MODELLING MAN-COMPUTER ORAL DIALOGUE IN NOISY ENVIRONMENT

Josef Psutka, Jiří Kepka, Luděk Müller, Zbyněk Tychtl

University of West Bohemia, Department of Cybernetics
Univerzitní 22, 306 14 Plzeň, Czech Republic
e-mail: psutka@kky.zcu.cz

ABSTRACT

A model of a voice controlled system will be presented in the paper. The behaviour of the system is modelled by a finite state process. The problem domain is supposed to be well bounded and very limited. An isolated word classification method is used for the recognition of user's control command or sequence of commands. A speech synthesizer is used to implement acoustic feedback control. As an illustration example its implementation and application for searching and updating database is described. Problems involved in classification and communication between the system and the user in noisy environment are treated. Optimization tradeoffs are proposed.

1 INTRODUCTION

Spoken language is a natural form of communication in many man-computer interfaces. Although speech recognition systems have been available for some time, their limitations have so far prevented their widespread use. Most frequently cited design issues in speech systems involve

- continuous versus isolated word speech recognition,
- large versus small vocabulary,
- speaker dependence versus speaker independence system,
- broad versus narrow grammar.

An often omitted issue of an extreme importance is the one of laboratory conditions versus real conditions. In noisy environment the recognition accuracy should be expected to decrease in dependence on the influence of noise and distortions. Moreover, highly interactive applications require real-time speech recognition speed, which is hard to achieve on a standard hardware, especially when higher-level knowledge should be utilized. As the result the isolated word recognition systems with limited vocabularies are preferred to perform satisfactory in real time and noisy environment.

Our research group has developed a real-time man-computer dialogue system that can be used for voice controlled searching (and sometimes also for updating) databases. The main part of this system – the isolated word recognizer – has been designed to be able to work in a noisy environment. The system is able to compare an unknown utterance against more than a hundred stored templates in real time and it provides the high possibility how to achieve optimal tradeoffs among the rate of recognized, unrecognized and misrecognized words.

2 BRIEF DESCRIPTION OF THE DIALOGUE SYSTEM

The dialogue between the user and the computer communication system is performed by means of a transmitter-receiver. The control mechanism supports the dialogue between the user and the database, see Fig.1. It may be modelled by a deterministic finite transducer

$$M = (Q,\ V,\ \Delta,\ \delta,\ q_0,\ F) \qquad (1)$$

where

- Q is a finite set of states representing the nodes in the problem state space;
- V is a finite set of input symbols representing a finite set of key-command words which corresponds to the vocabulary of the recognition system;
- Δ is a finite set of output symbols each of which corresponds to some action;
- δ is a mapping from $Q \times V$ into $\Delta \times W$ representing all acceptable state transitions, i.e. it describes all acceptable ways of the dialogue between the user and the database system, where W is a set of synthesized messages;
- q_0 is the initial state of the dialogue;
- $F \subseteq Q$ is the set of final states representing the final states of the dialogue.

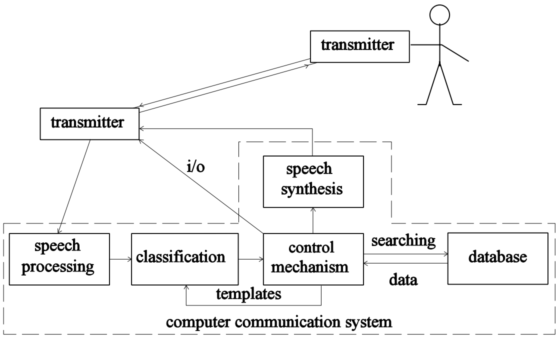

Fig. 1 An example of voice controlled database searching.

Actions of the control mechanism support the dialogue between the user and the database. They are responsible for e.g.

- reading database items and initiate the synthesizer;
- updating database items (if possible);
- providing the proper vocabularies (templates) for the classifier according to the state of the dialogue;
- asking the user for the decision how the dialogue shall continue, etc.

3 HUMAN-COMPUTER VOICE INTERACTION IN NOISY ENVIRONMENT

Because the application of communication systems in noisy environment deserve our attention we have to solve at least two problems involved in classification.

1. In the course of communication extraneous "words" can appear at the input of the system as the result of noise, extraneous speech, and other distortions. Using only the minimum distance criterion such a "word" is misclassified i.e. it is considered to belong to the current vocabulary.

2. Although the processed word belongs to the vocabulary, it can still be misclassified due to acoustical similarity to other words and/or additional noise and distortions.

If we want database items to be searched and changed by voice, the rate of misrecognized words is desired to be made as small as possible. That is the reason, why the above described system when operating in noisy environment should synthesize and report on most of classification results, so that the user could either confirm them by saying the chosen confirmating command (*CC*) or reject them by saying the rejecting command (*RC*). Then in the former case, either a corresponding path in the database is chosen or some database item is changed. In the latter one, the user is asked to repeat the misrecognized word.

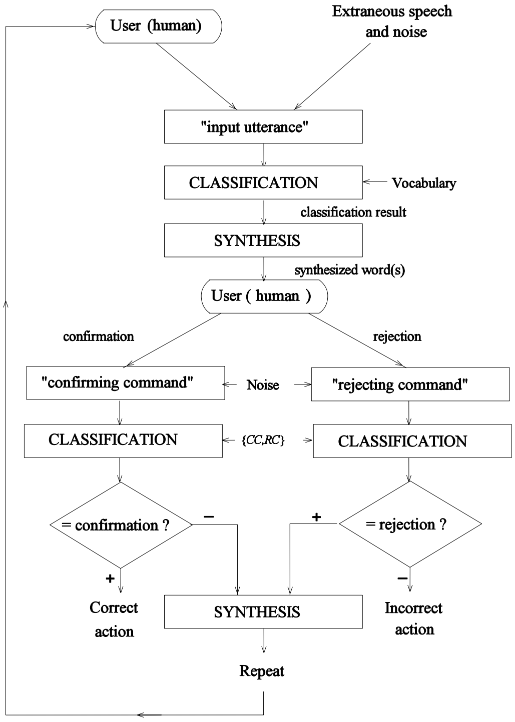

Fig. 2 The principle of communication in the voice controlled system.

Even when the system synthesizes and announces classification results, still the demand for extremly low misrecognition rates of key-command words such as "*CC*", "*RC*" or some other, for example "*error*", "*help*", etc. has to be solved to prevent incorrect actions, see Fig. 2.

When a classification result is to be either confirmed or rejected, only the vocabulary {*CC*, *RC*} is utilized. If a user's utterance of "*CC*" is misrecognized, then it is regarded to be "*RC*" instead of "*CC*" and vice versa. In the former case, as the result, the user is asked to repeat the utterance again, although this has been correctly recognized. In the latter case, some incorrect action is invoked (either an incorrect path is chosen or some database item is incorrectly changed) instead of asking the user to repeat the misclassified word again. The rate of the two cases should be made as small as possible. On the other hand, it should be noted that while the former case usually results only in a relatively small lost of the time (the process has to be repeated), in the latter case, the incorrect action can sometimes result in consequences the correction of which by voice is either very difficult or quite impossible. That means that in some states of communication the system should rather ask the user to repeat the word than accept an uncertain classification result.

Another problem is that when extraneous speech and/or noise is loud enough to be considered an input word, the classification is performed, the result is synthesized and the user is asked to take a decision on its correctness, although he has been silent in fact.

In the next section, the possibilities how to achive optimal tradeoffs among the rate of recognized, unrecognized, and misrecognized words are treated.

4 OPTIMAL TRADEOFFS

The method based on threshold distances has been proposed and implemented. The distances computed between an unknown token and stored templates are treated as random variables. For each word from the vocabulary a threshold distance is set on the basis of a training process so that the distance computed between an utterance of the word from the vocabulary and its nearest template is less than the corresponding threshold distance with the desired rate, see Fig. 3. If threshold distances for all words are computed to achive a desired rate, e.g. 99%, then cca. one percentage of user's utterances of these words will exceed the corresponding threshold distances. Some utterance of this word will be unrecognized (if $D_{rr} > t_r$ for every *r* from the vocabulary). But it should be stressed that at the same time the entire majority of extraneous words will also exceed all current thresholds and will be rejected.

That means that for every "input word" (either a user's utterance or an extraneous word loud enough) the distance to the nearest template in the current vocabulary is firstly computed and then it is compared with the corresponding threshold distance. If it is less than the threshold, the "input word" is regarded as from the current vocabulary, see Fig.4. Otherwise, it is treated as an extraneous word and very short sound (sound label) is synthesized so that the user can be quickly informed that either some extraneous speech was loud enough to get to the input of the classifier or that the utterance is treated as an extraneous word (noise). In the latter case, the user is expected to repeat the unrecognized word.

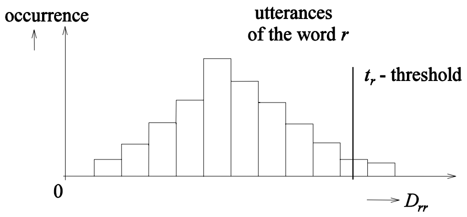

Fig. 3 Histogram.

The less the threshold distances, the larger the rate of unrecognized utterances of the words from the vocabulary and the less the probability for extraneous speech and/or distortions to be regarded as from the vocabulary. Therefore, an optimal tradeoff has to be chosen.

For example, let the incorrect replacement of "*RC*" by "*CC*" be some result in some error action, see Fig. 2. Two smallest distances between the utterance of an unknown input word (either "*CC*" or "*RC*" are expected) and the nearest template of "*CC*" and the nearest template of "*RC*" are computed, respectively. Let us denote them d_{CC} and d_{RC}. Let K is some threshold value, $K>0$. Then an unknown utterance u is classified:

$$u = \begin{cases} RC & d_{CC} - d_{RC} > 0 \\ CC & d_{CC} - d_{RC} + K \leq 0 \;\equiv\; d_{RC} - d_{CC} - K \geq 0 \\ repeat & d_{CC} - d_{RC} + K > 0 \;\wedge\; d_{CC} - d_{RC} \leq 0 \end{cases} \qquad (2)$$

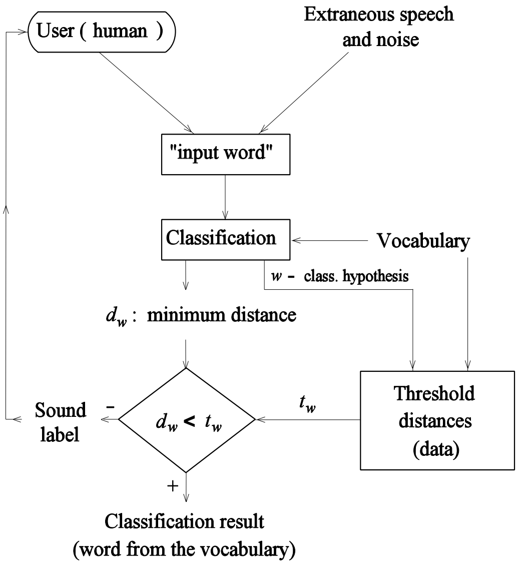

Fig. 4 Classification process using threshold distances.

The threshold value K in (2) can be set on the basis of the training process so that the maximal accepted rate M_r of unrecognized utterances of CC can not be exceeded, i.e.

$$P(repeat \mid u=CC) = M_r = P(d_{CC} - d_{RC} \leq 0 \;\wedge \\ \wedge\; d_{CC} - d_{RC} + K > 0)\,. \qquad (3)$$

In this manner the conditions for utterances of CC to be recognized as CC are made more difficult, because

$$P(CC \mid u=RC) = P(d_{RC} - d_{CC} - K \geq 0 \mid u=RC) \leq \\ \leq P(d_{RC} - d_{CC} \geq 0 \mid u=RC). \qquad (4)$$

Generally, the value of these thresholds should be set so that the optimal tradeoffs between the number of recognized, unrecognized and misrecognized words can be achieved.

It should also be noted that this tradeoff may vary according to the state of the dialogue. Especially in the states, when a classification result of the user's utterance is expected to be confirmed, the threshold distances for CC and RC should be set with a great care. Especially, the threshold distance for the word CC should prevent extraneous speech from being classified CC, because in this way a misclassified "input word" can be confirmed (the user usually does not reply immediately). In distinct states of communication distinct vocabularies can be used. That means that distinct threshold distances can be generally set for each word which belongs to at least two vocabularies.

Practical experience with the developed system has shown that if a processed word belongs to the current vocabulary and is misclassified using minimum distance criterion, then in the most cases the second best candidate obtained by the classification process is just the correct one. Therefore, two smallest distances between an utterance of an unknown word and two nearest templates of distinct words from the current vocabulary are the most important features for final decision making. The less the difference between them, the higher misclassification probability.

To explain this problem more detailed let d_a and d_b be the two smallest distances between an utterance u of an unknown input word and two nearest templates of the word a and b, $a \neq b$, from the current vocabulary V. In other words, it holds

$$d_a + d_b = \min \{d_x + d_y \mid x, y \in V, x \neq y\}, \qquad (5)$$

where d_c, $c \in V$, is the distance to the nearest template of c. Let $f(d_a, d_b \mid a)$ and $f(d_a, d_b \mid b)$ be two-dimensional density functions. Let us suppose that a correct classification result is either *a* or *b*. If we also assume that both the costs of an error of different kinds and the prior probabilities (P_a, P_b) are identical the using Bayes rule an unknown utterance *u* will be classified

$$u = \begin{cases} a, & \text{if } f(d_a, d_b \mid a) > f(d_a, d_b \mid b) \\ b & \text{otherwise .} \end{cases} \tag{6}$$

It is known that for equal cost, the Bayes classifier is essentially a maximum posterior probability classifier; e.g. for *a* :

$$\begin{aligned} P(a \mid d_a, d_b) &= \frac{f(d_a, d_b \mid a)\, P_a}{f(d_a, d_b)} = \\ &= \frac{f(d_a, d_b \mid a) P_a}{f(d_a, d_b \mid a) P_a + f(d_a, d_b \mid b) P_b} \ . \end{aligned} \tag{7}$$

As we want classified results to be reliable we can use the following equation to replace (6) by taking into account the trade-off between the substitution (misrecognition) rate and the rejection (unrecognition) rate

$$u = \begin{cases} a, & \text{if } P(a \mid d_a, d_b) > P(b \mid d_a, d_b) \wedge \\ & \quad \wedge P(a \mid d_a, d_b) \geq \alpha \\ b, & \text{if } P(b \mid d_a, d_b) > P(a \mid d_a, d_b) \wedge \\ & \quad \wedge P(b \mid d_a, d_b) \geq \alpha \\ \text{unrecognized} & \text{otherwise} \end{cases} \tag{8}$$

with $0.5 < \alpha \leq 1$ being a threshold.

5 EXPERIMENTAL RESULTS

Two vocabularies were used for experiments: V_1 = {info, soubor (file), edit, hledání (search), okna (windows), volba (choice), text, kursor (cursor), blok (block)}, V_2 = {obnov (undelete), přenes (move), kopíruj (copy), vlož (insert), zobraz (view), vymaž (delete), pravý (right), levý (left), menu, ven (quite)}.

Each word from V_1 were pronounced 200 times by one speaker. The templetes obtained were processed by a clustering algorithm and ten "optimal" templetes of each word were found. A threshold distance for each word from V_1 was computed to reach 2% rate of unrecognized utterances of this word. Then 2000 utterances of the word from V_2 (200 utterances per each word) were brought to the input of the classifier. 99.05% of utterances of these (extraneous) words were rejected. The words from both vocabularies were pronounced by the same speaker. In another experiment the words from V_2 were pronounced by the other speaker. In this case 99.9% of utterances of these words were rejected. Let us mention that these experiments were performed in relatively quiet environment.

If samples of real extraneous and distant speech combined with noise and other distortions (they must be loud enough) were brought to the input of the classifier, the rate of correctly rejected "utterances" of extraneous words would have still increased, because of a large dissimilarity to the user's speech than in the two above described experiments. On the other hand, it should be noted that all threshold values should be trained in the environment where the system is planned to work. In a noisy environment, the threshold distances will increase in dependence on the influence of noise and other distortions and as the result the probability for extraneous words to be regarded as from the current vocabulary will proportionally increase.

In the third experiment only a threshold α, α = 0.8, was used and 2000 utterances of the words from V_1 were classified according to (8). The recognition rate achieved was 99.5% while the rest 0.5% of utterances remained unrecognized. Without α the recognition rate achieved was 99.8% but 0.2% of utterances (5 utterances) were misclassified. It should be noted here that the value of α may be made dependent on *a* and *b* and also on the state of dialogue between the user and the system.

6 CONCLUSION

The problem area that has just been discussed has a practical background. Formerly we received a task from the research section of the Czechoslovak railways that could be mentioned as an application of such a man-computer voice communication system. The problem is that at each main railway station data on wagons in formed train sets are checked and, if necessary, also updated. Data on wagons include such items as destination, number, weight, length, state, direction, kind, etc. The dialogue between two railwaymen is performed by means of transmitter-receivers. One railwayman goes along a train set and transmits data on wagons to the railway station building, where the other railwayman compares received data with corresponding database items. If necessary, the items are updated. The aim of our effort has been to partly automate this process.

Although the task has never been implemented in practice because of financial difficulties of the Czech(o-slovak) railways we have obtained much practical and very valuable knowledge by solving this problem, for example in such domains as the dialog modelling in a transmitter-receiver mode, the classification in noisy environment, the adjustment of optimal tradeoffs among the rate of recognized, unrecognized and misrecognized words, etc.

REFERENCES

[1] KEPKA J., PSUTKA J.: Human-Computer Voice Communication in Noisy Environment. In: 3rd IFIP WG-7.6 Working Conference on Optimization. Prague 1994, pp.94-97.

[2] KUHN M.H., TOMASCHEWSKI H.H.: Improvements in isolated word recognition. IEEE Transact. on ASSP, **31**, 1983, No.1, pp.157-167.

[3] RABINER L.R.: Application of Voice Processing to Telecommunications. Proceedings of the IEEE, **82**, 1994, No.2, pp.199-228.

[4] RICH C., KNIGHT K.: Artificial Intelligence. McGraw-Hill, Inc. New York, London 1991.