SPEAKER-INDEPENDENT WORD RECOGNITION TECHNIQUES FOR CONTROL OF A VOICE MESSAGING SYSTEM

V. Gupta and P. Mermelstein

Bell-Northern Research
Nuns' Island, Quebec

ABSTRACT

Word recognition allows flexible control of new services on the telephone network such as voice messaging. We describe an experimental voice messaging system that is controlled by means of speaker-independent recognition of a small vocabulary over the dial-up network. We review the recognition performance attained for native English and French Canadian speakers of English and consider the impact of varying accents on the design of access protocols. Our simulations lead us to estimate that voice access is feasible for most members of the potential user population. Some comments on the design of spoken man-machine dialogues conclude the presentation.

RÉSUMÉ

La reconnaissance des mots isolés nous permet de commander l'accès aux services nouveaux sur le réseau téléphonique comme l'enregistrement et la transmission de messages parlés. Nous décrivons un système expérimental d'enregistrement et de transmission de messages parlés commandé par la reconnaissance des mots d'un vocabulaire restreint. Nous présentons des résultats sur le taux de reconnaissance atteint pour des locuteurs anglophones et francophones parlant anglais. Les accents des locuteurs modifient la valeur du taux de reconnaissance qu'on peut atteindre et influence la conception des protocoles d'accès. Les simulations que nous avons effectuées nous permettent de conclure que le système de reconnaissance verbale est utilisable pour la plupart des usagers potentiels. Nous concluons cette communication sur quelques remarques concernant la conception des dialogues homme-ordinateur.

INTRODUCTION

Word recognition represents today a viable input medium for man-machine systems due to the flexibility with which it can be incorporated into the user-machine dialogue and the wide availability of terminals, the large variety of telephone sets. Two important requirements exist for satisfactory design of such systems, a high speaker-independent recognition capability as well as the design of protocols that minimize communications problems in the presence of misrecognition of words. We review experimental results on speaker-independent recognition of a 20-word vocabulary for a sizeable population of native and foreign-accented speakers and indicate the problems that prevent rapid attainment of significantly lower error rates. Considerations in the design of communication protocols include dynamic alteration of the vocabulary acceptable at various points in the dialogue, verification of inputs where the risk associated with misrecognition is high and the use of error-protection digits to input digit-string identifiers. These points are illustrated in the context of the design of the dialogue for the control of voice message system accessed over the switched telephone network.

SPEAKER-INDEPENDENT RECOGNITION OF SPOKEN WORDS

Most recognition systems follow the classical pattern recognition paradigm of selecting appropriate representations of the test and reference data, defining a similarity metric that approximates the probability that a test item belongs to a particular reference category, and applying a decision rule that identifies the test item with the reference category to which it is most similar or rejects the test item because the recognition cannot be made with sufficient confidence. The following choices thus appear critical:

1. Selection of the reference data and its compact representation.
2. Representation of the test data in a form suitable for comparison (feature extraction).
3. Metric for computation of the similarity between the test and reference items.

A ubiquitous recognizer must possess information in the form of data or rules concerning all acceptable acoustic and phonetic variations of the words in the vocabulary. This idealized goal is approximated in practice by collecting reference data, i.e., labelled productions of the words in the vocabulary, from a large number of speakers including representatives of the major accent groups that will be included in the potential user population. Speaker-independence can only be verified for the reference speakers by treating one subset of the reference group as if it were the test group and training the system on the reference data that do not include that subset. If recognition performance is comparable across subsets of speakers used to train and test the systems, a valid claim for speaker-independence may be made. Such speaker-independence will only be achieved in general if both subsets contain adequate representatives from the different dialect and accent groups. Of course, if the user population is to include men, women and children, appropriate representative speakers from each linguistic grouping of each subset of the user population must be included.

To achieve economical storage of the reference data and simplify the comparison process between test and reference data, it is important to cluster the reference words and retain only a simplified representation of each cluster. Such a representation may be a reference word typical of the members of the cluster or one that corresponds to some generalized average of the constituents of the cluster. Use of an averaging process generally leads to somewhat lower error rates since it allows one to maximize the average similarity between the members of the cluster and its representative and thereby minimize the likelihood of misrecognizing the test data.

The clustering operation requires the definition of a distance measure between reference words. To achieve the highest recognition rate, this distance or dissimilarity metric should be the same as used for computing the distance between an unknown word and the reference words in the recognition process. The clustering technique we have found best in our experimental studies is known as the complete-link clustering algorithm [1]. It divides the reference productions of a given word into a specified number of clusters such that the maximum within-cluster distance is minimized. The clustering computation is carried out independently for each word without regard to the speaker who spoke the particular word. The technique requires computation of the matrix of distances between each pair of reference tokens of the same word, and is thus lengthy and generally implemented on a computing system that is more powerful than the actual recognition hardware.

Representation of Test and Reference Words

Test and reference words are represented in the form of energy in time-frequency space, a representation that is known to produce negligible information loss for the purpose of recognizing isolated words. Time and frequency resolution for the representation are selected as a result of studying the tradeoffs between the increased cost of storage of reference items and complexity of the distance computations with increasing resolution and loss of acoustic information leading to degraded recognition performance. In our experimental system we have employed spectra computed every 12.8 ms. A particularly compact spectral representation with practically no loss of information significant for recognition is achieved by transforming the frequency spectra as computed or measured by a bank of bandpass filters into vectors of seven mel-based cepstrum coefficients. These coefficients correspond to the cos-transform of the log-power spectra spaced linearly with frequency up to 1 kHz and logarithmically thereafter [2]. For systems operating on the telephone network, use of an appropriate speech detector that accurately locates the start and end of each word and thereby separates the word from the background noise is an indispensable requirement.

Similarity or Distance Computation Between Words

Variation in the duration of individual sounds is the major obstacle to be overcome in comparing for similarity between different productions of the same or different words. The comparison process we find best is a two-stage process consisting of linear and non-linear alignment stages. The first stage pseudo-linearly changes the time scale of the pattern with the smaller number of frames by duplicating frames at regular intervals so that the number of frames in the tokens to be compared is equalized. To obtain a more precise alignment, a dynamic programming procedure [3] is used that allows a change in the time scale in one region of the word if it is compensated by opposite changes in a second region. Among the permitted alignment paths, the dynamic programming procedure selects the path that minimizes the cumulative distance between aligned frames of the two tokens. The distance between pairs of frames is computed as the Euclidean distance between the cepstrum arrays. The final distance or dissimilarity is then given by the average frame distance between the aligned tokens. An unknown word is compared with all reference words of all words allowed as inputs at that point in the dialogue and recognized as belonging to the category of the most similar reference word if the decision can be made with sufficient confidence. The confidence measure used is the ratio of the distance to the most similar word and the distance to the most similar reference form of the next most similar word.

Recognition Experiments with Native Speakers and French Canadian Speakers of English

To assess the importance of accent on the performance of the word recognizer we have tested its performance separately on two disjoint groups of speakers, one including male and female native Canadian speakers of English, the second including French Canadian speakers of English whose accent varied significantly within the group. Reference data in each case were generated from 30 different male and female speakers from the same population group. All data were recorded over individually switched local dial-up connections and speakers were instructed to speak the words in response to randomized spoken prompts. As shown in Fig. 1, significant differences can be noted in recognition rate depending on the number of reference templates employed for each vocabulary word. Not only is the recognition rate lower for the non-native speakers but an asymptotic performance level is reached only with many more templates per word. Both results appear to be due to the greater between-speaker variation within the non-native speaker group. This result emphasizes the importance of including foreign-accented speakers among the speakers used to generate the reference data. Analysis of the variation of recognition rate with individual speakers shows that 80% of the French Canadians and 95% of the English Canadians yield better than 90% word recognition accuracy, a performance threshold which if not reached would result in a system that would be very difficult to employ in practical applications.

In order that a system be readily usable by a large population of speakers, training on individual users at the start of every conversation or storing data about individual speakers' speech characteristics cannot be considered. Unfortunately our knowledge of the detailed acoustic variations between different speakers' productions of specified words is limited and this represents the main impediment to better speaker-independent word recognition today. Although clustering entire words is a practically useful technique, through more precise representation of the acoustic variations we expect that many of the current limitations will be overcome in the future.

The above recognition rates hold for a 20-word vocabulary, the ten digits and some ten frequently used control words such as "yes", "no", "stop", "insert", "recall", that may be used in a variety of man-machine communication applications. Two-way "yes/no" discrimination is found successful better than 99% of the time. Thus limiting the recognition wherever possible to the responses meaningful in the particular context is an important tool to minimize misrecognition as well as computation time.

The recognition accuracy for strings of digits drops rapidly with any decrease in the raw digit recognition capability discussed above. Thus at 90% digit-recognition the recognition of arbitrary three-digit identifiers is only 73%. Through addition of a fourth check-digit, and constraining valid digit strings to add to zero modulo ten, most single digit errors can be corrected by substituting the most likely four-digit sequence that meets the imposed code constraint. This permits an increase in the digit-string from 73% to roughly 95%. Since we can associate a confidence measure with any recognition decision, we can ask the user to verify decisions that would drop below a confidence threshold. For example, accurate knowledge of the user code of a recipient is essential for proper operation of the messaging system, therefore, address code will be played back for "yes/no" verification relatively frequently. Since "yes/no" recognition is highly accurate, false acceptance is thus avoided and new input is required only when the input was correctly rejected.

CONSTRUCTION OF MAN-MACHINE DIALOGUES

We now wish to consider the construction of a spoken dialogue for control of a voice message system. The same principles may be appropriate for use in other information retrieval applications, thus we will attempt to generalize to situations beyond the specific messaging application. The dialogue is in the form of spoken prompts from the system and spoken words input by the user. To minimize the language understanding problem, the dialogue is designed to be under the control of the system. The dialogue can be modelled as a transition network where the next state is always determined by the current state and the input from the user.

The experimental voice message system we have designed allows users to dial our computing facility, identify themselves, and input or retrieve messages. Much like a text-message system, retrieval is selective. Messages appear to the user in reverse sequential order and he may specify the messages he wants to be played back. Although a practical system may accept calls forwarded from busy or unanswered telephones and deliver messages to specified destinations, the experimental system does not yet have such capabilities.

For high user acceptance and cost-effective employment of the available access ports to the message system, the verbal dialogue should be as brief as possible. In contrast to text prompts, the user cannot review the prompt except by failing to input a response. To eliminate any doubt on the part of the user as to the spoken input required, the instructions should be as specific as possible. Thus the design of the dialogue represents a compromise between brevity and clarity and dialogues that meet both those requirements generally lead to satisfied users.

The function of the prompt messages is to acknowledge input by the user and request additional input. When requesting new input, it is important that the prompt specify the choice of words available for the response. Thus the prompt query, "Do you wish to insert or recall a message?" implicitly specifies that the only legal responses at that point are "insert" and "recall". The user may be given printed instructions so as to familiarize himself with the system's capabilities, but he cannot be expected to refer to written instructions while using the system. By including and emphasizing the response word "insert" and "recall" in the prompt message, the response choices are identified. The dialogue may be structured into a series of "yes/no" responses, e.g., "Do you wish to insert a message?" "No","Do you wish to recall a message?" "Yes"..., but this tends to lengthen the conversation and thereby the system occupancy per transaction.

In order for the system to function with only an isolated-word recognition capability, the information requested by the system is always limited to what can be supplied by a single word or a sequence of digits of prespecified length. A limited time, normally 2 seconds, is allotted for any input. If no speech is detected within that interval the prompt is repeated. A repeated absence of input results in an abort of the dialogue. Rapid time out for speech input is necessitated by the heavy computing resources assigned to real-time recognition of input. This situation is markedly

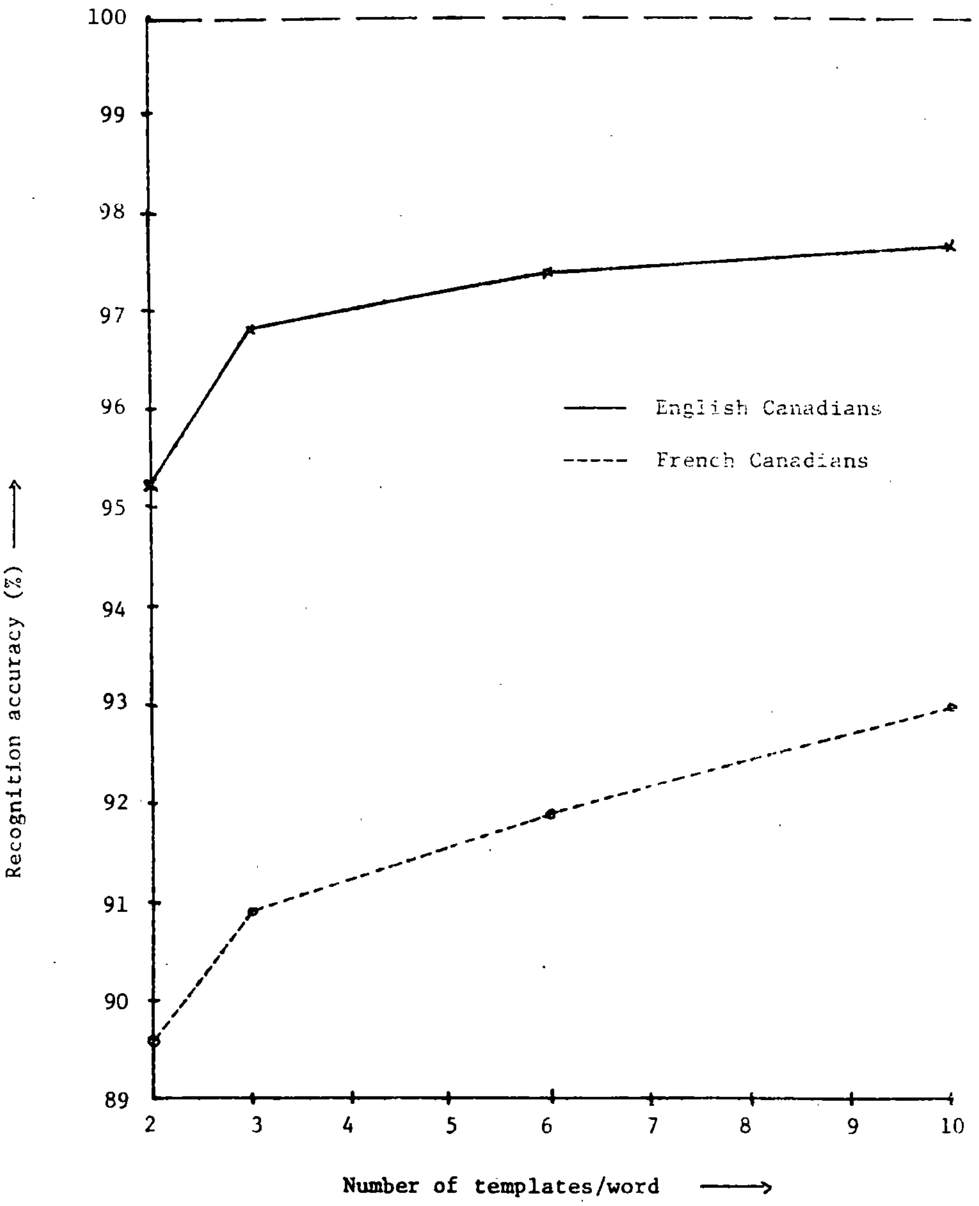

Fig. 1 Average word recognition accuracy as a function of number of templates/word.

different for text input where characters can be buffered on an interrupt basis until an input command line is terminated.

The design of the dialogue should reflect the differences between normal and exceptional situations. For example, in order to achieve storage economies it is desirable that users not save their messages except in special circumstances. Therefore, the user is not asked whether he wants to delete the message, the normal procedure, but rather whether he wants to save the particular message.

The input of strings of digits to identify senders and recipients of messages is controlled by means of short prompt tones. The tones are useful to prevent the speaker from talking before the prompt is finished and thereby avoid possibly losing the starting section of his input. Additionally they serve to separate the users' successive digits and thereby ease the task of recognition. The affect of neighbouring words on pronunciation of individual words is thus greatly reduced. Additionally, the rate of speaking is thereby artificially slowed so as to enhance the likelihood of clear articulation.

CONCLUSIONS

Man-machine communication with the aid of word recognition over the switched telephone network is technically feasible today. However, great care must be exercised in the design of such systems in order that they enjoy high user acceptance. The spoken information channel is different from that represented by written or keyed text. It is vital that the differences be kept in mind when designing protocols for spoken man-machine dialogues.

Such systems are in their infancy today. One limitation is the restricted recognition capability. However, careful dialogue design can overcome many of the limitations in actual recognition capability and lead to acceptable systems today. Improved recognition techniques and faster signal-processing logic will undoubtedly lead to even higher speaker-independent recognition capabilities. As these enhancements are attained, we can expect to be able to relax the currently highly structured conversational styles so as to arrive at a freer style of speech communication between man and machine.

REFERENCES

1. P. Hansen and M. Delattre, "Complete-link Clustering Analysis by Graph Colouring", J. Amer. Statist. Ass., vol 73, 397-403, 1978.

2. S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoustics, Speech and Signal Proc., ASSP-28, 357-366, 1980.

3. V. Gupta and P. Mermelstein, "Effects of Accent on the Performance of an Isolated-Word Recognizer", J. Acoust. Soc. Am., (submitted for publication).