the error rate of SVM was the lowest [2]. Therefore, we also use SVM to classify mental states.

In this paper, we propose a method that uses acoustic features to detect three types of mental states. In the method, an SVM uses eight acoustic features (Maximum of the pitch contour, Mean of the pitch contour, Dynamic range of the pitch contour, Maximum of the power contour, Mean of the power contour, Speaking rate, Ratio of upslope of the pitch contour, Ratio of upslope of the power contour) to detect three types of mental states (sincerity, confidence, and doubtfulness/acceptance).

II. Support Vector Machine for Classification [6]

In this section, we briefly review some fundamentals relevant to the use of an SVM for classification problems. For further details, refer to [7, 8, 9].

Given a training set of N data points $\{y_k, x_k\}_{k=1}^N$, where $x_k \in \Re^n$ is the k-th input pattern and $y_k \in \Re$ is the k-th output pattern, the support vector method approach aims at constructing a classifier of the form:

$$y(x) = sign[\sum_{k=1}^{N} \alpha_k y_k \psi(x, x_k) + b] \tag{1}$$

where α_k are positive real constants and b is a real constant. For $\psi(\cdot,\cdot)$ one typically has the following choices: $\psi(x, x_k) = x_k^T x$ (linear SVM); $\psi(x, x_k) = (x_k^T x + 1)^d$ (polynomial SVM of degree d); $\psi(x, x_k) = \exp\{-\| x - x_k \|^2 / \sigma^2\}$ (RBF SVM), $\psi(x, x_k) = \tanh[\kappa x_k^T x + \theta]$ (two layer neural SVM), where σ, κ, and θ are constants.

The classifier is constructed as follows. One assumes that

$$\begin{cases} w^T \varphi(x_k) + b \geq 1, & if \;\; y_k = +1 \\ w^T \varphi(x_k) + b \leq -1, & if \;\; y_k = -1 \end{cases} \tag{2}$$

which is equivalent to

$$y_k[w^T\varphi(x_k)+b]\geq 1, \quad k=1,...,N \tag{3}$$

where $\varphi(\cdot)$ is a nonlinear function which maps the input space into a higher dimensional space. However, this function is not explicitly constructed. In order to make it possible to violate (3), in case a separating hyperplane in this higher dimensional space does not exist, variable ξ_k is introduced such that

$$\begin{cases} y_k[w^T\varphi(x_k)+b]\geq 1-\xi_k, \quad k=1,...,N \\ \xi_k \geq 0, \quad k=1,...,N. \end{cases} \tag{4}$$

According to the structural risk minimization principle, the risk bound is minimized by formulating the optimization problem:

$$\min_{w,\xi_k} \mathrm{J}_1(\mathrm{w},\xi_\mathrm{k}) = \frac{1}{2}w^T w + c\sum_{k=1}^{N}\xi_k \tag{5}$$

subject to (4). Therefore, one constructs the Lagrangian

$$L_1(w,b,\xi_k;\alpha_k,v_k) = J_1(w,\xi_k) - \sum_{k=1}^{N}\alpha_k\{y_k[w^T\varphi(x_k)+b]-1+\xi_k\} - \sum_{k=1}^{N}v_k\xi_k \tag{6}$$

by introducing Lagrange multipliers $\alpha_k \geq 0, v_k \geq 0 (k=1,...,N)$. The solution is given by the saddle point of the Lagrangian by computing

$$\max_{\alpha_\kappa, v_k}\min_{w,b,\xi_k} L_1(w,b,\xi_\kappa;\alpha_k,v_k). \tag{7}$$

One obtains

$$\begin{cases} \dfrac{\partial L_1}{\partial w} = 0 \to w = \sum_{k=1}^{N}\alpha_k y_k\varphi(x_k) \\ \dfrac{\partial L_1}{\partial b} = 0 \to \sum_{k=1}^{N}\alpha_k y_k = 0 \\ \dfrac{\partial L_1}{\partial \xi_k} = 0 \to 0 \leq \alpha_k \leq c, k=1,...,N \end{cases} \tag{8}$$

which leads to the solution of the following quadratic programming problem:

$$\max_{\alpha_k} Q_1(\alpha_k;\varphi(x_k)) = -\frac{1}{2}\sum_{k,l=1}^{N} y_k y_l \varphi(x_k)^T\varphi(x_l)\alpha_k\alpha_l + \sum_{k=1}^{N}\alpha_k \tag{9}$$

such that

$$\begin{cases} \sum_{k=1}^{N}\alpha_k y_k = 0 \\ 0 \leq \alpha_k \leq c, k=1,...,N. \end{cases}$$

The function $\varphi(x_k)$ in (9) is related then to $\psi(x,x_k)$ by imposing

$$\varphi(x)^T\varphi(x_k) = \psi(x,x_k), \tag{10}$$

which is motivated by Mercer's Theorem. Note that for the two layer neural SVM, Mercer's condition only holds for certain parameter values of κ and θ.

The classifier (1) is designed by solving

$$\max_{\alpha_k} Q_1(\alpha_k;\psi(x_k,x_l)) = -\frac{1}{2}\sum_{k,l=1}^{N} y_k y_l \psi(x_k,x_l)\alpha_k\alpha_l + \sum_{k=1}^{N}\alpha_k \tag{11}$$

subject to the constraints in (9). One does not have to calculate w or $\varphi(x_k)$ in order to determine the decision surface. Because the matrix associated with this quadratic programming problem is not indefinite, the solution to (11) will be global [10].

Furthermore, one can show that hyperplanes (3) satisfying the constraint $\| w \|_2 \leq a$ have a VC-dimension h which is bounded by

$$h \leq \min([r^2 a^2], n)+1 \tag{12}$$

where $[\cdot]$ denotes the integer part and r is the radius of the smallest ball containing the point $\varphi(x_1),...,\varphi(x_N)$. Finding this ball is done by defining the Lagrangian

$$L_2(r,q,\lambda_k) = r^2 - \sum_{k=1}^{N}\lambda_k(r^2 - \| \varphi(x_k) - q \|_2^2) \tag{13}$$

where q is the center of the ball and λ_k are positive Lagrange multipliers. In a similar way as for (5) one finds that the center is equal to $q = \sum_k \lambda_k\varphi(x_k)$, where the Lagrange multipliers follow from

$$\max_{\lambda_k} Q_2(\lambda_k;\varphi(x_k)) = -\sum_{k,l=1}^{N}\varphi(x_k)^T\varphi(x_l)\lambda_k\lambda_l + \sum_{k=1}^{N}\lambda_k\varphi(x_k)^T\varphi(x_k) \tag{14}$$

such that

$$\begin{cases} \sum_{k=1}^{N}\lambda_k = 1 \\ \lambda_k \geq 0, k=1,...,N. \end{cases}$$

Based on (10), Q_2 can also be expressed in terms of $\psi(x_k,x_l)$. Finally, one selects a support vector machine with minimal VC dimension by solving (11) and computing (12) from (14).

We propose a method to calculate mental states by using an SVM. The SVM learns acoustic features in voice data. The acoustic features are explained in chapter 3.

III. ACOUSTIC FEATURES

In order to calculate mental states of the speaker, we use the following eight features.

- $f0_{max}$: Maximum of the pitch contour
- $f0_{mean}$: Mean of the pitch contour
- $f0_{range}$: Dynamic range of the pitch contour
- $power_{max}$: Maximum of the power contour

- $power_{mean}$: Mean of the power contour
- *speaking-rate* : Speaking rate
- $f0_{slope}$: Ratio of the sample number of the upslope to that of the downslope for the pitch contour
- $power_{slope}$: Ratio of the sample number of the upslope to that of the downslope for the power contour

The use of various acoustic features for analyzing emotion from speech has been reported in many studies [3, 4, 11, 12, 13]. From these features, we chose to use six basic acoustic features ($f0_{max}$, $f0_{mean}$, $f0_{range}$, $power_{max}$, $power_{mean}$, and *speaking-rate*). By way of comparison, Chuang's method for recognizing emotion from speech and text uses 33 acoustic features including $f0_{slope}$ and $power_{slope}$ [14].

The values of the acoustic features are calculated as follows. First of all, power and pitch values are extracted for each sampling time. Next, silent parts and voiceless parts are cut based on pitch value. Then, maximum, minimum, and mean of the pitch contour are calculated. Dynamic range indicates the ratio of $f0_{max}$ to $f0_{min.}$ Maximum and mean of power are also calculated in the same way. Voice activity time divided by the number of moras in the utterance is defined as *speaking-rate*.

$f0_{slope}$ and $power_{slope}$ represent not only the slope but also the shape of each vibration in the contour. Fig. 1 shows the difference between these parameters. In this figure, each part shows the vibration of a contour. In order to show how the parameters are used, we assume that the length and the amplitude of these two contours are the same. In part A, the length of the upslope contour is longer than that of the downslope contour, while the opposite is shown in part B. The ratio of upslope to downslope is 3.14 (22 upslope samples to 7 downslope samples) in part A and 0.26 (6 upslope samples to 23 downslope samples) in part B [14].

IV. EXPERIMENTATION

A. Test Data

The results obtained in interviewing patients who had consulted doctors revealed that the patients prefer doctors who are warm-hearted, confident, easy to talk to, not repetitious, and who listen closely and carefully to patient's utterances [15]. On the basis of these results, the following test data were prepared for the following utterances:

- Whether the utterance, "*ARIGATOU* (Thank you.)" was spoken sincerely (A) or not (B)
- Whether the utterance, "*KAZEDESU* (You've got a cold.)" was spoken confidently (A) or not (B)
- Whether the utterance, "*SOUDESUKA* (Really?/I see.)" indicated doubtfulness (A) or acceptance (B)

Ten male university students spoke all six utterances five times each. Three subjects checked the recorded data whether were ambiguous or not. Data that all three subjects classified correctly were used for experimentation. Data that two subjects classified correctly and the other did not classify were also used. Table 1 shows the volume of test data obtained.

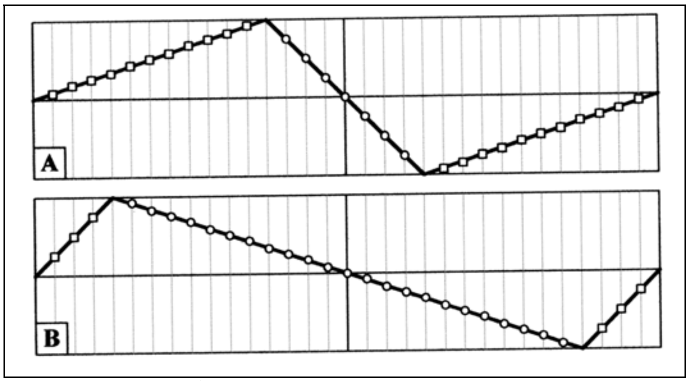

Figure 1. The ratio of the up-slope sample number to the down-slope sample number. Two contours with the same wavelength are shown in parts A and B; the square symbols indicate down-slope samples.

TABLE I. Test Data Volume

Utterance	A	B	Total
ARIGATOU	27	39	66
KAZEDESU	47	34	81
SOUDESUKA	42	51	93

Before inputting the acoustic features into SVM, the values of the features were standardized. The following standardization process was used for $f0_{max}$. Variable *j* is the number of speakers in the range [1, 10]. *i* indicates an utterance. Therefore, $f0max_{ij}$ indicates *f0max* of the *i*th data by speaker *j*. First, the mean of all $f0_{max}$ is calculated by using (15). Next, the mean of $f0_{max}$ for each speaker is calculated by using (16).

Finally, standardized $f0max'_{ij}$ is calculated by using (17). The other features are also standardized in the same way.

$$\overline{f0_{\max}} = \frac{1}{i*j}\sum_i\sum_j f0_{\max ij} \tag{15}$$

$$\overline{f0_{\max j}} = \frac{1}{i}\sum_i f0_{\max ij} \tag{16}$$

$$f0_{\max}'_{ij} = f0_{\max ij} + (\overline{f0_{\max}} - \overline{f0_{\max j}}) \tag{17}$$

B. Experimental Results

The eight acoustic features described in section 3 were calculated for all voice data. The SVM used the Leave-One-Out method to learn and classify the data. All possible combinations of the eight features were examined. Table 2 shows the experimental results. These results are the best values obtained in all combinations of the features.

TABLE II. Experimental Results

Mental State	**Accuracy**	**Recall**
Sincerity	0.79	0.82
Insincerity	0.87	0.80
Confidence	0.90	0.94
Uncertainty	0.91	0.85
Doubtfulness	0.80	0.83
Acceptance	0.84	0.84

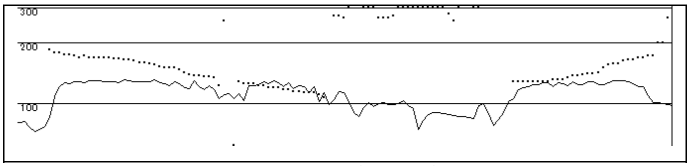

Figure 2. Pitch contour of “Doubtfulness” utterance

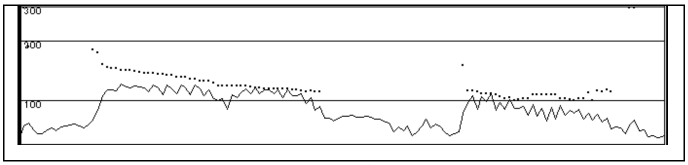

Figure 3. Pitch contour of “Acceptance” utterance

The best combination of the acoustic features to classify “sincerity” was (*$f0_{max}$*, *$power_{max}$*, *$power_{mean}$*, *speaking-rate*, *$f0_{slope}$*), and that for “confidence” was (*$power_{max}$*, *speaking-rate*, *$f0_{slope}$*, *$power_{slope}$*). The best combination to classify doubtfulness was (*$f0_{slope}$*, *$power_{slope}$*) and that to classify acceptance was (*$f0_{max}$*, *$f0_{mean}$*, *$power_{max}$*, *$power_{mean}$*, *speaking-rate*, *$f0_{slope}$*, *$power_{slope}$*).

Fig.2 and Fig.3 show the pitch contours for “doubtfulness” and “acceptance,” respectively. In the former the pitch rises at the end, but in the latter it does not. This kind of acoustic feature can be detected by using *$f0_{slope}$*.

V. Conclusion

In this paper, we proposed a method to calculate a speaker’s mental states (sincerity, confidence, and doubtfulness/acceptance) using eight acoustic features (*$f0_{max}$*, *$f0_{mean}$*, *$f0_{range}$*, *$power_{max}$*, *$power_{mean}$*, *speaking-rate*, *$f0_{slope}$*, *$power_{slope}$*). The features were learned by an SVM and experiments confirmed that accuracy and recall rates of 0.79-0.91 and 0.80-0.94 respectively were obtained. This method was combined with our previously reported method that uses grammatical features to detect inadequate utterances. The precision and recall rates obtained by combining these methods were 0.84 and 0.49, respectively.

We had previously proposed a method to detect inadequate utterances in medical communication based on eight grammatical features [1]. In future work, we will attempt to ascertain the best combinations of the acoustic and grammatical features to improve precision and recall rates.

References

[1] R. Wada, K. Mera, Y. Kurosawa, and T. Takezawa, Inadequate utterance detection method in medical communication based on grammatical features using SVM, Proceeding of IEEE SMC Hiroshima Chapter Young Researchers’ workshop, pp.61-64, 2011. (in Japanese)

[2] B. Schuller, G. Rigoll, and M. Lang, Speech recognition combining acoustic features and linguistic information in a hybrid support vector machine – belief network architecture, IEEE International Conference on Acoustics, Speech, and Signal Processing 2004 (ICASSP 2004), Vol. 1, pp.577-580, 2004.

[3] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, Designing emotional speech corpus and its acoustic analysis : discrimination of one emotion appeared during the dialog over voice chat system for MMORPG, IPSJ SIG Technical Report, Vol. 2008, No. 12, pp.133-138, 2008. (in Japanese)

[4] S. Yoshikawa, S. Makimoto, H. Kashioka, and N. Campbell, Intention classification in non-verbal utterance based on acoustic features, IPSJ SIG Technical Report, Vol. 2008, No. 12, pp.1750180, 2008. (in Japanese)

[5] Z. Callejas, and R. López-Cózar, Influence of contextual information in emotion annotation for spoken dialogue systems, Speech Communication, Vol. 50, pp.416-433, 2008.

[6] J.A.K. Suykens and J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters, vol. 9, no. 3, pp.293-300, 1999.

[7] V. Vapnik, The nature of statistical learning theory, Springer-Verlag, 1995.

[8] V. Vapnik, Statistical learning theory, John Wiley, 1998.

[9] V. Vapnik, The support vector method of function estimation, In Nonlinear modeling: advanced black-box techniques, J.A.K. Suykens, J. Vandewalle (Eds.), Kluwer Academic Publishers, pp.55-85, 1998.

[10] R. Fletcher, Practical methods of optimization, Chichester and New York: John Wiley and Sons, 1987.

[11] R. Banse and K. R. Scherer, Acoustic profiles in vocal emotion expression, Journal of Personality and Social Psychology, vol. 70, no. 3, pp.614-636, 1996.

[12] Y. Arimoto, S. Ohno, and H. Iida, An estimation method of degree of speaker’s anger emotion with acoustic and linguistic features, Journal of Natural Language Processing, vol.14, no. 3, pp.131-145, 2007. (in Japanese)

[13] A. Iida, S. Iga, and M. Yasumura, Study of emotion in speech : Findings from perceptual experiments, IPSJ SLP Technical Report, Vol.97, No.16, pp.113-118, 1997. (in Japanese)

[14] Z. Chuang and C. Wu, Multi-modal emotion recognition from speech and text, Computational Linguistics and Chinese Language Processing, vol.9, no. 2, pp.45-62, 2004.

[15] M. Lloyd and R. Bor, Communication skills for medicine, Churchill Livingstone, 2000.

[16] M. Kurisu, K. Mera, Y. Kurosawa, and T. Takezawa, Inadequate utterance detection method in medical communication based on acoustic features, Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing, pp.639-641, 2012. (in Japanese)