Hindawi
Mobile Information Systems
Volume 2022, Article ID 6095859, 9 pages
https://doi.org/10.1155/2022/6095859

Research Article

Keyword Detection of Japanese Media Teaching Based on Support Vector Machines and Speech Detection

Bo Qiu

Qiqihar University, Qiqihar 161006, China

Correspondence should be addressed to Bo Qiu; qiubo0312@qqhru.edu.cn

Received 21 March 2022; Revised 9 June 2022; Accepted 20 June 2022; Published 7 July 2022

Academic Editor: Imran Shafique Ansari

The keyword detection of Japanese speech in streaming media has a certain effect on our study of Japanese information and a certain promotion effect on Japanese teaching. Currently, there is a problem of stability in the detection model of Japanese speech keywords. In order to improve the detection effect of Japanese speech keywords in streaming media, based on SVM, this study constructed a detection model of Japanese speech keywords in streaming media based on support vector machine. Moreover, this study analyzes the problem of SVM probability output and the comprehensive problem of SVM confidence, etc. In addition, by comparing the effect of confidence synthesis with the arithmetic average method, we found that the confidence obtained by SVM can obtain a higher recognition rate under the same rejection rate and improve the overall performance of the system. Finally, this study uses the difference comparison test to analyze the performance of the model proposed in this study. The research results show that the algorithm proposed in this paper has good performance and can be used as a follow-up system algorithm.

1. Introduction

Human communication is inseparable from speech, and speech has become an important way for human communication. Making the machine understand human speech is the research direction of speech recognition technology in the past few decades, and the potential market in this respect is huge. At present, people can use speech recognition technology to order the machine to do some simple and cumbersome work or human beings cannot get involved in the environment [1]. In recent years, speech recognition has made considerable progress. The reason lies in the improvement of speech models (acoustic models, language models, etc.) and recognition algorithms by scientific researchers and the development of the current state of the art in speech recognition-related fields. However, there is still a big gap in the use of speech recognition technology to achieve civilianization, and the speech recognition systems currently developed are all used in specific fields and cannot be restricted to different fields. To solve this problem, we still need to improve the speech model and the recognition algorithm. For language communication between human beings, a sentence that others say may not be fully understood, but as long as we remember a few keywords in this sentence, we can roughly guess the meaning of this sentence. In terms of speech recognition technology, for a continuous speech stream, we do not need to recognize every word in the continuous speech stream, but we only need to grab a few common keywords to guess the meaning of the entire speech stream. Moreover, when humans say a sentence, they often highlight keywords, which makes the pronunciation of the keyword part in the continuous speech stream is accurate and clear [2]. The process of identifying keywords in a continuous speech stream is called keyword speech recognition. As can be seen from the above description, keyword speech recognition is much simpler than continuous speech recognition, but its breadth of market application is not worse than that of continuous speech recognition. Keyword recognition, as the name suggests, is to identify some keywords in the continuous speech stream. However, among these speeches to be recognized, it is impossible to have only the speech of human speaking; there may also be human breathing, sneezing, or noise in the external environment (car horn sound, construction site construction sound, etc.)

[3]. The inclusion of these nonhuman speeches increases the difficulty of keyword recognition. There are already formed keyword speech recognition systems in our daily life, such as our commonly used speech mobile phone dialing function, speech search function, and automatic speech response function.

Japanese speech has some difficulties in recognition process compared to English. Especially in extracting keywords, it is more difficult to extract Japanese speech. Some keywords in streaming media can be used to interpret the information, so Japanese speech keyword detection needs to be performed through speech recognition technology.

2. Related Work

The literature proposed the concept of "given word." Shortly thereafter, Christiansen and others proposed the concept of "keywords" [4]. After that, literature used LPC technology to locate keywords [5]. Since the 1980s, researchers began to study keyword detection algorithms based on dynamic time bending. The literature implemented a keyword detection system using template connection methods [6]. After that, the literature implemented a keyword detection system based on the HMM model for a small number of telephone terms [7]. At the same time, the literature proposed a performance evaluation benchmark for the keyword detection system [8]. After 90 years of the last century, based on HMM technology, the key research of keyword detection technology is to combine other pattern recognition methods to improve performance and method improvement and improve search recognition algorithm to increase speed. During this period, CMU's School of Computer Science, MIT's Lincoln Laboratory, and Dragon Systems reported their research results [9–11]. At the same time, when keyword detection technology based on the filler model template is widely recognized, it has been proposed that keyword detection is based on large vocabulary continuous speech recognition, mainly by inputting the results of continuous speech recognition after acoustic decoding, and then performing keyword detection [12, 13]. This research is mainly based on N-best structure. Researchers use N-best as the input for keyword detection. This text-based retrieval has high accuracy. However, N-best itself is a pruned form, and considering the choice of the optimal path, it may not guarantee the minimum word error rate. The end result is that the keyword may not be in N-best. There are also good application studies on keyword detection technology in China [14]. So far, research in the field of keyword detection is in the ascendant, new ideas and new algorithms in this area are constantly emerging, and keyword detection technology has also been well developed. However, how to find a more efficient and accurate detection method has always been the target of keyword detection, and we need to continue to move forward on this target.

Speech matching is an application in the field of speech recognition. Speech recognition is to convert speech recognition into text, and speech matching is to compare two speeches to determine whether they have the same meaning. It can be seen that speech recognition and speech matching are very similar. Speech matching methods are mainly divided into two types: one is to use the speech recognition method to convert two speeches into text and then use the string matching method to compare the strings; the other is to use the acoustic model to convert the speech signal into a vectorized feature matrix for matching. In the early 2000s, speech recognition was still dominated by traditional methods, such as hidden Markov models and feedforward artificial neural networks [15]. Today, many aspects of speech recognition have been replaced by a type of deep learning called Long Short-Term Memory Model (LSTM) [16]. Around 2007, the LSTM trained by Connection Timing Classification (CTC) began to outperform traditional speech recognition in some applications. In 2015, Google Speech Recognition used CTC-trained LSTM to play a 49% leap in performance, and this technology can now be used by all smartphone users through Google Speech. In the long history of speech recognition, shallow and deep artificial neural networks are exploring applications in speech recognition, but the performance of these methods has not exceeded the Gaussian mixture model-hidden Markov model (GMM-HMM) technology. At that time, the artificial neural network had many problems, including the gradient disappearance problem, which caused its bottleneck in speech recognition and matching. It was not until 2010 that various end-to-end neural network models became popular and became competitive models, which reduced the workload of feature engineering. In the model based on speech recognition, after the speech is converted into text, the similarity matching algorithm of the string is often used. One of the representative algorithms is Edit Distance. Editing distance is a way to quantify the mutual relationship between two different character strings (such as words) by calculating the minimum operand required to convert one character string to another [17]. Editing distance finds applications in natural language processing, where automatic spelling correction can determine candidate corrections for misspelled words by selecting a dictionary with a lower distance from related words from the dictionary. Speech signals belong to time series. In the field of time series analysis, dynamic time warping (DTW) is one of the algorithms used to measure the similarity between two time series, which may change in speed [18]. For example, even if one person walks faster than another person, or there are acceleration and deceleration during observation, DTW can also detect the similarity of walking. The DTW algorithm has been applied to various time series, such as video, audio, and graphic data. To be more precise, the DTW algorithm can be used to analyze any data that can be converted into a linear sequence. For example, its widely known application in speech recognition is to cope with different speaking speeds.

3. Feature Extraction

The calculation of MFCC considers the auditory characteristics of the human ear, and there are no assumptions, so this parameter has good recognition performance and ability to resist noise [19].

The calculation process of MFCC is shown in Figure 1.

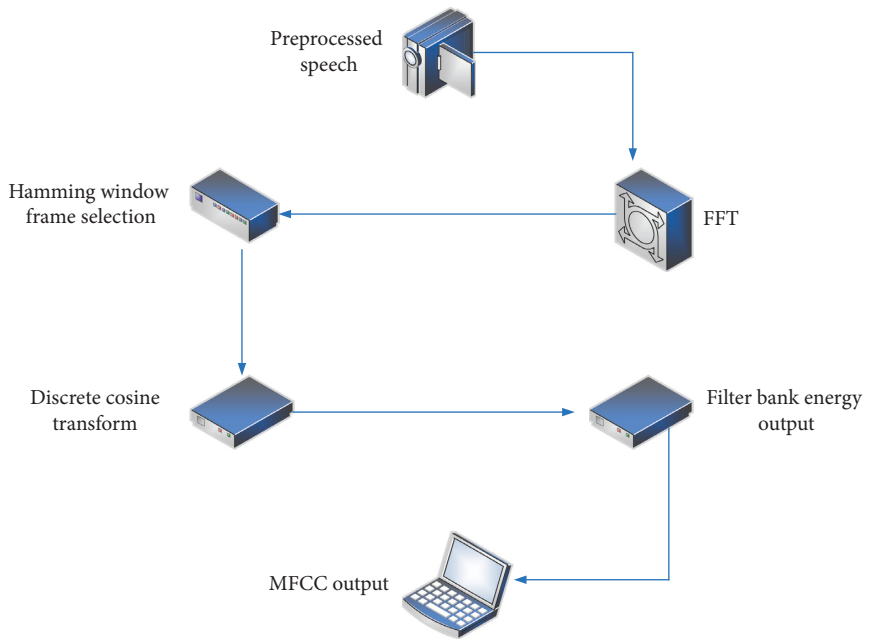

Figure 1: MFCC solution process.

(1) The speech signal becomes a short-time signal after windowing. Then, FFT is used to convert these time-domain signals $\theta(M_k)$ into frequency-domain signals $X(m)$, and the short-term energy spectrum $P(f)$ is calculated from the frequency-domain signals.

(2) $P(f)$ is converted into $P(M)$ on the Meier coordinate by the frequency spectrum on the frequency axis, and this conversion can be completed by the following formula [20]:

$$F_{\text{mel}} = 3322.23\lg(1 + 0.001) f_{Hz}. \tag{1}$$

(3) In the Meyer frequency domain, a triangular bandpass filter is added to the Meyer coordinates to obtain the filter bank $H_m(k)$, and the energy spectrum $P(M)$ on the Meyer coordinates is calculated through the output of this filter bank:

$$\theta(M_k) = \text{In}\left[\sum_{k=1}^{k} |X(k)|^2 H_m(k)\right] k = 1, 2, \ldots, k. \tag{2}$$

Among them, K represents the total number of filters.

(4) $\theta(M_k)$ represents the output energy of the k-th filter. Then, the Mel frequency spectrum $C_{\text{mel}}(n)$ can be obtained by the modified inverse discrete cosine transform:

$$C_{\text{mel}}(n) = \sum_{k=1}^{k} \theta(M_k)\cos\left(n - (k - 0.5)\frac{\pi}{K}\right), \tag{3}$$

$$n = 1, 2, \ldots, p.$$

In the above formula, p is the order of the MFCC parameter.

Figure 2 shows the structure of a typical HMM model $\lambda = \{A, B, \pi\}$; among them, A represents the state transition probability matrix, B is the probability density function vector, π is the probability matrix, and B is the probability density function vector. We can think of HMM as a finite automaton in the discrete time domain: at each discrete time t, it can only be in one of the finite states. The HMM in Figure 2 is a first-order model of state transition from left to right. There are six states: the first state is the entrance state, and the sixth state is the exit state. At any moment, the state of the automaton is only related to the state at the previous moment, and the state transition probability is a quantity independent of time.

In the process of speech recognition, the training of language models is also important. The training of language models is obtained through statistics and analysis of a large amount of text data. There are about 60,000 commonly used vocabularies. According to these vocabularies, language model training is performed, and the training model is applied in recognition. The N-Gram model is generally used for description, as shown in the following formula:

$$p(W_{n+1}|W_1^n) = P(W_{n+1}|W_{n-N+1}^n). \tag{4}$$

The idea of the N-Gram model is that the probability of the occurrence of the word $N + 1$ is only related to the most recent N words and has nothing to do with other words, that is, the hidden Markov model of order N. The training of language models may have the problem of sparse data, which is similar to the training of acoustic models. Therefore, in order to improve the promotion ability of the model, various smoothing algorithms are generally used to estimate the parameters of the language model.

The purpose of continuous speech recognition is to search for an optimal sequence in the grid formed by all possible subword sequences. For the keyword detection

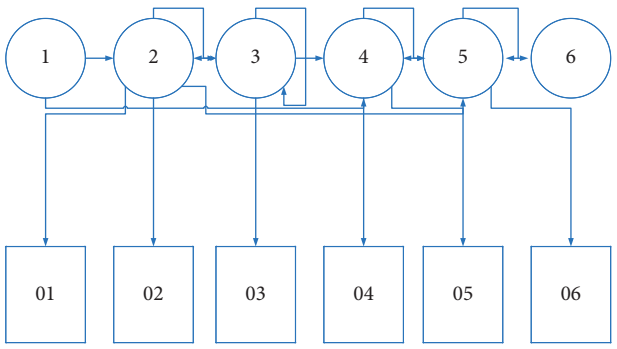

Figure 2: Topological structure of a typical HMM model.

system, it only cares whether the keywords appear in this sequence. However, for HMM, the most commonly used decoding algorithm is the Viterbi algorithm. The Viterbi algorithm solves the problem of how to determine an optimal state sequence $Q = q_1^* q_2^* \dots q_T^*$ when an observation sequence $O = o_1 o_2 \dots o_T$ and a model $\lambda = \{A, B, \pi\}$ are set.

The speech stream in the keyword recognition system contains keywords and nonkeywords, and the detection results are also divided into keywords and nonkeywords, so it will have four combined results, as shown in Figure 3.

Recognition rate is defined as the percentage of the number of keywords correctly recognized by the system and the total number of keywords actually appearing in the speech. It is also commonly referred to as the recall rate, and its definition is shown in formula (5). The definitions of missed recognition rate and recognition rate are just opposite, and their definitions are shown in formula (6).

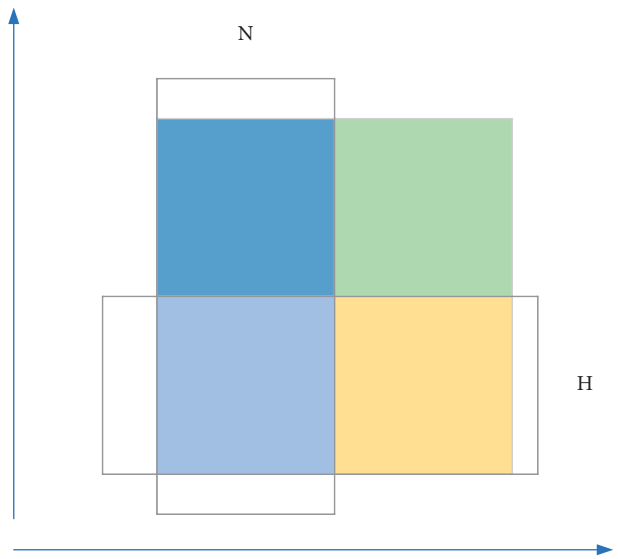

Figure 3: Relationship between detection results and reference results.

$$\text{Detection Rate} = \frac{\text{Number of key words detected by the system}}{\text{The actual number of voices}} \times 100\%, \tag{5}$$

$$\text{Missingrate} = \frac{\text{Number of key words not detected by the system}}{\text{The actual number of voices}} \times 100\%. \tag{6}$$

In practical applications, some keyword detection systems sometimes take a long time to work, and the number of real keyword occurrences is difficult to determine. At this time, we can define an evaluation standard called false alarm rate, which is defined as in the following formula:

$$\text{False alarm rate} = \frac{\text{Identify wrong result key words}}{\text{Hours of voice} \times \text{total number of detected key words} \times C} \times 100\%. \tag{7}$$

Among them, the total number of detected keywords refers to the total number of keywords in a given keyword table rather than the number of times the keywords appear. C is a constant, and its function is to make the false alarm rate and the rejection rate on the same scale. Since the recognition rate of keywords will be different under different false alarm rates, looking at the recognition rate under a certain false alarm rate cannot objectively measure the performance of the system. The current commonly used measurement method is the receiver operating characteristic curve, denoted as ROC curve. Figure 4 shows an example of a ROC curve. The ROC curve can intuitively and comprehensively reflect the performance of the keyword recognition system.

In practical applications, the ROC curve is only effective for measuring the performance of the system when the false alarm rate is less than 10. Therefore, in practical systems, we rarely choose point operation points with a false alarm rate greater than 10.

4. Confidence Research-Related Technology

We can define the confidence as a function $C(X)$, where X is the element in the event space $\{X_1, X_2, \dots, X_k, \dots\}$. If the

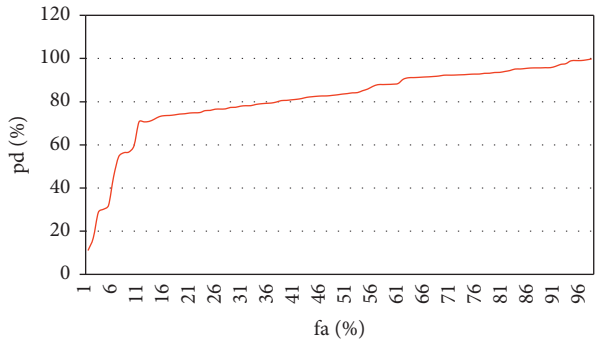

Figure 4: ROC curve.

probability of event X_1 is higher than the probability of event X_2, then $C(X)$ satisfies $C(X_1) > C(X_2)$. In speech recognition, if the reference model of speech is W and the speech signal that can be sensed is $O = \{O^1, O^2, \ldots, O^n, \ldots, O^N\}$, then the confidence $C(O|W)$ of O relative to W represents the degree of confidence that the speech O is generated by the speech model W.

In speech recognition, for the confidence $C(O|W)$ of the speech hypothesis result O obtained from a certain W in the speech model set, the output result O of the speech recognizer has two states H_0 and H_1 for the hypothesis H where the reference speech model W exists.

Null Hypothesis H_0: the hypothesis result O is generated by the speech model W; that is, the recognition result is correct.

Alternative Hypothesis H_1: the hypothesis result is generated by models other than the speech model W; that is, the recognition result is wrong.

To understand confidence from the perspective of hypothesis testing is to divide the test statistics into two different domains, the acceptance domain and the rejection domain, and then establish a critical point. When the test statistics fall on the side of the critical point, the hypothesis is rejected; otherwise, the hypothesis is accepted.

We can also understand confidence from the perspective of pattern recognition. If it is assumed that the speech model W is category 1, then all models $\overline{W}$ except W are category 2. The problem of judging whether the speech is generated by the speech model W becomes the recognition problem of judging whether the speech O belongs to class 1 or class 2. We assume that the identification function of the problem is $D(O)$ and that it satisfies the following:

(1) If $D(O) \geq 0$, then $O \in W$,

(2) If $D(O) < 0$, then $O \in \overline{W}$.

The recognition function at this time is equivalent to the confidence, that is, $C(O|W) = D(O)$.

Many speech recognition algorithms regard speech recognition as a pattern classification problem and obtain the word sequence $\widehat{W}$ with the highest similarity through the maximum posterior probability. In other words, $\widehat{W}$ is the sequence that maximizes the posterior probability $p(W|X)$ under the given conditions of the observation sequence X, which can be described by the following formula:

$$\begin{aligned}\widehat{W} &= \arg\max_{w\in} \sum p(W|X) \\ &= \arg\max_{w\in} \sum \frac{p(W|X)\cdot p(W)}{p(X)} \\ &= \arg\max_{w\in} \sum p(X|W)\cdot p(X).\end{aligned} \tag{8}$$

$\sum$ represents all possible sequence sets, $p(W)$ is the language model probability score of the hypothetical word sequence W, $p(X)$ is the probability value of the observation sequence X, and $p(X|W)$ represents the probability that the observation sequence of the recognition result sequence W is X.

Based on the calculation method of the posterior probability of the word graph, the speech recognizer generates a word graph χ for each speech segment X. The word graph is a directed acyclic structure with weights, which is composed of nodes and arcs. Among them, arc $[w]_s^e$ means that the start node is s, the end node is e, the hypothesis of a word is w, and the weight of the arc is recorded as $B(w)_s^e$. After that, we can record the complete hypothetical path of a speech segment X as $C = \left\{[w]_{s_1}^{e_1}, \ldots, [w]_{s_n}^{e_n}\right\}$. Then, the probability of this complete path is

$$p(C|\chi) = \prod_{i=1}^{n} B(w_i)_{s_i}^{e_i} \cdot p(w_i|h_i). \tag{9}$$

h_i represents the hypothesis before the word w_i on this path, and $p(w_i|h_i)$ is the corresponding N-Gram language model score. Therefore, the posterior probability $p(a|\chi)$ of arc $a = [w]_s^e$ is

$$p(a|\chi) = \frac{\sum_{C\in\chi, a^C} p(C|\chi)}{\sum_{C\in\chi} p(C|\chi).} \tag{10}$$

It is the ratio of the number of all complete paths including arc a to the number of all paths.

The purpose of finding these features is to distinguish between correct and wrong results. We can extract some features of the recognition results according to the design methods of the two types of classifiers; that is, we can use a large number of feature samples of wrong recognition results to train model M_e, and a large number of feature samples of positive recognition results to train model M_c. Moreover, for a recognized candidate result, the feature Y is selected first, and then the "distance" of this feature to M_c and M_e is calculated, which are defined as $d_c(Y)$ and $d_e(Y)$, respectively. Then, the formula for calculating the distance difference is

$$d(Y) = d_c(Y) - d_e(Y). \tag{11}$$

The result $d(Y)$ calculated by the above formula can be used as a confidence measure of the recognition result; that is, the higher the score of the model trained on the correct result and the lower the score of the model trained on the

wrong recognition result, the higher the confidence and the higher the probability that the result is correct.

The likelihood of the "positive model" is the model likelihood when it is assumed that the recognition result is correct. This value can be obtained during the identification process. The likelihood of "inverse model" means that the recognition result is wrong and its acoustic features are substituted into the inverse model, and then the likelihood is calculated recently.

H_0: when W is correct, it is brought into the "positive model" to get $p(X|W)$.

H_1: when W is wrong, it is brought into the "antimodel" to get $p(X|\widetilde{W})$.

$$LR = \frac{p(X|W)}{p(X|\widetilde{W})}. \tag{12}$$

In the above formula, $\widetilde{W}$ is a model obtained from non-W samples, that is, the abovementioned full-background model, inverse model, or competition model.

In the keyword detection system, the false alarm rates are defined as follows:

$$\begin{aligned} FR &= \frac{FR}{\text{Match}} \times 100\%, \\ FA &= \frac{FA}{KW \cdot HR \cdot M} \times 100\%. \end{aligned} \tag{13}$$

In the above formula, KW represents the size of the keyword vocabulary, HR represents the number of hours of speech to be detected, and M represents the average maximum number of false positives per keyword per hour.

In continuous speech recognition systems, the false alarm rates are usually defined as follows:

$$\begin{aligned} FR &= \frac{FR}{\text{Match}} \times 100\%, \\ FA &= \frac{FA}{\text{Sub} + \text{Ins}} \times 100\%, \\ CER &= \frac{FR + FA}{\text{Num}} \times 100\%. \end{aligned} \tag{14}$$

CER represents the confidence error rate, among them, and Num is the number of words in the recognition result of the system. When the threshold of confidence is constantly changing, FR and FA will also change. At this time, a balance must be taken. In general, the CER of the sum of FR and FA is used as the error rate of confidence.

5. Some Problems That SVM Needs to Solve When Solving Confidence

For different confidence source scores, different methods can be used to solve. For calculating the confidence of the larger unit as a whole from a small confidence source, the main methods are average methods, such as harmonic average, arithmetic average, and geometric average:

$$\begin{aligned} CMh(w) &= \frac{N}{\sum_{i=1}^{N} 1/CM_i}, \\ CMa(w) &= \frac{1}{N}\left[\sum_{i=1}^{N} CM_i\right], \\ CMg(w) &= \exp\left(\frac{1}{N}\left[\sum_{i=1}^{N} \log(CM_i)\right]\right). \end{aligned} \tag{15}$$

In the above formula, N is the total number of subsequences in the word, CM_i is the confidence score of the i-th subunit, and CMh, CMa, CMg are the combined confidence values obtained by harmonic average, arithmetic average, and geometric average, respectively. The results obtained with different confidence methods are also different. Therefore, we can use these confidence values to determine the choice of recognition results.

For a given hypothetical hit word, different units and different sources of confidence information can get a vector of confidence scores:

$$v_n(w) = (CM_1, CM_2, \ldots, CM_N). \tag{16}$$

Among them, CM_i is the score of a certain confidence information source (phonon), and after normalization over time, it becomes

$$v_n(w) = \left(\frac{CM_1}{t_1}, \frac{CM_2}{t_2}, \ldots, \frac{CM_N}{t_N}\right). \tag{17}$$

Among them, C_i is the confidence of the i-th phoneme, and t_i is the duration of the i-th phoneme.

If the phoneme's confidence vector is used as the input vector of an SVM, then the output of the SVM can be a certain representation of the word's confidence after a certain conversion. Since confidence decision is a typical binary classification problem, and SVM is a typical binary classifier, and SVM performs well in the field of pattern recognition, SVM can be used to synthesize confidence.

At the same time, the phoneme confidence is sorted to form a vector v_s, $v_s(w) = F_s(v_n(w))$, which can reflect the phonetic confidence distribution of a word as a whole. Since the vector v_s is related to the number of phonemes in the word, the dimension of v_s will change. In this study, words are classified based on the number of phonemes, and words with the same number of phonemes are used as a classification and share an SVM classifier. The resulting confidence score is C_{SVM}.

$$C_{SVM} = \sum_{S_{L(W)}} \alpha^0_{i,L(w)} y_{i,L(w)} K\left(x_{i,L(w)}, v_s(w)\right) + b^0_{L(w)}. \tag{18}$$

Among them, $L(w)$ represents the number of phonemes of the word m, S is the set of support vectors, $\alpha^0_{i,L(w)}$, $y_{i,L(w)}$, $x_{i,L(w)}$, and $b^0_{L(w)}$ are the parameters of the SVM classifier, which are all obtained by training.

There is a problem when using SVM to solve the confidence; that is, the relationship between the distance between the test pattern and the classification interval and the

posterior probability is not very clear. This problem needs to be solved emphatically. In the identification problem, it is necessary to get a clear probability output. However, the output result of using SVM is often binary; that is, it is either 0 or 1, which does not reflect the probability that the test sample belongs to a certain category, that is, the probability of belonging to a certain category. If the output can reflect its probability characteristics, that is, the size of the probability belonging to a certain class, then it will bring aspects to the decision of the confidence, so as to be better applied to the actual system and to improve the system's overall performance.

According to the knowledge of SVM introduced above, the simple form of SVM output is

$$y = \text{Sign}(f(x)). \tag{19}$$

In the above formula, if $f(x)$ is greater than 0, the result y is a positive sample. However, if $f(x)$ is less than 0, then the result y is a negative sample. Of course, the sample needs to be normalized during the calculation; that is, the sample point closest to the classification surface meets $\|f(x)\| = 1$. The sample points above the classification surface need to satisfy $m\ f(x) = 0$, and the remaining sample points need to satisfy $f(x) = \pm d\|w\|$. d in the formula represents the distance between the sample point x and the classification surface. The positive and negative signs indicate that the sample point is on both sides of the classification surface. Finally, we can use the Sigmoid function to probabilistically output the SVM results. The specific conversion method is as follows:

$$\begin{aligned} p(y = 1|f(x)) &= \frac{1}{1 + e^{\alpha f(x) + \beta}}, \\ p(y = -1|f(x)) &= 1 - p(y = 1|f(x)). \end{aligned} \tag{20}$$

In the above formula, the parameters α and β can control the shape of the Sigmoid function, and $f(x)$ is the output value of the sample x obtained in the support vector machine. Using the above formula, the results obtained by SVM can be output probabilistically. Of course, sometimes the training data will cause the distance estimation to deviate seriously, so when evaluating the parameters of the Sigmoid function, a cross-validation set must be added. The amount of data in this set is determined by the amount of training data. After solving the above probabilistic output problem, we can use SVM to solve the confidence problem, and then use the confidence to confirm the results to make the recognition results more effective.

6. Model Performance Analysis

In order to improve the performance of this research model, the research model is analyzed through control experiments. The Japanese speech of a certain streaming media is the research object, and 60 sets of speech data characteristics are counted as basic data, and the speech keyword detection is carried out using this research model. In order to improve the performance comparison effect, this study compares the traditional neural network algorithm with the algorithm proposed in this study. The algorithm proposed in this study is named PSVM, and the neural network algorithm is named NN. First, the comparison of algorithm keyword detection accuracy is performed, and the results are shown in Table 1 and Figure 5.

Table 1: Statistical table of the accuracy of keyword detection (%).

	NN	PSVM		NN	PSVM
1	48	92	31	51	95
2	31	87	32	37	89
3	43	87	33	48	95
4	35	88	34	38	90
5	36	85	35	45	91
6	32	89	36	35	88
7	36	87	37	35	89
8	36	94	38	53	93
9	37	88	39	55	95
10	48	92	40	56	95
11	41	89	41	38	92
12	54	94	42	44	89
13	36	89	43	38	93
14	46	93	44	38	88
15	38	88	45	49	93
16	55	89	46	36	95
17	53	93	47	46	86
18	41	96	48	53	88
19	45	91	49	43	92
20	45	95	50	33	93
21	39	90	51	33	89
22	40	92	52	52	93
23	42	89	53	54	94
24	32	86	54	50	90
25	38	92	55	35	89
26	49	96	56	44	86
27	53	92	57	39	94
28	43	91	58	34	95
29	45	89	59	53	95
30	55	89	60	45	86

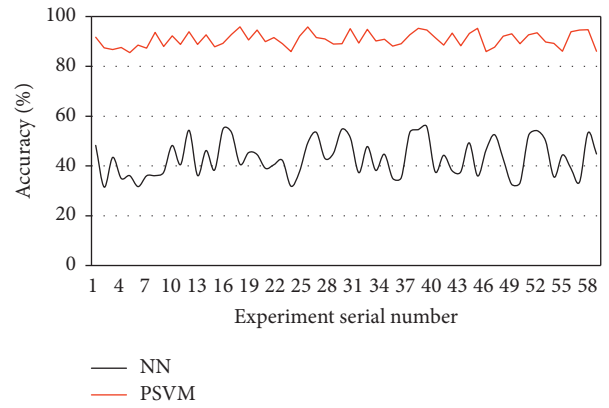

Figure 5: Statistical diagram of the accuracy of keyword detection (%).

Afterwards, a comparative analysis of keyword detection speed is performed, and the results are shown in Table 2 and Figure 6.

Table 2: Statistical table of the speed of keyword detection (ms).

	NN	PSVM		NN	PSVM
1	243	78	31	248	83
2	201	57	32	199	70
3	223	78	33	217	76
4	196	78	34	219	85
5	208	60	35	195	66
6	209	60	36	207	69
7	257	51	37	259	79
8	209	55	38	210	80
9	248	76	39	219	81
10	222	82	40	205	73
11	255	56	41	206	71
12	247	52	42	198	79
13	224	84	43	227	73
14	236	58	44	242	67
15	211	87	45	214	82
16	256	81	46	245	81
17	218	73	47	242	78
18	205	51	48	222	76
19	258	56	49	230	72
20	221	75	50	195	61
21	195	53	51	226	78
22	253	56	52	226	75
23	227	84	53	234	87
24	248	74	54	250	52
25	211	50	55	191	59
26	201	52	56	255	69
27	244	56	57	211	54
28	259	76	58	214	81
29	246	83	59	239	81
30	253	80	60	201	79

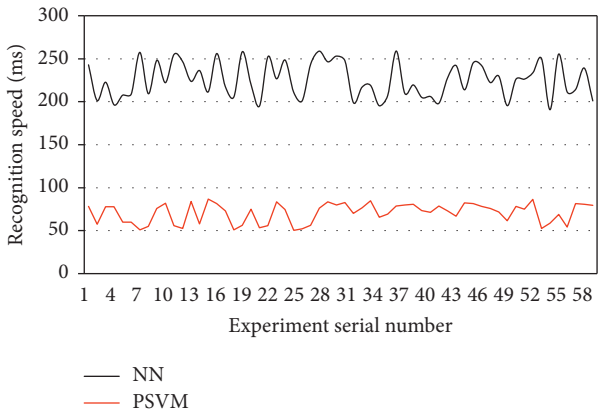

Figure 6: Statistical diagram of the speed of keyword detection (ms).

It can be seen from Tables 1 and 2 and Figures 5 and 6 that the recognition accuracy of the algorithm proposed in this study is distributed between 85% and 100%, which meets the actual detection accuracy requirements and far exceeds traditional algorithms. In terms of recognition speed, the algorithm proposed in this study controls the recognition time within 100 ms, while the recognition time of traditional algorithms exceeds 200 ms. This shows that the algorithm proposed in this study has a certain effect regardless of the recognition speed and recognition accuracy.

7. Conclusion

In this study, a support vector machine-based streaming Japanese speech keyword detection model is constructed, and SVM is used to solve the confidence. Moreover, some problems that need to be solved by this method are discussed: SVM probabilistic output problem, SVM confidence synthesis problem, etc. Compared with the effect of arithmetic averaging on the confidence synthesis, the confidence obtained by SVM can obtain a higher recognition rate under the same rejection rate and improve the overall performance of the system. Moreover, this study proposes a new retrieval algorithm. The algorithm is a detection algorithm based on word graph structure. The algorithm uses the Trie tree to store the keywords that need to be detected and then retrieves the keywords on the Trie tree through the nodes in the word graph, which can be applied to some occasions with specific requirements. In addition, compared with continuous speech recognition, it also has the characteristics of fast speed, high detection rate, and strong practicality, so it has broad application prospects and good research value. In order to improve the overall performance of the system, it is necessary to remove some of the incorrect recognition results or select an appropriate operating point for recognition.

Data Availability

The datasets used or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author declares that he has no conflicts of interest.

Acknowledgments

This paper was supported by 2019 Ministry of Education Industry-University Cooperation Collaborative Education Project (201902318007) and Educational Science Research Project of Qiqihar University (GJSKYB202020).

References

[1] R. Rhodes, "Aging effects on voice features used in forensic speaker comparison," *International Journal of Speech Language and the Law*, vol. 24, no. 2, pp. 177–199, 2017.

[2] Q. K. Ngoc and H. T. Duong, "A review of audio features and statistical models exploited for voice pattern design," *Computer Science*, vol. 03, no. 2, pp. 36–39, 2015.

[3] M. Sarria-Paja, M. Senoussaoui, and T. H. Falk, "The effects of whispered speech on state-of-the-art voice based biometrics systems," *Canadian Conference on Electrical and Computer Engineering*, vol. 2015, pp. 1254–1259, 2015.

[4] A. Leeman, H. Mixdorff, M. O'Reilly, M. J. Kolly, and V. Dellwo, "Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison," *International Journal of Speech Language and the Law*, vol. 21, no. 2, pp. 343–370, 2015.

[5] A. K. Hill, R. A. Cárdenas, J. R. Wheatley et al., "Are there vocal cues to human developmental stability? Relationships between facial fluctuating asymmetry and voice attractiveness," *Evolution and Human Behavior*, vol. 38, no. 2, pp. 249–258, 2017.

[6] M. Woźniak and D. Połap, "Voice recognition through the use of Gabor transform and heuristic algorithm," *Nephron Clinical Practice*, vol. 63, no. 2, pp. 159–164, 2017.

[7] T. Haderlein, M. Döllinger, V. Matoušek, and E. Nöth, "Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples," *Logopedics Phoniatrics Vocology*, vol. 41, no. 3, pp. 106–116, 2015.

[8] S. S. Nidhyananthan, K. Muthugeetha, and V. Vallimayil, "Human recognition using voice print in LabVIEW," *International Journal of Applied Engineering Research*, vol. 13, no. 10, pp. 8126–8130, 2018.

[9] C. T. Herbst, S. Hertegard, and D. Zangger-Borch, "Freddie Mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics," *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 1–10, 2016.

[10] J. Al-Tamimi, "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations," *Laboratory Phonology*, vol. 8, no. 1, pp. 1–40, 2017.

[11] P. Laukka, H. A. Elfenbein, N. S. Thingujam et al., "The expression and recognition of emotions in the voice across five nations: a lens model analysis based on acoustic features," *Journal of Personality and Social Psychology*, vol. 111, no. 5, pp. 686–705, 2016.

[12] F. Mousavizadeh, K. Maghooli, E. Fatemizadeh, and M. S. Moin, "Liveness detection in face identification systems: using zernike moments and fresnel transformation of facial images," *Indian Journal of Science and Technology*, vol. 8, no. 8, p. 523, 2015.

[13] S. Orlandi, C. A. R. Garcia, and A. Bandini, "Application of pattern recognition techniques to the classification of full-term and preterm infant cry," *Journal of Voice*, vol. 30, no. 6, pp. 656–663, 2015.

[14] C. C. Hsu, K. M. Cheong, T. S. Chi, and Y. Tsao, "Robust voice activity detection algorithm based on feature of frequency modulation of harmonics and its DSP implementation," *IEICE - Transactions on Info and Systems*, vol. E98.D, no. 10, pp. 1808–1817, 2015.

[15] P. H. Kumar and M. N. Mohanty, "Efficient feature extraction for fear state analysis from human voice," *Indian Journal of Science & Technology*, vol. 9, no. 38, pp. 1–11, 2016.

[16] F. L. Malallah, K. N. Y. M. G. Saeed, S. D. Abdulameer et al., "Vision-based control by hand-directional gestures converting to voice," *International Journal of Scientific & Technology Research*, vol. 7, no. 7, pp. 185–190, 2018.

[17] S. Sleeper, "Contact effects on voice-onset time in Patagonian Welsh," *Journal of the Acoustical Society of America*, vol. 140, no. 4, p. 3111, 2016.

[18] G. Mohan, K. Hamilton, A. Grasberger, A. C. Lammert, and J. Waterman, "Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures," *Journal of the Acoustical Society of America*, vol. 137, no. 4, p. 2302, 2015.

[19] T. G. Kang and N. S. Kim, "DNN-based voice activity detection with multi-task learning," *IEICE - Transactions on Info and Systems*, vol. E99.D, pp. 550–553, 2016.

[20] HaNa Choi, S. W. Byun, and S. P. Lee, "Discriminative feature vector selection for emotion classification based on speech," *The Transactions of the Korean Institute of Electrical Engineers*, vol. 64, no. 9, pp. 1363–1368, 2015.