

A 1000 WORD SPEECH RECOGNITION SYSTEM USING A SPECIAL PURPOSE CMOS-PROCESSOR.

W. Drews*, R. Laroia*, J. Pandel*, A. Stoelzle*

ABSTRACT.

This paper describes a speech recognition system for speaker dependent isolated word recognition. The system consists essentially of two processing boards: a feature extraction board using a TMS320C25 signal processor, and a pattern matching board involving a special purpose CMOS-processor capable of matching with up to 1000 words in real time. A 80186 microprocessor, also contained on the pattern matching board, provides control and data transfer between the speech recognition system and an arbitrary host computer.

I. INTRODUCTION.

Speech recognition and speech synthesis can be considered as special fields in digital signal processing which have attracted increasing interest in the last years. Speech recognition systems can be characterized by the permitted type of speech, i.e., isolated words or continuous speech, and by their degree of speaker dependence. Among these options, to date only the problem of speaker dependent isolated word recognition seems to be sufficiently well understood for a large vocabulary size and therefore amenable to practical applications. Thereby, a word by word comparison between stored reference words and a spoken unknown word has to be performed on the basis of so-called feature vectors, which can be considered as a discrete-time digital representation of speech.

Several possibilities exist for the choice of these feature vectors.[1]. The most important requirements are elimination of redundancy from the speech while preserving distinctive features of similarly sounding words, and the existence of a computationally simple distance metric. In the presented speech recognition system we use a spectral representation of the speech in terms of short-time log-converted energies within fifteen frequency bands. Every 20 ms a feature vector is yielded from a digital filter bank, which is implemented on a TMS320C25 signal processor. During a training phase, the feature vectors of the reference words are stored in a template file within the host computer, which must be downloaded to the template memory of the pattern matching board before starting the recognition phase (Fig. 1). During the recognition phase, the word distance between a spoken unknown word and a reference word is calculated by either the accumulated Euclidean or city-block distance (programmable) of their feature vectors. This task is performed using the so-called dynamic-time-warp algorithm (DTW-algorithm) [2], which provides a time alignment between words of different duration and number of feature vectors, respectively. Since this algorithm requires a very high computation rate for a large vocabulary size and real time processing, a dedicated CMOS-processor was designed for this purpose. Finally, the index of the reference word with the smallest accumulated distance to the unknown one is sent to the host computer, which prints the corresponding word and may execute a related program.

II. FEATURE EXTRACTION FILTER BANK.

In order to determine the spectral energy distribution of the speech signal, a digital filter bank was designed for splitting the speech signal into several frequency bands. The sampling rate at the input should be $F = 12.8$ kHz, allowing for higher frequency components to be analyzed. Due to the frequency characteristic of the human ear, the analysis bandwidth should increase towards higher frequencies ("critical band" theory [3]).

*Siemens Central Research and Development Laboratories, ZFE ME 22, Otto-Hahn-Ring 6, 8000 Munich 83, Germany.

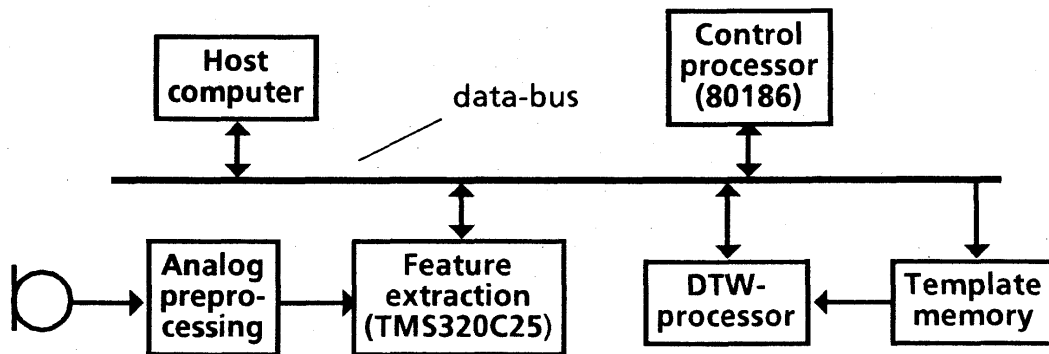


Fig. 1: Functional block diagram of the speech recognition system.

The straightforward solution of implementing 15 separate band-pass filters is not realizable with a general purpose signal processor due to the relatively high input sampling rate, which in this case would have to be equal to the operating rate of all band-pass filters. A frequency analysis by a sliding discrete Fourier transform or a polyphase filter bank [4,5] has also not been found useful, because these methods provide a linear spacing of analysis bands, so that the widths of all analysis bands would be determined by the smallest required bandwidth in the lower frequency range, which would result in an unnecessary large number of analysis bands and a corresponding large size of a fast Fourier transform.

Therefore, a multistage filter bank with a tree structure involving half-band filters [5] has been chosen, which allows a successive subdivision of analysis bands by factors of 2 respectively, along with a sampling rate decrease of the same factor. These half-band filters, basically arrangements of two filters having mirror-image symmetry about one quarter of the operating rate, can be efficiently realized as wave digital filters [6] yielding an optimum dynamic range and easily maintainable forced response stability [7,8]. Fig. 2 shows the basic structure of each half-band filter, where the sampling rate decrease is performed via an alternating switch and y_1 and y_2 denote the low-pass and high-pass filtered output signals, respectively. S_1 and S_2 denote all-pass circuits which can be realized as a cascade of wave digital all-pass sections.

Due to the losslessness of the all-pass branches in Fig. 2, the property of energy conservation holds for this filter structure. It can be shown that the signal energies

$$W_x = \sum_n x^2(nT) \quad W_{y_1} = \sum_m y_1^2(2mT) \quad W_{y_2} = \sum_m y_2^2(2mT) \quad (1)$$

are related by

$$W_{y_1} + W_{y_2} = \frac{1}{2}W_x \quad (2)$$

which means that the total energy of the input signal $x(nT)$ is retained in the output signals y_1 and y_2 except of a constant scaling factor. This property, which is independent of the actual filter coefficients, results in the fact, that the sum of the energy samples of all feature vectors reflect the total energy of the speech signal, avoiding any frequency gap or duplicate covering of any frequency interval.

The required filter bank is easily constructed by connecting additional half-band filters to both y_1 - and y_2 -outputs of the first filter and continuing this procedure with the four newly obtained output signals. Thereby, the mentioned tree structure of the filter bank is realized, where after each stage a doubling of the number of analysis bands along with a halving of the respective bandwidths and sampling rates occurs. Though, this splitting of frequency bands is not uniformly continued over the whole speech spectrum, which would otherwise result in a uniform and therefore less efficient filter bank, as has been mentioned. Instead of this, after a sufficient frequency resolution no further halving of bands is performed, depending on the respective

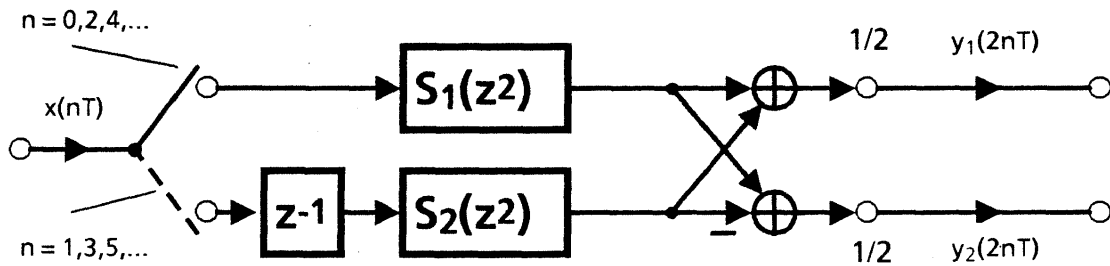


Fig. 2: Basic structure of a wave digital half-band filter with sampling rate decrease by 2.

frequency range. Table 1 shows the so obtained 40 analysis bands, which offer sufficient flexibility for a collection of 15 components forming the feature vectors. These components are calculated for 15 characteristic frequency bands (Tab. 2) by squaring, weighting, and adding up the output signals of the underlying analysis bands, averaging of the so obtained energy signals over about 30 ms, and log-conversion according the A-law.

frequency range/Hz	analysis bandwidth/Hz	number of channels
0-800	50	16
800-1600	100	8
1600-3200	200	8
3200-6400	400	8

Tab. 1: Analysis bands obtained from multi-stage filter bank.

channel	frequency range/Hz	bandwidth/Hz
1	50-200	150
2	200-350	150
3	350-500	150
4	500-650	150
5	650-800	150
6	800-1000	200
7	1000-1200	200
8	1200-1500	300
9	1500-1800	300
10	1800-2200	400
11	2200-2600	400
12	2600-3200	600
13	3200-4000	800
14	4000-4800	800
15	4800-6000	1200

Tab. 2: Analysis bands corresponding to the components of the feature vectors.

Besides the calculation of the feature vectors, the determination of the beginning and the end of a word is carried out on the signal processor. Thereby, the algorithm described in [9], after some modifications, has been used, which is based on the observation of the short-time average magnitude function and the short-time zero crossing rate of the speech waveform.

III. DYNAMIC-TIME-WARP PROCESSOR.

A full-custom integrated CMOS-processor [10] has been designed for calculating the vector distances between the feature vectors of a spoken unknown word and all reference words stored in a template memory (Fig. 1). Two alternative distance metrics have been implemented in this processor. Let $\mathbf{U}(i) = (U_1(i), \dots, U_{15}(i))^T$ (T denotes transposition) be a vector of the unknown word for the i -th time slot, and let $\mathbf{T}(j) = (T_1(j), \dots, T_{15}(j))^T$ be a vector of a specific reference word for the j -th time slot, then the distance between $\mathbf{U}(i)$ and $\mathbf{T}(j)$ can be calculated according

$$d_{i,j} = \sum_{n=1}^{15} (U_n(i) - T_n(j))^2 \quad \text{or} \quad d_{i,j} = K_s(i,j) \sum_{n=1}^{15} |U_n(i) - T_n(j)| \quad (3a, b)$$

where K_s in (3b) designates a programmable scaling factor which may vary for different time slots depending on special properties of the respective speech segments. Fig. 3 shows the functional

block diagram of the data path for the distance computation, which operates with a clock rate of 10 MHz and represents with 80 MOPS the computational most intensive part of the chip.

Following the computation of the d_{ij} -terms the DTW-algorithm attempts to find a warp path within the (i,j) -plane of time slots, for which the accumulated values d_{ij} yield a minimum (Fig. 4). The accumulation is recursively computed according

$$D_{i,j} = \min(D_{i-1,j}, D_{i-1,j-1}, D_{i,j-1}) + d_{i,j} \text{ for } i=1, \dots, n; \quad j=1, \dots, m$$

$$\text{with } D_{0,j} \Big|_{j=1, \dots, m} = D_{i,0} \Big|_{i=1, \dots, n} = \infty; \quad D_{0,0} = 0 \quad (4)$$

Thereby, n and m are the total numbers of the feature vectors of the two compared words, so that $D_{n,m}$ yields the desired word distance. In order to compute the word distances in real time, i.e., all word distances with respect to the stored reference words should be evaluated after a short response time following the end of the unknown word, the computation of the D_{ij} has been realized in a column order [10,11]. This requires, in view of (4), the storing of all values D_{ij} of the last column in an external scratch-pad memory.

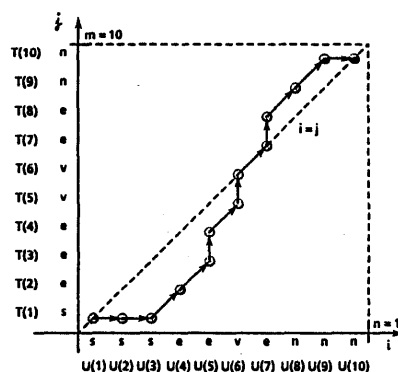
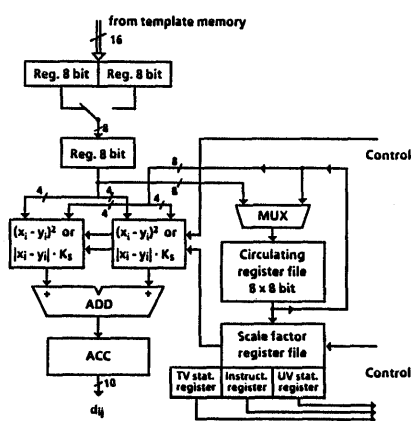


Fig. 3: Data path for distance computation.

Fig. 4: Example of a warp path between two words.

External control of the processor is provided by an additional 16th component attached to each feature vector. This component contains the state information to be stored in a special state register after entering the processor. Internal control is performed by finite state machines.

The processor contains about 10500 transistors on an area of 17,6 mm² including 40 pads [10] and was fabricated with a 2 μ -CMOS technology at Siemens.

References

- [1] B.A. Dautrich, L.R. Rabiner, T.B. Matin, "On the effects of varying filter bank parameters on isolated word recognition", IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-31, 1983, pp. 793-807.
- [2] H. Sakoe, S. Chiba, "A dynamic programming approach to continuous speech recognition", Proc. 7th Int. Conf. on Acoustics, Budapest, 1971, pp. 65-68.
- [3] B. Scharf, "Critical Bands", in Foundations of modern auditory theory, edited by J.N. Tobias, (Academic, New York).
- [4] L.R. Rabiner, R.W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Englewood Cliffs, New Jersey 07632.
- [5] R.E. Crochiere, L.R. Rabiner, "Multirate Digital Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey
- [6] W. Wegener, "Wave digital directional filters with reduced number of multipliers and adders", Archiv für Elektronik und Übertragungstechnik, vol. 33, 1979, pp. 239-243.
- [7] A. Fettweis, K. Meerkötter, "Suppression of parasitic oscillations in wave digital filters", IEEE Trans. Circuits and Systems, CAS-22, 1975, pp. 239-246.
- [8] L. Gazsi, "Explicit formulas for lattice wave digital filters", IEEE Trans. Circuits and Systems, CAS-32, 1985, pp. 68-88.
- [9] L.R. Rabiner, M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances", Bell System Technical Journal, vol. 54, 1975, pp. 297-315.
- [10] W. Drews, L. Laroia, J. Pandel, A. Schumacher, A. Stölzle, "A CMOS-processor for a 1000 word speech recognition system", Proc. IEEE Custom Integrated Circuits Conf., Portland 1987, pp. 559-562.
- [11] R.A. Kavalier, M. Lowy, Hy Murveit, R.W. Broderson, "A dynamic-time-warp integrated circuit for a 1000-word speech recognition system. IEEE Journal of Solid-State Circuits, SC-22, 1987, pp. 3-14.