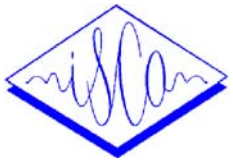
ISCA Archive
http://www.isca-speech.org/archive



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

AN INTELLIGENT TELEPHONE ANSWERING SYSTEM USING SPEECH RECOGNITION

Lobanov, B.M., Brickle, S.V., Kubashin, A.V., Levkovskaja, T.V.
Institute of Engineering Cybernetics, Academy of Science of Belarus
Surganov St. 6, 220011 Minsk, Belarus
Tel. +375 172 685295, Email lobanov@novcom.bas-net.by

ABSTRACT

The computer system described in this paper answers incoming telephone calls and employs speaker-independent speech recognition to identify callers. The users of the system can define caller-specific treatment and change this treatment using a graphical user interface. Apart from relating and receiving spoken messages, the system also offers advanced telephony features such as paging and call forwarding, providing the required subscription services are available from the telephone company. Standard interfaces to the telephony and audio hardware are used, so that the system runs on a desktop PC equipped with a voice-enabled modem.

1. INTRODUCTION. SYSTEM IN BRIEF

This report describes a telephone answering computer system using speech recognition, called Cognitel, which has been built to meet the needs of the small-office or home-office based professional, to perform some secretarial-type functions and effectively organize incoming telephone calls. In contrast to comparable voice-mail systems, it offers a high degree of customization of treatment for individual callers.

Cognitel answers incoming telephone calls using PC-standard voice-modem hardware and engages the caller in a dialogue using recorded prompts and speaker-independent discrete-word speech recognition. The caller is asked which recipient he/she wishes to speak with, and his/her own name. Using this information, Cognitel reacts to the caller according to predefined instructions. The owner of the system instructs it how to treat each of his or her expected callers in advance using a graphical user interface.

Cognitel supports a number of recipient mailboxes, typically one for each system user. Each recipient mailbox can have a list of known callers. These are the contacts of the system users, with whom Cognitel interacts on their behalf. The names of both callers and recipients must be trained in advance of incoming calls if they are to be recognized. Training is performed using a microphone attached to the Computer's sound card, assisted by the graphical user interface. Each recipient may configure the following actions for his/her callers:

- Play a pre-recorded message to the caller,
- Record a message from the caller,
- Call the recipient's pager after the call,
- Transfer the call to a different telephone number.

Combinations of these actions are also allowed, wherever possible. The paging and call-forwarding options require that suitable hardware and telephony services are available to the recipient.

Cognitel also includes a remote-access interface. Using this, a recipient can call into the system, and using spoken commands or DTMF telephone tones, he/she can navigate the list of messages, which have been received on his/her behalf. Entry to the remote-access interface is allowed after entry of a numerical password, using DTMF tones.

Messages recorded from callers are displayed in the graphical user interface, where the recipients can replay them. Each recorded message is stored with recordings of the caller's responses to the system dialogue. These samples may be used to add data to the word models in the system, or to add new word models, thus improving the speech recognition quality with use.

2. SPEECH RECOGNITION SUBSYSTEM

It is clear from the previous section that the main task of Cognitel's speech recognizer is spoken name recognition. The nature of personal names, particularly in North-American English, is such that the pronunciation is often not obvious from the written form. Moreover, names may originate from a wide variety of European, African, Asian or Hispanic backgrounds. This variety, and the corresponding scarcity of suitable training data, make HMM-based recognizers, by far the most commonly used today, unsuitable for the task of name recognition. More suitable, at least for relatively short lists of names, is pattern-based name recognition using a Dynamic Time Warping (DTW) algorithm for matching patterns.

The speech recognizer must also deal with several problems inherent in telephone-based speech recognition. When a caller interacts with an isolated word recognizer, he/she often adds some non-speech sounds (such as breathing, lip smacks etc.), so the DTW algorithm used must be capable of rejecting these sounds. Another problem is the variation in the electro-acoustical characteristics of the channel (microphone or telephone line) and the characteristic variations of the caller's voice. Channel independence is especially important, as training is normally performed using the computer's microphone, but recognition must be achieved from the telephone line. Methods of speech signal preprocessing, analysis and normalization which guarantee good speaker and channel independence are used to combat these difficulties.

To minimize the inconvenience to the users, only a small number of training samples can be requested for each name. A means of creating statistically significant word-patterns from a small number of samples was developed for this purpose.

Since it is expected that calls will be received from callers not known to the system, and it is not practical to train all possible names, a rejection technique for unknown caller names is used.

The configuration of the speech recognition module is shown in Figure 1.

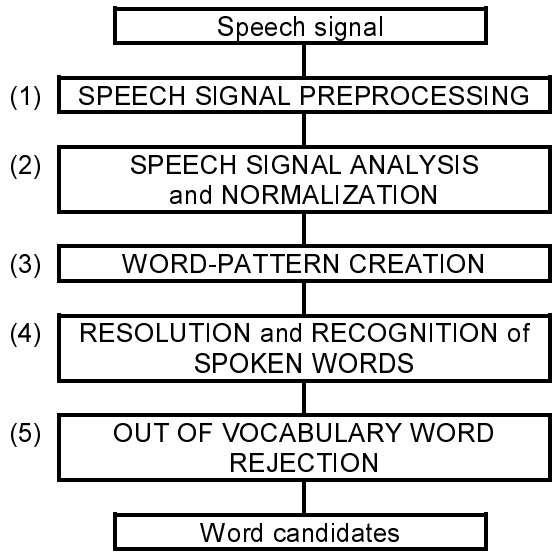


Fig.1. Block diagram of general algorithm of the speech recognition module

Speech Signal Preprocessing.

This stage attempts to identify the position of speech signals in an incoming audio stream. Speech is identified by its relatively high energy, i.e. a high amplitude and a long duration, and its modulated nature, i.e. comparatively rapid changing of amplitude with the syllable frequency. These characteristics are used to exclude non-speech signals such as short clicks and long stationary noise from the recording.

Speech signal analysis and parameter normalization.

Since a very small number of word samples is permissible for training, the speech signal analysis and parameterization must provide potentially maximal speaker and channel independence . Different types of speech signal analysis, including filter banks, cepstral and formant analysis, were tested [1]. It was found that formant analysis provided the best results with respect to speaker and channel independence. Formant analysis involves the extraction of the formant frequencies F1, F2, F3 and formant amplitudes A1, A2, A3 as functions of time. These six formant parameters and their corresponding first derivatives are taken as the input for name recognition. A normalization procedure is also carried out in order to compensate for different kinds of voice variations.

Word-Pattern Creation.

Initially the training samples of the word are time-aligned in nonlinear fashion by DTW. Quasi-statistical parameters of the word-pattern are then calculated. For each nth time-frame and for each pth formant parameter the quasi-statistical parameters of word sample distribution are determined by calculating the set of centers of gravity CG: CGc(p, n) - the center of the distribution, CGl(p, n) - the left deflection from the center, CGr(p, n) - the right deflection from the center. The advantage of this method is that training is permissible with a minimum of a single word sample. The procedure of word pattern creation is described in detail in [2].

Recognition and resolution of spoken words.

The algorithm of spoken word resolution and recognition in a running signal is based on an original modified dynamic programming method called Start-Point Free Dynamic Time Warping (SPF DTW). The theoretical ground of the algorithm was first given in [3]. Although the algorithm was developed long ago, it was not used as at the time as it seemed computationally too complex. With modern computer equipment it can however be used successfully. The main advantage of the SPF DTW is that it not only gives an evaluation of the word location in a continuous signal but also evaluates the time of its beginning and end. The algorithm is described in [4].

Out of vocabulary name rejection.

Cognitel is intended for use on a wide variety of voice modem hardware for which no suitable calibration is possible. This makes the use of threshold-based methods for rejecting bad recognition candidates difficult to use effectively. Instead, garbage models are mixed with the valid name candidates. The garbage models for first name recognition are made from the last names, and vice versa. If either first or last name candidates are one of the

garbage models, or if the first name-last name pair is not valid, the choice is rejected. These two methods proved to be effective in rejecting names not known to the system.

3. SPOKEN USER INTERFACE

The voice interface for incoming calls is shown schematically in Figure 1. Some of the stages may be omitted as shown, if the Caller ID signal from the telephone company is used instead of speech recognition.

If more than one recipient mailbox is in use, Cognitel first asks the user "Who do you want to speak to ?". If only one mailbox is set-up, it is assumed that this is the intended recipient. The dialogue for recipient identification is shown schematically in Figure 2. The dotted lines show that if no name can be identified after 2 attempts, a default recipient mailbox is assumed.

The caller identity is always required. Cognitel asks the caller "Who is calling, please ?" to prompt the caller for his/her name. The dialogue for caller identification is shown schematically in Figure 3. Unlike the recipient dialogue, there is no confirmation of caller names; These are assumed to be confidential. After two attempts the caller is assumed to be unknown or new.

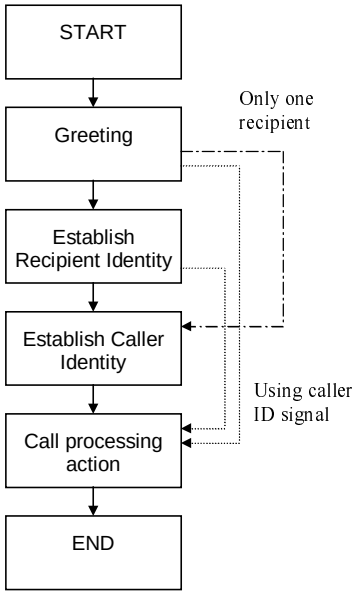


Figure 1: Caller interface

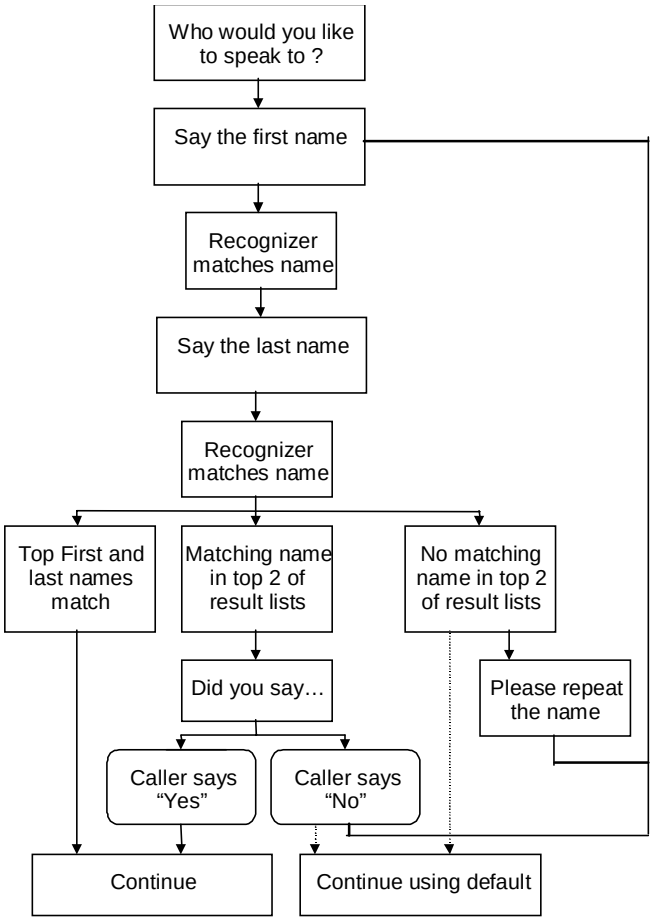


Figure 2: Recipient identification dialogue

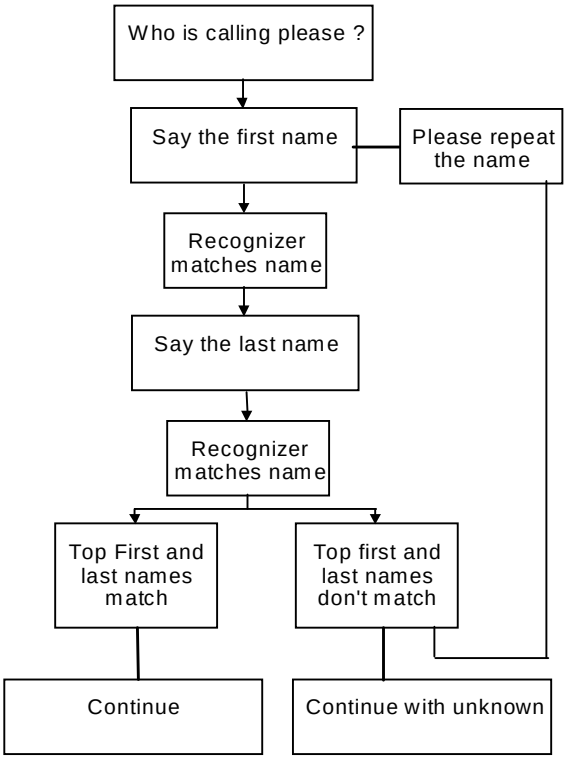


Figure 3: Caller identification dialogue

Both recipient and caller identification make use of the N-best results from the recognition stage to control the dialogue. For recipients, a name is considered valid if it is found in the top two of the N-best lists for first and last names. If the names are not top of both lists, the dialogue asks the user to confirm the choice.

Once the identity of the caller and recipient have been established, one of the "call processing actions" listed in section 1 is executed. The action to use for each recipient/caller pair is configured using a graphical user interface.

Each recipient mailbox has a "Default" action. This is the action which is carried out for callers for whom no special action is configured. There is also a "System default" action, which is used in cases where the recipient mailbox cannot be recognized. This simply plays a prompt, asking the caller to leave a voice message.

By pressing the "*" (STAR) key on the telephone keypad., and entering a numeric password, system users can listen to their messages remotely. This interface uses spoken commands such as "Go Forward", "Play Message" to navigate a list of messages. Other actions, such as retraining names and command words, and searching for messages from certain callers, are also included in the remote command set.

4. SUMMARY. ONGOING WORK

This paper has presented Cognitel, a telephone answering system which uses standard voice modem hardware to answer incoming calls, and identify the intended recipient of the call, and the name of the caller. If the recipient is not available to take the call, Cognitel interacts with the caller, according to pre-programmed instructions.

The use of speech recognition makes it possible for the system to know the identity of the caller, and so offer personalized treatment without caller and recipient having to agree in advance on mailbox entry codes or suchlike.

One obvious weakness of the system is that it is not secure; Simply identifying the spoken name does not guarantee that the caller is who he claims to be. It is intended for home or small-office use, however, where it is assumed that such matters will not present a significant risk.

The speech recognition system described here is essentially very simple, using DTW methods to achieve robust, real-time recognition over the phone. The training procedures used are simple and quick, so the inconvenience to the user is minimized. Further training can increase the quality of the speech recognition.

Cognitel is currently undergoing field trials, and as a result of these, some changes are to be expected. Likely future directions include the development of a connected word recognizer and syntax processor, and extension of the remote access interface to offer a wider range of configuration options.

With a connected word recognizer and the ability to evaluate syntax, many more forms of address will be supported, as well as nicknames for family members, for example.

A richer functionality for remote access would allow the call-processing actions to be defined and edited remotely, and messages to be forwarded to different locations, so that the system can be used effectively even if the user is away for a long period of time.

REFERENCES

1. Lobanov B.M., Levkovskaja T.V. Comparative investigation of two speech signal analysis methods. In *Digital image processing.* Proceedings of the Institute Engineering Cybernetics, Academy of Science of Belarus, Minsk, 1997, pp. 79-84.

2. Lobanov B.M., Levkovskaja T.V. An algorithm for speaker-independent word-pattern creation for speech recognition. In *Digital image processing.* Proceedings of the Institute Engineering Cybernetics, Academy of Science of Belarus, Minsk, 1997, pp. 147-153.

3. Lobanov B.M., Slucker G.S, Tizik A.P. Automatic Recognition of Sounds Combination in Running Speech Signal. (in Russian),*Trudy NIIR*, No 4, Moscow, 1969, pp.67-75.

4. Lobanov B.M., Levkovskaja T.V. Recognition of words and word sequences in running speech. In *Digital image processing.* Proceedings of the Institute Engineering Cybernetics, Academy of Science of Belarus, Minsk, 1997, pp. 154-161.