Confidence-based Utterance Selection for a Recognizer-free Spoken Dialogue System

Akinori Ito
aito.spcom@tohoku.ac.jp
Graduate School of Engineering, Tohoku University
Sendai, Miyagi, Japan

ABSTRACT

This work aims to develop a spoken dialogue system without a speech recognizer, which is used for a CALL system of endangered languages for language revitalization. A problem in realizing such a system is the accuracy of choosing a most-matched example utterance because of the difference in voice characteristics. Thus, this paper proposes a method to choose the selection candidate among multiple candidates selected by different speech features. The selection is based on the confidence measure calculated from the distance values between the input and the database utterances. We conducted an experiment to choose a candidate among those by MFCC and PPG and obtained a 2.6-point accuracy improvement.

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Information systems** → **Speech / audio search**.

KEYWORDS

Spoken dialogue system, Language revitalization, Computer-assisted language learning, Confidence measure, Mel-frequency cepstral coefficients, Phonetic posteriorgram

ACM Reference Format:
Akinori Ito. 2023. Confidence-based Utterance Selection for a Recognizer-free Spoken Dialogue System. In *2023 15th International Conference on Machine Learning and Computing (ICMLC 2023), February 17–20, 2023, Zhuhai, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3587716.3587796

1 INTRODUCTION

Many languages in the world are at risk of extinction. Campbell and Below listed more than 3,000 languages as endangered [1], which is more than half of the existing languages. Not only documenting the endangered languages [2] but there are also attempts to increase the number of speakers of that language, which is called "language revitalization" [3].

Education is an essential part of language revitalization [4]. Since it is difficult to find language teachers for an endangered language, the computer-assisted language learning (CALL) system is expected to play an important role in language revitalization [5].

ICMLC 2023, February 17–20, 2023, Zhuhai, China

ACM ISBN 978-1-4503-9841-1/23/02.
https://doi.org/10.1145/3587716.3587796

There have been several CALL systems for endangered languages, such as Hawaiian [6], Welsh [7], Cornish, Manx, and Ainu [8]. There are four skills to learn: reading, writing, listening, and speaking. However, the current CALL systems for those languages provide only basic functionalities. For example, Leokï, the system for Hawaiian [6], provides basic vocabulary training and reading materials. Although some CALL systems of major languages provide functionalities of speaking training, such as pronunciation training [9,10] and conversation training [11], it is rare for a CALL system of an endangered language to provide such functionalities. One reason these functions are difficult to provide is that these functions require an accurate speech recognizer and high-quality speech synthesizer. Since the development of speech recognizer and speech synthesizer needs a large amount of speech data, it is difficult to develop such a system for low-resource languages.

We have developed a method to realize a spoken dialogue system without a speech recognizer to develop a spoken dialogue system for low-resource or zero-resource languages [12]. This system directly matches the input speech signal to the utterances in the database, which is a language-independent process. Therefore, the system can be used for any language, even if the language has no speech recognizer.

One problem in this system is that the recognition accuracy is still not high. Therefore, in this paper, we describe an idea to increase the accuracy of utterance selection by combining multiple speech features.

2 RELATED WORK

This work is based on our previous work on developing a spoken dialogue system without a speech recognizer [12]. Before describing the proposed method, we explain the existing dialogue systems.

Figure 1 shows a basic structure of an ordinary spoken dialogue system that uses both a speech recognizer and a synthesizer. First, the automatic speech recognizer transcribes the input speech. Then the dialogue manager calculates the response to the input from the transcription. Finally, the generated response sentence is converted to a speech signal using the speech synthesizer, and the system outputs the response speech to the user. The retrieval-based dialogue system is a kind of dialogue system that has an example-response database. This kind of system [13] finds the most similar example to the input utterance from the database. The response sentence associated with the selected example sentence is used as the response speech. The basic system measures the similarity between two sentences by superficial similarity; other systems calculate similarity by neural networks [14].

As stated above, the existing retrieval-based dialogue system uses the speech recognizer to transcribe the input speech, which

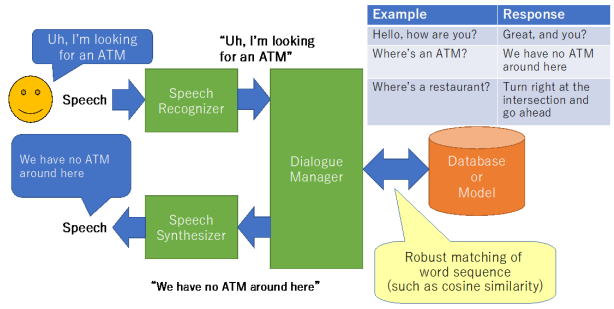

Figure 1: The basic structure of an ordinary spoken dialogue system.

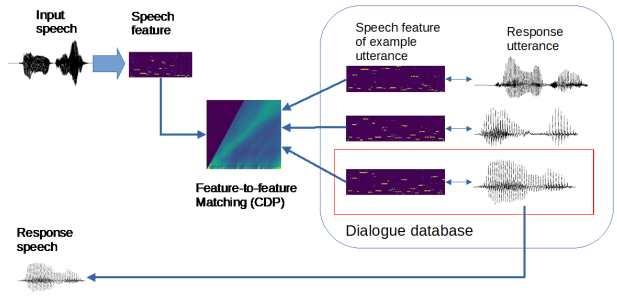

Figure 2: The framework of the proposed dialogue system

is difficult for minority languages with limited language resources. Therefore, our system employs direct speech-to-speech matching instead of transcribing the input speech [12]. Figure 2 shows the framework of the system. First, the input utterance is converted to the speech feature, such as the mel-frequency cepstral coefficients (MFCC) or the phonetic posteriorgram (PPG) [15]. The problem is that the MFCC depends on both the speaker and the content, while the PPG depends on both the language and the content. Since we apply this method to an endangered language, the feature should be independent of both the speaker and the language. To do that, we can combine the PPG of the multiple languages to reduce the language dependency of the feature [16].

The feature-to-feature matching is based on continuous DP (Dynamic Programming) matching (CDP) [17], which can detect occurrences of a short pattern from longer patterns. After matching the input feature and all features in the database, the database feature with the highest similarity (the smallest distance) is selected as the candidate, and the associated response utterance is played.

3 THE SPEAKER-DEPENDENCE OF THE FEATURE

As described above, the MFCC depends on both the speaker and the content. Therefore, when we compare two features of MFCC, the similarity between the two features is affected by both the similarity of the speech content and the speaker's voice. The speaker-independent features, such as PPG or the bottleneck feature [18], effectively remove the speaker dependency of the feature. However,

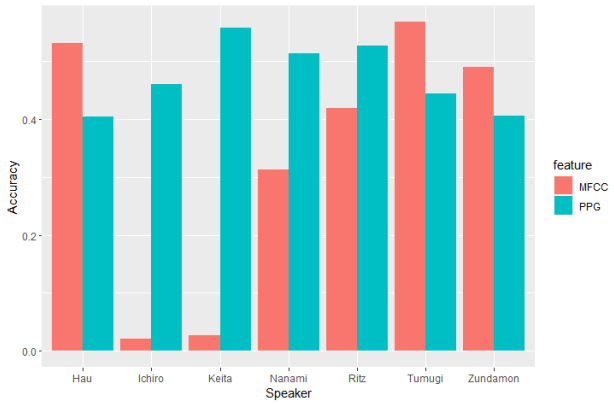

Figure 3: Accuracy for a female speaker's query.

because the accuracy of PPG estimation is not perfect, it deteriorates the accuracy when the characteristics of the two speakers are similar [12].

Figure 3 shows the accuracy of a female speaker's 482 queries to the database utterances by seven speakers [12] (the detailed experimental conditions will be described in Section 5). In this result, the two database speakers (Ichiro and Keita) are male, and the others are female. We can confirm that we could not recognize the utterance when the database speakers were male, and the feature was MFCC, whereas the PPG improved the accuracy. However, when the gender of the user and the database speakers were the same (female in this case), the PPG results were not necessarily better than the MFCC. For example, MFCC showed better results for three speakers (Hau, Tumugi, and Zundamon), while PPG was better than MFCC for the other female speakers. Therefore, we can improve the utterance selection accuracy if we know whether the speaker's characteristics of two utterances are similar or not.

One idea to do that is to extract speaker characteristics such as the i-vector [19,20] or d-vector [21] and measure the speaker similarity. However, this approach requires extra computation for speaker similarity calculation in addition to utterance matching. Therefore, we employed a different way, inspired by the classical confidence measure estimation for isolated word recognition [22].

4 ESTIMATION OF CONFIDENCE

The confidence measure of word recognition was used to reject the recognition result when the confidence value was low [22]. Using this concept, we compute the confidence of the matching result based on the distance between the input and database utterances using multiple features.

Let u be the user's input utterance, $u(1), \ldots, u(N)$ be the utterances in the database, $f(u)$ be the feature vector sequence of utterance u, and $D(f(u), f(v))$ be the distance between $f(u)$ and $f(v)$ calculated by the CDP. Let us denote

$$D(u, v|f) = D(f(u), f(v)). \tag{1}$$

Then we calculate

$$\mathbf{D}_K(u, f) = (D(u, u(i_1)|f),\ D(u, u(i_2)|f), \ldots, D(u, u(i_K)|f)) \tag{2}$$

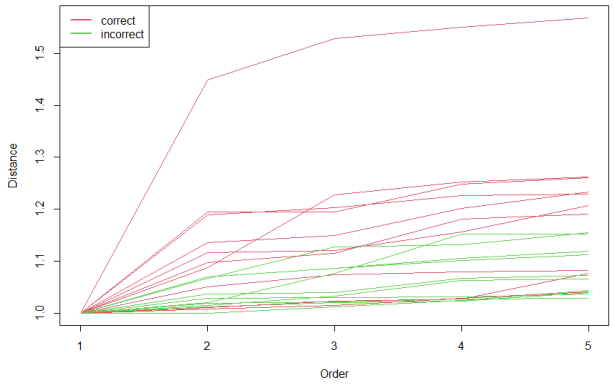

Figure 4: The sorted distances.

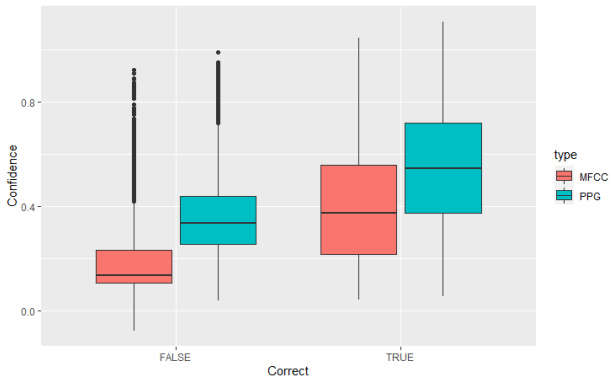

Figure 5: Distribution of the confidence.

where $i_1, \ldots, i_N$ are the indices so that the distance values are sorted in ascending order,

$$D\left(u, u\left(i_1\right)|f\right) \leq D\left(u, u\left(i_2\right)|f\right) \leq \ldots \leq D\left(u, u\left(i_K\right)|f\right)). \quad (3)$$

Therefore, $\mathbf{D}_K(u, f)$ is a list of K smallest distances between u and the database utterances using feature f.

Figure 4 shows examples of $\mathbf{D}_5(u, MFCC)$, normalized by the smallest distance. The green lines are data for incorrect results, and the red lines are for correct results. This example shows that the correct samples tend to have higher values, which suggests that we can classify the detection result as either "incorrect" or "correct." Thus, we calculate the confidence $C\left(\mathbf{D}_K(u, f)\right)$ so that it becomes 0 when $\mathbf{D}_K(u, f)$ is incorrect and 1 when correct. If we have multiple features $f_1, \ldots, f_F$, we can choose the best feature as

$$\widehat{f} = \underset{f \in \{f_1, \ldots, f_F\}}{\arg\max}\ C(\mathbf{D}_K(f)). \quad (4)$$

Then we take the detection result using $\widehat{f}$.

5 EXPERIMENT

5.1 Conditions

According to our previous work [12], we conducted a simulation experiment using 482 Japanese sentences for spoken dialogue systems [23,24]. Each sentence has an ID, and the system selects the response utterance using the ID. When two or more sentences share the same ID, these sentences share the same response utterance. For example, the two sentences "*Anata ni shitsumon ga arimasu*" (I have a question for you) and "*Anata ni kikitai koto ga arimasu*" (I want to ask you something) share the same ID.

We generated six kinds of speech signals using speech synthesizers: one male voice by Microsoft SAPI, one male and one female by Microsoft Azure Cognitive Service TTS, and five females by VOICEVOX[1]. Thus, we generate $482 \times 8 = 3856$ utterances in total.

When we conduct an utterance selection experiment, we pick two speakers, one is the user, and the other one is the database speaker. Then we choose one utterance u of the user and match the utterance to all the utterances $u(1), \ldots, u(482)$ of the database speaker. We excluded the same sentence to simulate the unknown input. Thus, the database utterance is selected as follows.

$$\widehat{k} = \arg \min_{k: u \neq u(k)} D(u, u(k)|f) \quad (5)$$

We regard the selection result to be correct when the IDs of u and $u(\widehat{k})$ are the same.

We used MFCC and PPG as the features. The PPG extractor was a 1-dimensional convolutional neural network (CNN) trained using the ASJ-JNAS database. See detail for [12].

Table 1: 2x2 table of the selection results by MFCC and PPG

	PPG incorrect	PPG correct
MFCC incorrect	12871	8140
MFCC correct	2007	3974

5.2 Confidence calculation

We calculated the confidence using $\mathbf{D}_5(f)$. We used XGBoost [25] as the classifier, implemented as an R library[2], with the default parameter setting ($\eta = 0.3$, maximum tree depth 6, no L1 regularization), and the number of epochs was 10. The experiment was based on two-fold cross-validation. Figure 5 shows the distribution of the confidence values. This result confirmed that the confidence values for correct candidates tended to be higher than those for incorrect candidates.

5.3 Sentence selection experiment

Based on the confidence in the previous section, we carried out an experiment to choose a sentence using both MFCC and PPG. Table 1 shows the summary of the selection results. For example, we had 8140 utterances correctly processed using the PPG, but the MFCC results were incorrect. From this table, if we could select the feature among MFCC and PPG optimally, the accuracy would become (2007+8140+3974)/26992=52.3%. This is the upper limit of the feature selection method.

When we had an input utterance u, we calculated both $\mathbf{D}_5(u, MFCC)$ and $\mathbf{D}_5(u, PPG)$. Then we calculated the confidence $C(\mathbf{D}_5(u, MFCC))$ and $C(\mathbf{D}_5(u, PPG))$ using XGBoost, and we took the candidate with higher confidence. Figure 6 shows the experimental results. This

[1]https://voicevox.hiroshiba.jp

[2]https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

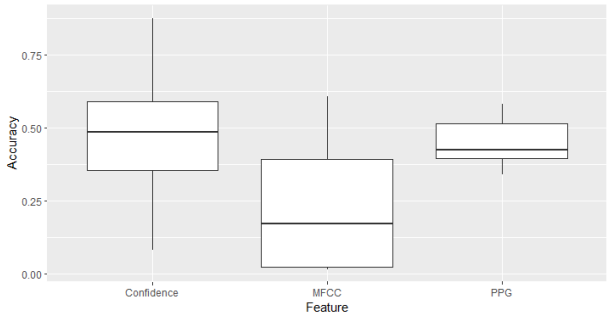

Figure 6: Distribution of accuracy by three features.

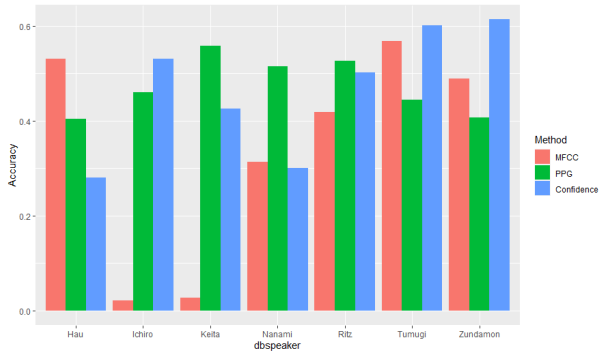

Figure 7: Accuracy for a female speaker's query (Effect of confidence-based method).

result shows that we could improve the selection result using confidence. The average accuracies were 22.2% for MFCC, 44.9% for PPG, and 47.5% for the confidence-based method. Although the confidence could improve the accuracy, Figure 6 also shows that the minimum accuracy by the confidence-based method was lower than that by PPG.

Figure 7 shows an example of the effect of the confidence-based utterance selection. The results of labels "MFCC" and "PPG" are the same as in Figure 3. These results suggest that the combination method did not work for several speakers (such as Hau, Keita, Nanami, and Ritz). We need further investigation to improve the accuracy of these speakers.

6 CONCLUSION

In this paper, we propose a method to improve the accuracy of the utterance selection for a recognizer-free spoken dialogue system. We used a list of inter-utterance distances as a feature to estimate the confidence of the user's input utterance. Then we compared the confidence values calculated by different features (MFCC and PPG in this paper) and took the utterance selection result with a higher confidence value. The experimental result revealed that the accuracy was improved from 44.9% (with PPG) to 47.5%.

In future work, we will investigate why the proposed utterance selection method did not improve the accuracy. Besides, we can combine other features, such as the bottleneck feature or PLP.

ACKNOWLEDGMENTS

Part of this work was supported by Yotta Informatics Project by MEXT, Japan, and JSPS KAKENHI JP19H05589.

REFERENCES

[1] Lyle Campbell and Anna Belew. 2018. *Cataloguing the world's endangered languages*, vol. 711. New York, USA: Routledge.
[2] Colette Grinevald. 2003. Speakers and documentation of endangered languages. *Language documentation and description*, 1, 52-72.
[3] Leanne Hinton. 2001. Language revitalization: An overview. In *Leanne Hinton and Ken Hale (eds.) The Green Book of Language Revitalization in Practice*, BRILL.
[4] Candace K. Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29(7), 1137-1151.
[5] Delaney Lothian, Gökçe Akçayir and Carrie D. Epp. 2019. Accommodating Indigenous People When Using Technology to Learn Their Ancestral Language. In *SLLL@ AIED* (pp. 16-22).
[6] Mark Warschauer, Keola Donaghy and Hale Kuamoʏo. 1997. Leokï: A powerful voice of Hawaiian language revitalization. *Computer Assisted Language Learning* 10(4), 349-361. https://doi.org/10.1080/0958822970100405
[7] Dylan V. Jones and Marilyn Martin-Jones. 2004. Bilingual education and language revitalization in Wales: Past achievements and current issues. In *James W. Tollefson and Amy B. M. Tsui (eds.) Medium of instruction policies: Which agenda? Whose agenda?* Lawrence Erlbaum Associates, Mahwah, New Jersey. 43-70.
[8] John C. Maher. 2014. Reversing Language Shift and Revitalization: Ainu and the Celtic Languages, *The Japanese Journal of Language in Society* 17(1), 20-35.
[9] Martha C. Pennington and Pamela Rogerson-Revell. 2019. Using technology for pronunciation teaching, learning, and assessment. *English Pronunciation Teaching and Research*. Palgrave Macmillan, London. 235-286.
[10] Jiang Fu, Yuya Chiba, Takashi Nose and Akinori Ito. 2020. Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, 116, 86-97.
[11] Serge Bibauw, Thomas François, Wim Van Den Noortgate and Piet Desmet. 2022. Dialogue systems for language learning: a meta-analysis. *Language Learning & Technology*, 26(1).
[12] Akinori Ito. 2022. Spoken dialogue system development without speech recognition towards language revitalization. In *Proc. International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, to appear.
[13] Ryuichi Nisimura, Akinobu Lee, Hiroshi Saruwatari and Kiyoshi Shikano. 2004. Public speech-oriented guidance system with adult and child discrimination capability. In *Proc. ICASSP*. vol. 1, pp. I–433.
[14] Chongyang Tao, Jiazhen Feng, Rui Yan, Wei Wu and Daxin Jiang. 2021. A Survey on Response Selection for Retrieval-based Dialogues. In *Proc. IJCAI*, pp. 4619-4626.
[15] Timothy J. Hazen, Wade Shen and Christpher White. 2009. Query-by-example spoken term detection using phonetic posteriorgram templates. In *Proc. Workshop on Automatic Speech Recognition & Understanding*. pp. 421–426.
[16] Satoru Mizuochi, Takashi Nose and Akinori Ito. 2022. Spoken term detection of zero-resource language using posteriorgram of multiple languages. *Interdisciplinary Information Sciences* 28(1), 1–13.
[17] Sei-ichi Nakagawa. 1989. Speaker-independent continuous-speech recognition by phoneme-based word spotting and time-synchronous context-free parsing. *Computer Speech & Language* 3(3), 277–299.
[18] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda and Ekaterina Egorova. 2012. The language-independent bottleneck features. In *2012 IEEE Spoken Language Technology Workshop (SLT)* , pp. 336-341.
[19] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788-798.
[20] Daniel Garcia-Romero and Carol Y. Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech*, 249-252.
[21] Ehsan Variani, Xin Lei, Eric McDermott, Ignacio Lopez Moreno and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *Proc. ICASSP*, pp. 4052-4056.
[22] E. Tsiporkova, F. Vanpoucke and H. Van hamme. 2000. Evaluation of various confidence-based strategies for isolated word rejection. In *Proc. ICASSP*, pp. 1819-1822.
[23] Yukiko Kageyama, Yuya Chiba, Takashi Nose and Akinori Ito. 2018. Analyses of example sentences collected by conversation for example-based non-task-oriented dialog system. *IAENG International Journal of Computer Science* 45(1), 285–293.
[24] Yukiko Kageyama, Yuya Chiba, Takashi Nose and Akinori Ito. 2018. Improving user impression in spoken dialog system with gradual speech form control. In *Proc. SIGdial*. pp. 235–240.
[25] Tianqi Chen and Carlos Guestrin. 2016. XBGBoost: A scalable tree boosting system. In *Proc. ACM SIGKDD*, pp. 785-794.