VOICE CONTROLLED MOBILE PHONE FOR CAR ENVIRONMENT

Ivan Bourmeyster[1], *Jamil Chaoui*[2], *Silvio Cucchi*[3], *Nicola Griggio*[3], *Alessandro Guido*[3], *Giuliano Moroni*[3], *Antonello Riccio*[3], *Marco Stanzani*[3], *Fabio Valente*[3]
[1]Alcatel Mobile Phones, [2]formerly at Alcatel Mobile Phones - 32, avenue Kléber, 92707 Colombes, France
Tel : +33 1 46 52 17 06 ; fax : +33 1 46 52 80 25
[3]Alcatel Corporate Research Centre - Via Trento, 30, 20059 Vimercate (Milano), Italy
Tel : +39 39 686 4077 ; fax : +39 39 686 3587

ABSTRACT

The development of an application of speech processing in a car environment is addressed. The main objective is to provide the user of a vehicular phone with a powerful and friendly bidirectional vocal interface. In particular, the paper focusses on the speech recogniser component of the interface as it was specifically designed and tuned to operate in the very hostile acoustic environment of a moving car.

The recogniser operates in a fully speaker dependent mode so enabling the user to store his/her personal agenda of frequent called parties.

For the training, three repetitions of each vocabulary word are recommended, although the performances remain still satisfactory with only two repetitions.

Reliable performance assessment was conducted with particular attention to the aspect of robustness of the recogniser against spurious noises. Standard procedures (SAM oriented) were used to guarantee the repeatability of any test. An outlook on future improvements is also given.

1 INTRODUCTION

Applications of speech recognition in the telecommunication environment are becoming quite popular nowadays thanks to the enhancement of basic algorithms and to the widely spread availability of powerful DSP platforms. Nevertheless, for some specific applications, even using well known and stable algorithms yet problems remain to be solved.

In this paper indeed the design and implementation of a DTW speech recogniser for car environment is described. The fundamental innovation lies in the front-end processing which has been proven to be extremely robust even in the very hostile acoustic environment envisaged for the final application.

Moreover the methodology adopted for the performance assessment and the interaction with the user have been thoroughly validated.

2 APPLICATION DESCRIPTION

It is commonly felt that normal usage of mobile phones in a car represents a potential cause of driver's lack of attention with obvious consequence in terms of safety. Hands-free mode is the solution adopted to reduce the potential risk at least during communications phases ; but in order to be really effective, the hands-free mode must be available to the user also to place a call : in this case speech recognition is mandatory.

In the intended application, the speech recogniser must handle isolated utterances and the operation has to be guaranteed both in quiet conditions and with car moving. Furthermore the speech recogniser must be trainable as far as the user is allowed to create his/her own personal agenda.

The solution adopted is to use a speaker dependent, isolated word speech recogniser to be trained by the user ; the vocabulary consists of digits, command words and names for the personal agenda. Moreover, for obvious safety reasons, the training phase must be done with the car stopped and therefore in an acoustic environment which is totally different when compared to the operating conditions. It means that, while reference templates are extracted from clean speech produced by a speaker in quiet conditions, during recognition very high car noise will corrupt the speech signal and will also affect the way the speaker will produce speech (Lombard effect).

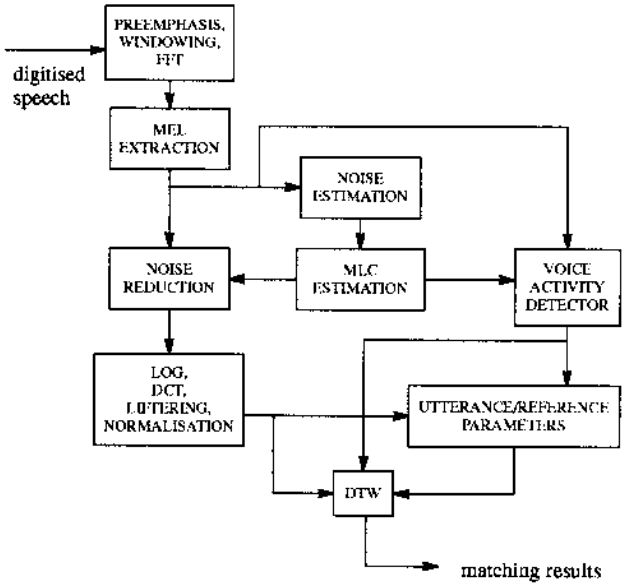

Figure 1 : Algorithm block diagram

3 ALGORITHM DESCRIPTION

The block diagram Figure 1 shows the main processing modules of the whole speech recogniser for both training and recognition phases.

3.1 Front-end processing

With reference to Figure 1, following steps are performed :

- Digitised speech, sampled at 8 kHz, is preemphasized and windowed.

- After FFT, a MEL extraction is performed. 30 coefficients over MEL scale are calculated from the square modulus of the 256 FFT coefficients [1].

- Noise estimation : noise estimation updating of each MEL band is in the form $N(t) = \alpha \times N(t-1)$, where α is function of signal to noise ratio $S(t) / N(t-1)$.

- Maximum Likelihood Coefficients (MLC) estimation : for each band a parameter is obtained as function of the signal to noise ratio of that band, in accord to the maximum likelihood function with soft decision scheme algorithm [2][3] :

$$mlc(i) = 0.5 \times \left(1 + \sqrt{\frac{S(i) - N(i)}{S(i)}}\right) \times P[H_1 \mid S(i)]$$

where $P[H_1 \mid S(i)]$ is the probability of speech, given the observed power $S(i)$ in the i-th band. $P[H_1 \mid S(i)]$ is a function of the signal to noise ratio too.
These parameters can be considered as a measure of the speech probability in each MEL band. They allow an integration of operation of VAD and noise suppression in the algorithm : in fact they are used both by the voice activity detector to determine the beginning and end of a word and by a module called *noise reduction* that modifies the MEL powers in accord to the MLCs. This modification reduces the noise components of the output spectrum and allows better matching between references and utterances.

- Voice Activity Detector : the MLC of each band is filtered in the time domain. It is used to determine the state (WAIT, STARTWORD or INWORD) of the associated mel-band by comparing it with a threshold (see Figure 2).

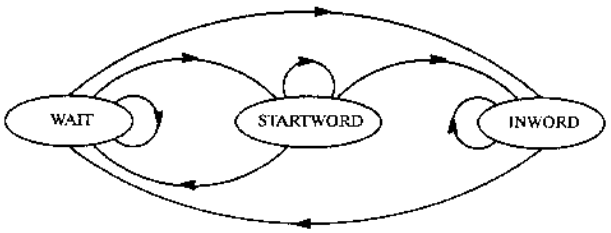

Figure 2 : MEL band state machine

Transition thresholds from a mel-band state to another one vary with the frequency and are periodically updated in accord to a long term estimation of the noise.

Transitions from a global state of the VAD to the following one are determined by the set of band states. There are 4 possible global states for the VAD : WAIT, STARTWORD, INWORD, MUC (see Figure 3). This state machine is used to determine the start and end point of a word. It has the ability of discarding a word and taking into account the Middle Utterance Closure (MUC), i.e. the consequence of a vocal-tract occlusion inside a word.

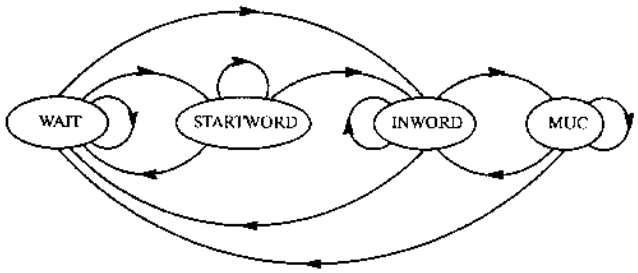

Figure 3 : Global VAD state machine

- Noise reduction : MLC factors are previously filtered in the frequency domain, then their squares are multiplied by the relative mel-band power.

- DCT : 12 coefficients from the 30 MEL coefficients are calculated and used by DTW for the matching between references and utterances.

- Liftering : a sine square liftering and a normalization are applied to the 12 DCT coefficients.

3.2 Dynamic Time Warping (DTW)

The Dynamic Time Warping is started and stopped by the VAD signal. Lower bound to the DTW path doesn't exist while the upper bound is of the form :

min { max [2*(*counter_frame* – *start_DTW*), *K*],
number of frames of the reference }

The score is the sum of the distances between utterances and references along Viterbi path, normalized by the modulus of utterance parameter frames.

A modified version of projective distance is used :

$$D = |utt| . \sqrt{1 - utt \times ref} \quad \text{if } \sqrt{1 - utt \times ref} < 1,$$
$$D = |utt| \quad \text{otherwise.}$$

The utterance scores are chosen evaluating the minimum value of the score, normalized by the utterance modulus, on the upper bound where the upper bound is equal to the number of frames in the reference. In this way minimum score can be placed before or after the VAD stop.

A weight that depends on the distance between the VAD end point and the frame where the minimum score occurs is

applied.

In the test phase, a given utterance, which is detected by the VAD, is used as input to the DTW algorithm and compared to all the references of a given vocabulary. The reference which gives the lowest score is chosen as the recognised word.

4 IMPLEMENTATION

The recogniser kernel has been implemented on the 16 bits fixed point digital signal processor core ST18950.

The kernel refers to an external file system able to manage the input/output : in detail, the references parameters (i.e. the parameters of the words to be recognised) consist of contiguous binary files whose pointers are passed to the kernel main functions onto stack.

The kernel main routine is supposed to be called by the man-machine interface when a recognition session starts : for each speech frame of the session, the parameters of the current utterance are provided, together with the partial best recognition scores and the VAD status.

Its operation can be roughly split into four steps :

a) Parameters extraction

First a preemphasis of the time-domain input samples is performed. The preemphasized samples are then submitted to a spectral analysis (256 points real DFT) : the output is the square moduli of the complex FFT of each frame. The frequency domain samples of each frame are then submitted to a MEL weighting and to a noise cancellation before passing to the cepstral analysis, suitable as the speech recognition algorithm input : at the output a set of (typically 12) normalized parameters is provided [4].

b) Noise estimation

Starting from a frequency domain noise estimation, a set of maximum likelihood coefficients is computed for each component of the MEL weighted signal : these coefficients are the input of both VAD (they are used to trigger the VAD module state machines) and parameters extraction (their square is used to cancel the noise from the MEL samples). Noise estimation is performed after the first 20 frames : during this time the noise buffer initilization is performed.

c) Voice Activity Detector (VAD)

Starting from the set of Maximum Likelihood Coefficients computed during the noise estimation process, the voice activity detection is performed : this task is devoted to trigger the speech recognition algorithm at proper time whenever an incoming word is estimated to arrive and to stop the recognition process whenever a pronounced word is estimated to end.

d) Speech Recognizer

Starting from an input database of templates (physically, a set of files descriptors provided by the MMI onto stack) and from the set of normalized parameters provided by the parameters extraction activity, the DTW algorithm progresses over the current set of references. After processing, the current three best scores of the complete database are available.

4.1 Trade off between time and memory needs

For better efficiency of implementation, DTW trellis is made progress for more than one frame (typically 4) with the same template : in this case the current template file is loaded (and unpacked) once for one session.

Supposing to directly access a reference file if the file is not fragmented – the time devoted to input is clearly the main bottleneck of the whole activity. The law for estimating the maximum number of reference words that can be processed in a single session by the recognizer kernel was studied and the results are shown in Figure 4.

Figure 4 indeed represents the degradation on the maximum number of templates that the present kernel is able to process, when passing from an average number of frame per word of 55 to 128 as a function of the number of instruction cycles required to process a single point of the DTW trellis : lines are plotted by using the typical values below. The required instruction cycle time of 25 ns has been slightly increased (up to 26.32 ns) in order to take into account background activities.

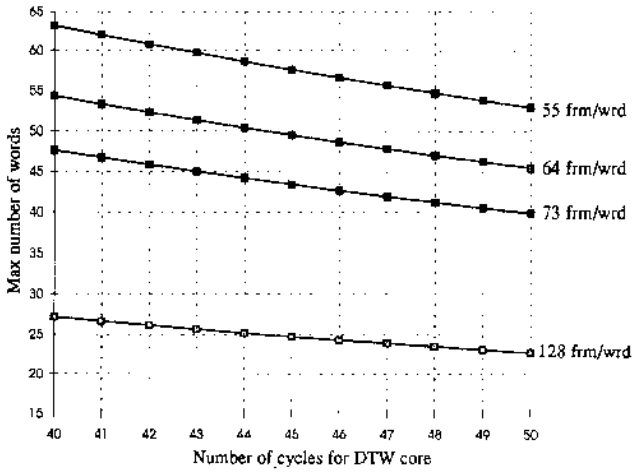

Figure 4 : DTW efficiency using different numbers of frames per word

The factors affecting the maximum number of templates are :

- the number of repetition of each word of the vocabulary (typically 3) ;
- the duration of a session (typically 60 ms, that is 4 frames of 20 ms with 25% overlap) ;
- the time required for executing the front end code (that is parameters extraction, End Point Detection, glue code, etc...) ;
- the number of instruction cycles to compute a point of

the DTW trellis ;
- the number of frames in a session (typically 4) ;
- the average number of frames per word.

5 ASSESSMENT METHODOLOGY

5.1 Data base

In order to build and assess the algorithms adopted, an appropriate database was recorded in the car environment. It consists of continuous recordings of isolated words divided into 5 groups ; the recordings contain commands, short names, long names (Christian names + family names), digits and spurious words uttered by different speakers at various speeds.

This database was generated in a SAM-like format in order to enable the adoption of all SAM standards, in particular to allow a repeatable and controllable analysis of the algorithms and of the results, i.e. a predictive diagnostic assessment. The recordings were made in a continuous way in order to test the recogniser even in the worst case when the recogniser is always active and false alarms (unwanted triggering of the recogniser due to spurious acoustic events, treated by the recogniser as a word) might be a strong performance degradation cause. This issue has to be taken particularly into account in the very noisy (with even negative SNR) car environment [5].

The duration of the isolated words database used for the tests is about 3 hours, with a total of nearly 3000 utterances. Assessment tests were conducted extensively on the available speech database (static tests) and also with real speakers in conjunction with an application simulator (dynamic tests).

5.2 Performance assessment

A correct recognition rate better than 95% was observed with the static database ; this figure is guaranteed independently of the type of speaker (male or female), the vocabulary (digits, commands, names), the car speed (from 0 to 130 km/h).

A complete recognition statistic with all the speakers in every conditions, which takes into account all the possible events occurring in the recogniser follows :
- Recognition : 95.8 %
- Substitutions : 2.6 %
- Complete Non Detection : 1.1 %
- Rejection : 0.5 %

where :
- a Substitution occurs when a wrong word is recognised
- a Complete Non Detection occurs when a word is not detected at all by the front end
- a Rejection occurs when a word is detected but discarded having a too high score (a fixed threshold is chosen).

With these results, a False Alarm rate (referred to the number of pronounced words) of 1.1% is obtained, i.e. the background noise present in the car causes 11 wrong recognitions every 1000 words when the temporal distance between two utterances of the database is at least 3 seconds and the recogniser is enabled all the time.

The robustness to background noise was tested with a background noise recording of the following type and duration :

Type	Duration	False Alarms / min
Radio	2h47	14.8
Traffic	2h20	1.4
Fan	2h01	0.06

6 FUTURE DEVELOPMENTS

Three main improvements of the recogniser are foreseen. Presently 3 templates per word are needed to guarantee good recognition performances, but a reduction to 2 templates and a parallel reduction in storage required for each of them will let more words to be used in the personal agenda.

Furthermore, in order to reach an even more hands-free and user-friendly usage, vocal activation and de-activation of the recogniser with the usage of a keyword are very important and a challenging target.

Finally, evolution towards speaker-independence will relieve the user of the initial training phase regarding the fixed words, i.e. the commands and the digits, which have presently to be trained to build the DTW templates. These fixed vocabularies should be stored in the form of HMM models, off-line built and ready to be used without a preparation phase done by the user. The personal agenda vocabulary should still be implemented with DTW templates thus leading to a mixed DTW-HMM solution.

REFERENCES

[1] A. Brancaccio, F. Ceglie, G. d'Acunzo, C. Pelaez, A. Riccio, F. Rigosi, "A comparative study of the influence of parameter processing on two different approaches for speech recognition in adverse environment", Proceedings of ESCA Workshop on Speech Processing in Adverse Conditions, 1992.

[2] R. J. McAulay, M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", vol. ASSP-28, no. 2, pp. 137-145, April 1980.

[3] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems", vol. II, pp. 363-366, ICASSP proceedings 1993.

[4] A. Brancaccio, C. Pelaez, "Experiments on Noise Reduction Techniques with Robust Voice Detectors in a Car Environment", Eurospeech 93, vol. 2, pp. 1259-1262.

[5] Johan Smolders, Tom Claes, Gert Sablon and Dirk Van Compernolle, "On the importance of the microphone position for speech recognition in the car", Proceedings of ICASSP 1994, vol. 1, pp.429-432.