ACOUSTIC CLASSIFICATION AND SPEECH RECOGNITION HISTORIES FOR ADAPTABLE SPOKEN LANGUAGE DIALOGUE SYSTEMS

Sebastian Fleissner, Xiaoyue Liu & Alex Fang

Dialogue Systems Group, City University of Hong Kong, Hong Kong
sfleissn@cityu.edu.hk; xyliu0@cityu.edu.hk; acfang@cityu.edu.hk

ABSTRACT

This paper proposes speech recognition histories and acoustic classification as a means for facilitating adaptation to non-native speech in spoken language dialogue systems. Apart from providing a detailed description of acoustic classification and recognition histories, a study is presented that applies and evaluates the two concepts in a concrete dialogue system used by non-native English speakers.

Keywords: speech recognition, dialogue systems

1. INTRODUCTION

Spoken language dialogue systems rely on speech recognition and text-to-speech engines to interact with human users. A primary concern of dialogue systems is to provide users with a smooth experience by employing various techniques, such as grounding [1] and dialogue (flow) management techniques, to achieve natural dialogues. However, despite such techniques, speech recognition errors can cause a dialogue to stall and require the user to repeat herself several times. This is especially a problem when users with a dialogue system that does not use their native language.

This paper proposes two techniques that are designed to help dialogue systems to learn from recognition mistakes and adapt to non-native pronunciations: speech recognition histories and acoustic classification. In addition to describing these two techniques in detail, this paper presents a study that examines how acoustic classification and recognition history can assist an English spoken language dialogue in adapting to the pronunciation of Chinese native speakers.

2. RELATED WORK

Various research has been conducted in the areas of recognition accuracy improvement and non-native speech recognition. Litman, et al. [4] present a machine learning approach for adapting to poor speech recognition, Oh, et al. [5] use pronunciation variability analysis for non-native speak recognition, and Bouselmi, et al. [2] describe a phonetic confusion-based acoustic model.

Another related area is the overall design and evaluation of spoken language dialogue systems and related technologies. Examples include methods for evaluation of speech recognition accuracy [7], usability [6], and discourse understanding [3].

3. ACOUSTIC CLASSIFICATION

Acoustic classification evaluates the phonetic similarity of two given words. For any given pair of words, the proposed classification algorithm assigns a value between 0.0 and 1.0, where 1.0 indicates that the two words are phonetically identical. The algorithm takes the following factors into account: pronunciation, syllable count, and word length. Phonetic transcriptions of English words are obtained from an external dictionary.

The following equations provide a formal definition of the algorithm.

$$S(w_1, w_2) = S_p(pron(w_1), pron(w_2)) \quad (1)$$

$$S_p(p_1, p_2) = \frac{\sum_{i=1}^{min(len(p_1),len(p_2))} S_l(p_1(i), p_2(i))}{max(len(p_1), len(p_2))} \quad (2)$$

Equations 1 and 2 define the phonetic similarity between two English words w1 and w2 in terms of two functions called S and S_p. The two functions utilize three helper functions called *pron*, *max* and *len*. The *pron* function returns the pronunciation (phonetic transcription) for a given English word, the *len* function returns the length of a word in phonetic characters, and the *max* function returns the greater one of two given numerical values. The expression $p_x i$ denotes access to the i^{th} phonetic character of a word. The definition of S_p furthermore contains a reference to the function S_1, which is defined by Equation 3. The functions S_c

and S_v used in the definition of S_l are defined by Equations 4, 5, 6, and 7.

$$S_l(x_1, x_2) = \begin{cases} S_c(x_1, x_2) & \text{if both } x \text{ are consonants} \\ S_v(x_1, x_2) & \text{if both } x \text{ are vowels} \\ 0.0 & \text{otherwise} \end{cases} \quad (3)$$

$$S_c(x_1, x_2) = \begin{cases} 1.0 & \text{if } x_1 = x_2 \\ 0.7 & \text{if } (x_1, x_2) \in scSet \\ 0.5 & \text{if } x_1(0) = x_2(0) \\ 0.0 & \text{otherwise} \end{cases} \quad (4)$$

$$S_v(x_1, x_2) = \begin{cases} 1.0 & \text{if } x_1 = x_2 \\ 0.7 & \text{if } (x_1, x_2) \in svSet \\ 0.5 & \text{if } x_1(0) = x_2(0) \\ 0.0 & \text{otherwise} \end{cases} \quad (5)$$

$$scSet = \{(`d', `t'), (`t', `d'), (`p', `b'), (`b', `p')\} \quad (6)$$

$$svSet = \{(`*', `a'), (`a', `*')\} \quad (7)$$

As indicated in Equations 4 and 5, the functions S_c and S_v assign discrete values between 0.0 and 1.0 depending on the similarities of consonants and vowels. These discrete values depend on the particular phonetic transcription system used, and can be adjusted to achieve optimal results. The vowel and consonant sets defined in Equations 6 and 7 cover the most common phonetically similar characters defined in the phonetic transcription system. Both sets can be adjusted and extended to accommodate other phonetic transcription systems.

4. RECOGNITION HISTORY

The proposed recognition history is a module that collects, analyses, interprets information generated by the speech recognition engine of a spoken language dialogue system. It records both successful and unsuccessful recognition results and provides functions to query recognition history to resolve problematic recognitions. In particular, the history includes the following information:

- Identity of the user interacting with the system, if available.
- Speech recognition results. A speech recognition result is typically a so-called n-best list that contains an ordered list of all words matching the user's input. Some speech recognition engines can be configured to attach a confidence score to each word specifying the likelihood of being the word spoken by the user.
- Mappings of speech recognition results to the words spoken by the user (the intended words), if sufficient information is available.

The recognition history collects single words, compound words, and phrases. To illustrate its functionality, consider a dialogue that requires an input whose possible values are ``cable", ``able", and ``label" and the user says ``able", then the resulting n-best list generated by the speech recognition engine can be any permutation of lists containing the full set or a subset of the three possible values. If the dialogue is designed to resolve recognition failures, it is able to determine the intended word spoken by the user, and this information can be to create a mapping between the intended word and all recognition results (n-best lists) in the recognition history.

Apart from storing data, the recognition history provides functions for generating augmented n-best lists, determining possible intended words for a given recognized word, and determining possible recognized words for a given intended word.

4.1. Possible intended word lookup

This function uses the recognition history to determine possible intended words for one or more recognized words. If user information is available, this function can be configured to only consider history (i.e. n-best list mappings) associated with a specific user. The function can be invoked by a dialogue when it becomes apparent that a result returned by the speech recognition engine is not the intended word spoken by the user.

For example, consider a scenario where the recognition history data contains one or more records of the word “label” being recognized as “table” when a certain user U interacts with a dialogue D. If user U interacts with dialogue D again, and the recognition result is “table”, then the possible intended word lookup function of the recognition history can be used to determine that the actual intended word is “label”.

4.2. Possible recognized word lookup

This function is the inverse of the possible intended word lookup. It can be used to determine possible words that are recognized by the speech

recognition engine, if a user speaks a certain intended word.

4.3. Augmented n-best list generation

The functions described in the previous two sections can be invoked by a dialogue after it has determined that a recognition is problematic. The augmented n-best list generator function uses the data in the history to extend and rearrange an n-best list received from the speech recognition engine to generate a new n-best list whose first entries are more likely to match the intended word spoken by the user. It is invoked by a dialogue as soon as a n-best list from the speech recognition engine is received and its parameters are the n-best list and a value that specifies the maximum length of the generated augmented n-best list.

The augmented n-best list generation function uses recognition history data to calculate rank scores for each n-best list entry and re-orders the list accordingly. The acoustic classification algorithm described in section 3 is used to determine and add additional, phonetically similar words to the n-best list.

The augmented n-best list generation function employs the following algorithm:

1. Create a new augmented n-best list by copying the n-best list received from the speech recognition engine.
2. Use acoustic classification to obtain phonetically similar words for each entry in the received n-best list.
3. Set rank score of phonetically similar words to -1.0 and append them to augmented n-best list.
4. For each word in received n-best list:
   (a) Access history to retrieve n-best list containing current word and mapped intended word.
   (b) Calculate a mean rank score based on position of word in retrieved history n-best lists and, if available, confidence score.
   (c) Add word to augmented n-best list.
5. Order augmented n-best list by rank score (descending)
6. If length of augmented n-best list exceeds maximum length parameter, truncate it.
7. Return augmented n-best list.

5. EVALUATION

This section presents an experiment designed to evaluate the ability of the proposed concepts to help dialogue systems to adapt to mispronunciations and non-native pronunciations. The experiment uses a concrete dialogue system that allows users to look up English terms defined in the glossary of the Hong Kong legal corpus. The dialogue interacts with its users as follows:

1. Ask user to speak single- or multi-word term.
2. Prompt user to confirm recognition.
3. Look up term and read its definition to user.

The dialogue system uses four randomly generated groups of 1500, 3000, 6000, and 12000 terms defined in the Hong Kong legal glossary. Each term group is a subset of the glossary and organized as a network linking terms that have been designated as phonetically similar by the acoustic classification algorithm described in section 3.

5.1. Subjects

The experiment uses ten native Mandarin Chinese speakers who speak English as a second language (5 males and 5 females) as subjects. Each subject is supplied with four lists containing 25 randomly selected terms from each of the four term groups and receives instruction on how to interact with the dialogue.

5.2. Experiment description and results

The experiment consists of several phases: a training and audio capture run ("Run 1") during which the users go through the term lists once to record their inputs and allow the recognition history to gather data, a phase where log files and recognition history are examined to identify problematic terms, and a second run ("Run 2") during which the previously recorded inputs are used to simulate a second run.

Figure 1: Number of Problematic Terms (Comparison).

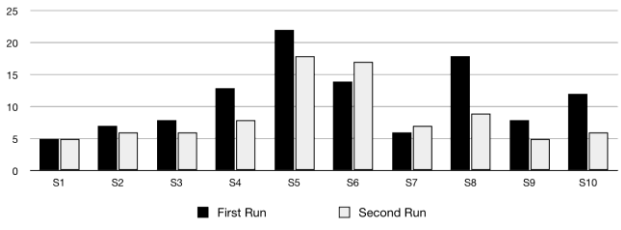

Table 1 shows the amount of problematic terms that were identified after the first run for each participant in for each of the four term groups. Overall, the numbers indicate that the range of problematic terms for each subject is between 5% and 22%, with an average of 11.3%.

Table 1: Number of Problematic Terms (Run 1).

	G_{1500}	G_{3000}	G_{6000}	G_{12000}	Sum
S1	1	1	0	3	5
S2	1	1	3	2	7
S3	1	3	2	2	8
S4	0	3	5	5	13
S5	8	3	4	7	22
S6	4	0	6	4	14
S7	1	0	3	2	6
S8	2	5	7	4	18
S9	2	3	1	2	8
S10	4	2	0	6	12
Avg	2.4	2.1	3.1	3.7	11.3

Table 2: Number of Problematic Terms (Run 2).

	G_{1500}	G_{3000}	G_{6000}	G_{12000}	Sum
S1	2	0	1	2	5
S2	1	0	3	2	6
S3	1	2	2	1	6
S4	0	3	2	3	8
S5	6	4	5	3	18
S6	5	2	5	5	17
S7	3	1	1	2	7
S8	0	3	4	2	9
S9	2	1	2	0	5
S10	1	2	0	3	6
Avg	2.1	1.8	2.5	2.3	8.7

Table 2 shows the number of problematic terms encountered during the second run and Figure 1 compares the number of problematic terms encountered in the first and second run. As the tables and figures show, the number of problematic terms the subjects encountered during this phase has been reduced. In particular, Figure 1 shows that, in comparison with the training run, eight of the subjects (S_1, S_2, S_3, S_4, S_5, S_8, S_9, and S_{10}) encountered fewer problematic terms than before, while the remaining two subjects (S_6 and S_7) encountered more problematic words. As indicated in table 2, the range of problematic terms for each subject has narrowed from 5% - 22% to 5% - 18%, and the overall average of problematic terms has decreased from 11.3% to 8.7%, which is a reduction of approximately 23%. In addition, the log files indicate that the overall number of inputs (i.e. total number of times the subjects has to repeat terms until they are successfully recognized) has been reduced.

6. CONCLUSION

The results of the experiment indicate that acoustic classification and recognition histories can help dialogue systems adapt to (mis-)pronunciations of its users. Without existing recognition history data, the overall number of misrecognized terms is approximately 29.8% higher than the overall number of misrecognitions that occur after the system has collected recognition history data for each user. Our future work will focus on combining the proposed concepts with grounding and dialogue management strategies to resolve non understanding and mis-understanding in a natural way.

7. REFERENCES

[1] Aust, H., Oerder, M., Seide, F., Steinbis, V. 1995. The philips automatic train timetable information system. *Speech Communication* 17(3-4), 249-262.

[2] Bouselmi, G., Fohr, D., Illina, I., Haton, J.P. 2006. Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints. *Proc. Interspeech 2006.*

[3] Higashinaka, R., Miyazaki, N., Nakano, M., Aikawa, K. 2004. Evaluating discourse understanding in spoken dialogue systems. *Transactions on Speech and Language Processing*, 1-20.

[4] Litman, D., Pan, S. 2000. Predicting and adapting to poor speech recognition in a spoken dialogue system. *Seventeenth National Conference on Artificial Intelligence*, 722-728.

[5] Oh, J.R., Yoon, J.S., Kim, H.K. 2007. Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication*, 49(1), 59-70.

[6] Salonen, E., Hartikainen, M., Turunen, M. 2004. Subjective evaluation of spoken dialogue systems using servqual method. *Proc. Interspeech 2004*, 2273-2276.

[7] Webster, J., Fang, A., Liu, X.Y., Li, W. 2009. Multi-factor evaluation of speech recognition accuracy for better dialogue system design. *Proc. PACLING 2009*, 274-282.