Consumer and Corporate Affairs Canada | Consommation et Corporations Canada

(11) (A) No. 1 232 686

(45) ISSUED 880209

(52) CLASS 354-54

(51) INT. CL.[4] G10L 9/02

(19) (CA) **CANADIAN PATENT** (12)

(54) Speech Recognition

(72) Lennig, Matthew;
Mermelstein, Paul;
Gupta, Vishwa N.,
Canada

(73) Granted to Northern Telecom Limited
Canada

(21) APPLICATION No. 473,198

(22) FILED 850130

No. OF CLAIMS 20

Canada

DISTRIBUTED BY THE PATENT OFFICE, OTTAWA
CCA 274 (11-82)

The invention relates to speech recognition.

In known speech recognizers the speech, coded in pulse code modulation (PCM) format, is pre-processed to render it in a form that is more closely related to the way that the human auditory system perceives speech. For example, the speech may be processed to give filter bank energies, cepstra, mel-frequency cepstra, or linear prediction coefficients. Recognition units, for example, words, or syllables, are then compared with each of a series of reference templates representing valid units. The template that is the closest match is deemed to be the unknown unit and the label of the unit corresponding to the template is provided at the output.

Although such recognizers are adequate for certain applications, they are not entirely satisfactory because they give an error rate that is unacceptable in some applications, especially speaker-independent, telephone-based, or large vocabulary applications. This is thought to be because the usual representation does not model sufficiently the response of the human auditory system.

According to the present invention, apparatus for recognizing speech comprises:-

(i) means for representing an unknown speech utterance as a sequence of parameter frames, each parameter frame representing a corresponding time frame of said utterance;

(ii) means for providing a plurality of reference templates, each comprising a sequence of parameter frames expressed in the same kind of parameters as the first-mentioned

parameter frames

each parameter frame of the first-mentioned (unknown) sequence and second-mentioned (reference) sequence comprising a set of primary parameters and a set of secondary parameters, each secondary parameter representing the signed difference between corresponding primary parameters in respective parameter frames derived for different time frames; and

(iii) means for comparing the sequence of parameter frames of the unknown utterance with each reference template and determining which of the reference templates most nearly resembles it.

Each parameter frame comprises a set of parameters selected according to the type of representation employed, for example filter bank energies, cepstra, mel-based cepstra or linear prediction coefficients.

Preferably, the time difference between centres of said different time frames is from 20 mS to 200 mS, preferably about 50 mS. Conveniently, the secondary parameter is derived from preceding and succeeding primary parameters, for example ±25 milliseconds or ± two frames.

It is also preferable to include a component representing change in amplitude or change in perceptual loudness as a secondary parameter for both the unknown utterance and the reference templates. Such a loudness component is not usually used in the primary parameters since absolute amplitude or absolute loudness is not effective in distinguishing words.

Generally, then, the innovation consists of

augmenting the set of primary short-time static parameters normally used for speech recognition with a set of dynamic secondary parameters representing change in each of the primary parameters over a short time interval (for example, 20 to 200 mS). Use of dynamic parameters in addition to primary parameters renders the distance measure or probability density function used to distinguish speech sounds more sensitive to important phonemic differences as opposed to other, irrelevant, acoustic differences.

Any kind of short-time spectral representation may be used as the set of primary parameters. Examples of such representations include filter bank energies, the cepstrum, the mel-frequency cepstrum, linear prediction coefficients, etc. Each of these representations estimates the magnitude or power spectrum over a time frame (typically between 2 and 50 mS) in terms of a small number of parameters (typically between 3 and 80).

If P_t is the vector of primary parameters computed at time t, time offsets a and b are chosen such that:-

$$20 \text{ mS} \leq a + b \leq 200 \text{ mS}$$

the dynamic parameter vector ΔP_t is defined to be the vector difference

$$\Delta P_t = P_{t+a} - P_{t-b}$$

The invention consists of using the ensemble of parameters P_t together with ΔP_t to represent the speech signal in the neighbourhood of time t. Probability density functions and distances are then defined in terms of this augmented parameter set consisting of both static (primary) and dynamic (secondary) parameters.

Alternatively, the above derivation may be expressed in terms of frame numbers. If Δt = the time difference between adjacent frames and if P_i = the primary parameter vector at frame i, then the dynamic parameter vector ΔP_i is defined as the vector difference

$$\Delta P_i = P_{i+\left\lfloor \frac{a}{\Delta t} \right\rfloor} - P_{i-\left\lfloor \frac{b}{\Delta t} \right\rfloor}$$

Preferably the parameters are mel-based cepstral coefficients in which case the primary coefficients $C_1,\ldots,C_n$ represent the spectral shape and the secondary parameters $\Delta C_1,\ldots,\Delta C_m$ represent change in spectral shape during the specified time interval. In addition, ΔC_0 may be included in the set of secondary parameters to represent change in loudness or amplitude.

An embodiment of the invention will now be described by way of example only and with reference to the accompanying drawings, in which:-

Figure 1 is a generalized block diagram of a speech recognizer; and

Figure 2 is a diagram representing the characteristics of a filter means of the speech recognizer.

In the speech recognition system illustrated in Figure 1, signal S_n represents a linear pulse-code-modulated (PCM) speech signal, which is the unknown or "input" utterance to be recognized. Signal S_n is applied to window means 10. In the window means 10, the signal S_n is divided into time frames, each of 25.6 milliseconds or 204 samples duration. In operation, each

frame is advanced by 12.8 milliseconds or 102 samples so that successive frames overlap by 50 per cent. Each time frame is then multiplied point-by-point by a raised cosine function and applied to filter means 12. This Hamming window attenuates spectral sidelobes.

A 256 point Fast Fourier Transform is performed on each time frame and results in a 128 point real power spectrum, $F_1,\ldots,F_N$, where N=128.

The filter means 12 effectively comprises a filter bank of twenty triangular filters, which determine the energy in a corresponding set of twenty channels spanning the range from about 100 Hz to about 4000 Hz for a PCM sampling rate f_s of 8 KHz. As illustrated in Figure 2, the channels are mel-spaced, with channel centre frequencies spaced linearly from 100 Hz to 1000 Hz at 100 Hz intervals and logarithmically from 1100 Hz to 4000 Hz.

For each time frame, the output of each filter channel is a weighted B_j derived in accordance with the expression:-

$$B_j = \sum_{i=1}^{N} W_{ij} F_i$$

where B_j is the jth mel-frequency channel energy output, F_i are the N spectral magnitudes $1 \leq i \leq N$ from the Fast Fourier Transform, and the W_{ij} are weights defined as:

$$W_{ij} = \begin{cases} 0, & i\Delta f \leq l_j \\ (i\Delta f - l_j)/(k_i - l_j), & l_j \leq i\Delta f \leq k_j \\ (h_j - i\Delta f)/(h_j - k_j), & k_j \leq i\Delta f \leq h_j \\ 0, & i\Delta f \geq h_j \end{cases}$$

for $1 \leq i \leq N$ and $1 \leq j \leq 20$

where $\Delta f = \frac{f_s}{2N}$

and where l_j, k_j, h_j for $1 \leq j \leq 20$ are the low, center, and high frequencies, respectively of each filter channel, given in Table 1.

The twenty log channel energies of the signal B_j are computed in means 14 according to the expression:-

$$L_j = \log_{10} B_j \text{ for } 1 \leq j \leq 20.$$

The outputs of the filter means and the means 14 are applied to means 16 for computing, respectively, perceptual loudness C_0, and the first seven mel-based cepstral coefficients C_1, C_2, ... C_7.

The perceptual loudness C_0 is the log of a perceptually weighted sum of the channel energies B_j obtained thus:

$$C_0 = 600 \log_{10} \sum_{j=1}^{20} v_j B_j$$

where $v_j \geq 0$ are chosen to correspond to perceptual importance. Suitable values for v_j are illustrated in Table 1 below.

FILTER NO.(j)	l_j Hz	k_j Hz	h_j Hz	LOUDNESS WEIGHT v_j
1,	0.,	100.,	200.	.0016
2,	100.,	200.,	300.	.0256
3,	200.,	300.,	400.	.1296
4,	300.,	400.,	500.	.4096
5,	400.,	500.,	600.	1.
6,	500.,	600.,	700.	1.
7,	600.,	700.,	800.	1.
8,	700.,	800.,	900.	1.
9,	800.,	900.,	1000.	1.
10,	900.,	1000.,	1150.	1.
11,	1000.,	1150.,	1320.	1.
12,	1150.,	1320.,	1520.	1.
13,	1320.,	1520.,	1750.	1.
14,	1520.,	1750.,	2000.	1.
15,	1750.,	2000.,	2300.	1.
16,	2000.,	2300.,	2640.	1.
17,	2300.,	2640.,	3040.	1.
18,	2640.,	3040.,	3500.	1.
19,	3040.,	3500.,	4000.	1.
20,	3500.,	4000.,	4600.	1.

TABLE 1.

The means 16 for obtaining the cepstral coefficients C_i functions by taking the cosine transform of the log energies, thus:-

$$C_i = \sum_{j=1}^{20} L_j \cos\left[\frac{i(j-\frac{1}{2})\pi}{20}\right]$$

where $1 \leq i \leq 7$.

For further information on computing the coefficients, the reader is directed to a paper by S.B. Davis and P. Mermelstein entitled "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics and Signal Processing, Vol. ASSP 28 No. 4 pp. 357-366 August 1980.

The output of means 16, which comprises the set of primary parameters $C_1,\ldots,C_7$ and the perceptually weighted loudness parameter C_0, is passed, every 12.8 milliseconds, to utterance endpoint detector 18. The word endpoints are detected by searching for minima of sufficient duration and depth in the perceptual loudness C_0 as a function of time frame number. Endpoint detection may be by one of various known methods, for example as disclosed in "An Improved Endpoint Detector for Isolated Word Recognition", L.F. Lamel, L.R. Rabiner, A.E. Rosenberg and J.G. Wilpon, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-29, No. 4, August 1981, p.777-785.

Thereafter the interword intervals or silences are removed, i.e., only the intervening utterance is transmitted. Optionally the number of parameter frames per utterance M may be

standardized, for example at M=32, by linearly spaced frame deletion or repetition.

The output of the endpoint detector 18 is a sequence of M mel-based cepstra, and is represented by the matrix:-

$$U = \begin{bmatrix} C_{1,0} & \cdots & C_{1,7} \\ \vdots & & \vdots \\ C_{M,0} & \cdots & C_{M,7} \end{bmatrix}$$

This output signal, or recognition unit representation, U, is applied to dynamic parameter computing means 20 which computes the dynamic parameters as:-

$$\Delta C_{i,j} = C_{i+c,j} - C_{i-d,j}$$

for $d+1 \leq i \leq M-c$, $0 \leq j \leq 7$,

where c is the leading frame separation, d is the lagging frame separation. In the specific case, c=d=2.

For $1 \leq i < d+1$

$\Delta C_{i,j} = C_{i+c,j} - C_{1,j}$; and

For $M-c < i \leq M$

$\Delta C_{i,j} = C_{M,j} - C_{i-d,j}$

These dynamic parameters take account of the human auditory system's propensity for perceiving change in the incoming stimulus.

The sequence of M parameter frames U' comprising primary (static) and secondary (dynamic) parameters, represented by the matrix:-

$$U' = \begin{matrix} C_{1,1},\ldots,C_{1,7}, & \Delta C_{1,0},\ldots,\Delta C_{1,7} \\ \vdots & \vdots \\ C_{M,1},\ldots,C_{M,7}, & \Delta C_{M,0},\ldots,\Delta C_{M,7} \end{matrix}$$

is applied from dynamics computing means 20 to dynamic time warping means 22.

It should be noted that the C_0 vectors relating to static loudness are not used.

A corresponding set of templates, including dynamic parameters and a dynamic loudness component $\Delta T_{.,0}$ is derived by means 24 of the form:-

$$T' = \begin{matrix} T_{1,1},\ldots,T_{1,7}, & \Delta T_{1,0},\ldots,\Delta T_{1,7} \\ \vdots & \vdots \\ T_{32,1},\ldots,T_{32,7}, & \Delta T_{32,0},\ldots,\Delta T_{32,7} \end{matrix}$$

The sequence of parameters for the templates is also applied to dynamic time warping means 22.

The "unknown" parametric representation U' is compared with each of the reference templates T' in turn and the time warp distance computed in each case. The unknown utterance is deemed to be the reference utterance corresponding to the template having the minimum warp distance.

The dynamic time warp computation may be as described by Hunt, Lennig and Mermelstein in a chapter entitled "Use of Dynamic Programming in a Syllable-Based Continuous Speech

Recognition System" in Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, D. Sankoff and J.B. Kruskal, eds. Addison-Wesley (Reading MA), pp. 163-187, 1983.

It has been found that a significant improvement in recognition accuracy is obtained by including dynamic parameters in the parametric representation. It has also been found that particularly good results are obtained when the dynamic parameters represent the change in the speech signal between time frames spaced by about 50 milliseconds.

Although described as applied to isolated word recognizers, the invention is also applicable to connected word recognizers and is also useful whether the recognizer is speaker-trained or speaker-independent.

Experimental results using the parameter set augmented with dynamic parameters as described above, in which the primary parameters were the first seven mel-frequency cepstral coefficients, resulted in about 20% reduction in recognition errors in speaker-independent connected digit recognition over the public switched telephone network. In addition to using dynamic parameters corresponding to the first seven mel-based cepstral coefficients, the eighth dynamic parameter corresponding to change in overall loudness further reduced errors by about 10%.

THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE PROPERTY OR PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS:

1. A method of recognizing an unknown speech utterance comprising the steps of:-

(i) representing said unknown speech utterance as a sequence of parameter frames, each parameter frame representing a corresponding time frame of said utterance;

(ii) providing a plurality of reference templates, each comprising a sequence of parameter frames expressed in the same kind of parameters as the first-mentioned parameter frames;

each parameter frame of the first-mentioned sequence and second-mentioned sequences of parameters comprising a set of primary parameters and a set of secondary parameters, each secondary parameter representing the signed difference between corresponding primary parameters in respective parameter frames derived for different time frames; and

(iii) comparing the sequence of parameter frames of the unknown utterance with each reference template and determining which of the reference templates most closely resembles the unknown utterance.

2. A method as defined in claim 1, wherein the time between the centers of different time frames is in the range of 20 to 200 milliseconds.

3. A method as defined in claim 2, wherein said time

is about 50 milliseconds.

4. A method as defined in claim 1, 2 or 3, including the step of computing a dynamic loudness component as a secondary parameter, and providing a corresponding dynamic loudness component in each of said parameter frames.

5. Apparatus for recognizing an unknown speech utterance in a speech signal comprising:-

(i) means for representing an unknown speech utterance as a sequence of parameter frames, each parameter frame representing a corresponding time frame of said utterance;

(ii) means for providing a plurality of reference templates, each comprising a sequence of parameter frames expressed in the same kind of parameters as the first-mentioned parameter frames

each parameter frame of the first-mentioned sequence and second-mentioned sequence of parameter frames comprising a set of primary parameters and a set of secondary parameters each secondary parameter representing the signed difference between corresponding primary parameters in respective parameter frames derived for different time frames; and

(iii) means for comparing the sequence parameter frames of the utterance with each reference template and determining which of the reference templates most nearly resembles the unknown utterance.

6. Apparatus as defined in claim 5, wherein said means for providing provides each said secondary parameter to represent the signed difference between primary parameters in respective parameter frames derived for time frames that are spaced by a time interval in the range of 20 to 200 milliseconds.

7. Apparatus as defined in claim 6, wherein the time frames are spaced by about 50 milliseconds centre-to-centre.

8. Apparatus as defined in claim 5, 6 or 7, comprising means for computing for both said unknown utterance sequence and said template sequence a dynamic loudness component as one of the set of secondary parameters.

9. Apparatus as defined in claim 5, wherein the means for providing includes means for computing the secondary parameters $\Delta C_{i,j}$ in accordance with the expression:-

$$\Delta C_{i,j} = C_{i+c,j} - C_{i-d,j}$$

for $d+1 \leq i \leq M-c$, $0 \leq j \leq 7$

where c is the leading frame separation and d is the lagging frame separation, both relative to the frame for which the dynamic parameter is being determined.

10. A method as defined in claim 1, wherein the secondary parameters are computed in accordance with the expression:-

$$\Delta C_{i,j} = C_{i+c,j} - C_{i-d,j}$$

for $d+1 \leq i \leq M-c$, $0 \leq j \leq 7$

where c is the leading frame separation and d is the lagging frame separation, both relative to the frame for which the dynamic parameter is being determined.

11. Apparatus as defined in claim 9, wherein said means for computing the secondary parameters does so in accordance with the expression:-

$$\Delta C_{i,j} = C_{i+c,j} - C_{1,j}$$
for $1 \leq i < d+1$

and in accordance with the expression:-

$$\Delta C_{i,j} = C_{M,j} - C_{i-d,j}$$
for $M-c < i \leq M$

12. A method as defined in claim 10, wherein the secondary parameters are computed in accordance with the expression:-

$$\Delta C_{i,j} = C_{i+c,j} - C_{1,j}$$
for $1 \leq i < d+1$

and in accordance with the expression:-

$$C_{i,j} = C_{M,j} - C_{i-d,j}$$
for $M-c < i \leq M$

13. A method as defined in claim 1, wherein neither of said different time frames comprises the time frame for which the corresponding primary parameter was derived.

14. A method as defined in claim 13, wherein said different time frames precede and succeed, respectively, the time frame for which the corresponding primary parameter was derived.

15. A method as defined in claim 14, wherein said different time frames are not consecutive with said time frame for which the corresponding primary parameter was derived.

16. A method as defined in claim 1, wherein said different time frames are not consecutive with said time frame for which the corresponding primary parameter was derived.

17. Apparatus as defined in claim 5, wherein neither of said different time frames comprises the time frame for which the corresponding primary parameter was derived.

18. Apparatus as defined in claim 17, wherein said different time frames precede and succeed, respectively, the time frame for which the corresponding primary parameter was derived.

19. Apparatus as defined in claim 18, wherein said different time frames are not consecutive with said time frame for which the corresponding primary parameter was derived.

20. Apparatus as defined in claim 5, wherein said different time frames are not consecutive with said time frame for which the corresponding primary parameter was derived.

2-1

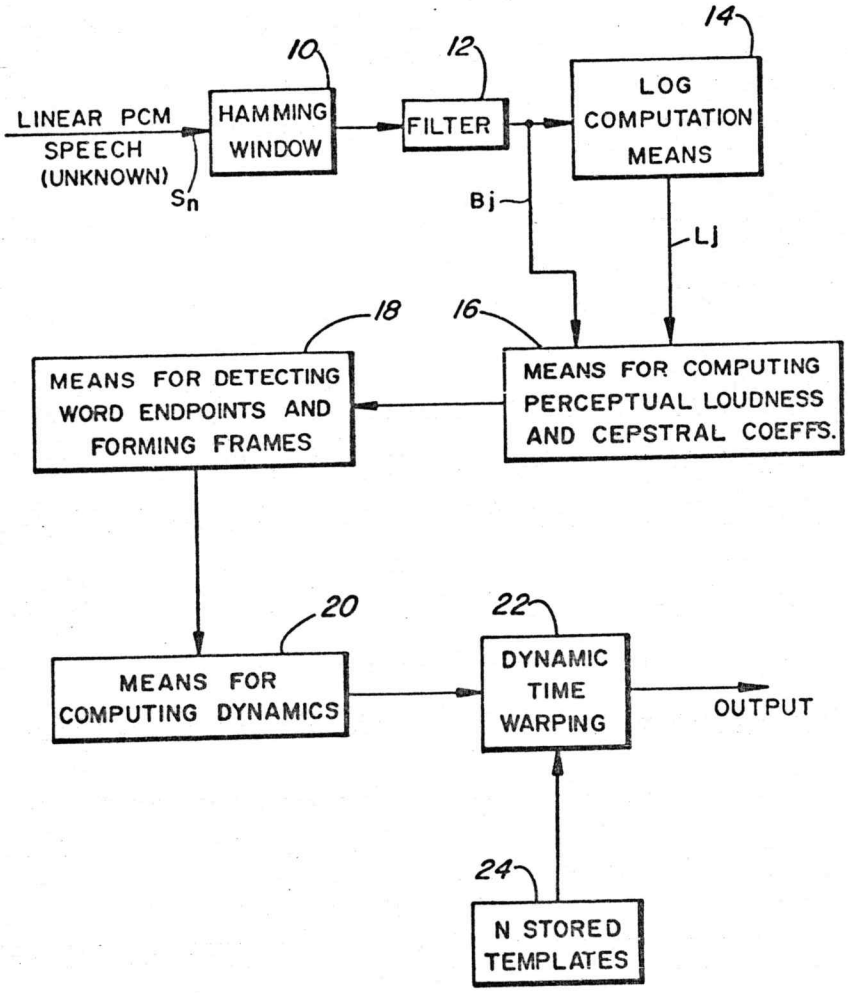

FIG. 1

J. Adams
Agent.

2-2

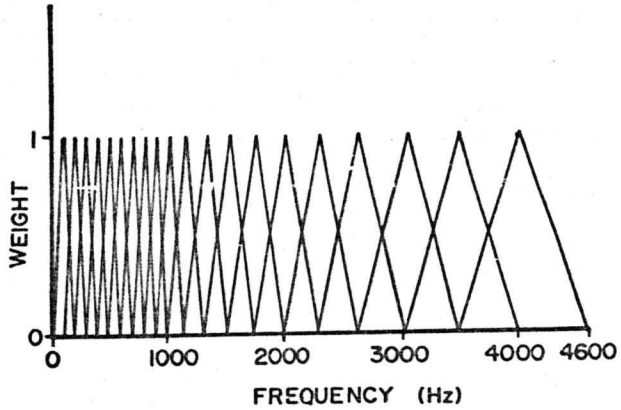

FIG. 2a

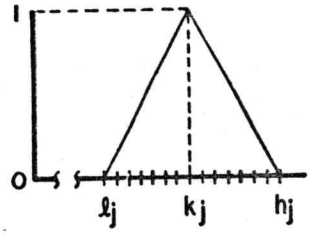

FIG. 2b

J. Addams
Agent.