Journal of Computer Science (Special Issue): 84-88, 2005
ISSN 1549-3636

Speaker Recognition Using Spectral Cross-correlation: A Fast Algorithm

Zoubir Hamici
Faculty of Electrical Engineering, Amman University
Amman University Post Office, P.O. Box, 129, Sweileh 19328, Jordan

Abstract: This study presents an original algorithm for computing the cross-correlation function applied for speech recognition A spectral correlation estimation algorithm based on the comparing the magnitude spectrum of the two signals is presented. The number of samples is reduced by a factor of two, after eliminating the image spectrum. A moving average filter is used to smooth the magnitude spectrum and a re-sampling is performed in the frequency domain, which reduces the spectrum size, by a factor of 8. The algorithm shows good results in recognizing the voice of a specific person, hence its application in speaker identification.

Key words: Speaker recognition, spectral cross-correlation, fast algorithm

INTRODUCTION

Speech recognition has been an active field of research during the last three decades. The rapid pace of business today requires employees and customers to have fast, constant access to information. Technology has provided a mechanism for efficient communication and information storage-but it has also introduced new levels of complexity into the business environment. Long learning curves impact productivity and complex interfaces make software difficult to use. Speech-enabled interfaces to computers can help solve important business problems and improve sensitive area security systems. Speech-enabled applications can help reduce the training costs of rapidly changing software products by providing a more intuitive user interface, allowing users to substitute complex drop-down menus commands with simple spoken commands. Customer service departments can provide customers with automated access to information and services since speech-enabled applications can eliminate the constraints of the telephone keypad and allow easier-to-use-automated systems.

Because speech is a learned function, any interference with learning ability may be expected to cause speech impairment. The most common interfering conditions are certain neuroses and psychoses, mental retardation and brain damage, whether congenital or acquired. Voice disorders, so-called dysphonias, may be the product of disease or accidents that affect the larynx. They may also be caused by such physical anomalies as incomplete development or other congenital defect of the vocal cords. Disorders of rate and rhythm are generally either psychogenic or have a basis in some neurological disturbance. Hence speaker-independent speech recognition system must have a robust algorithm to identify a special phrase pronounced by the speaker and permits the access to specific location or the use of a given device. A computer application was developed which activates a hardware device based on the result of recognition (Fig. 1). The predefined-recorded phrase is compared in real-time with a speech-recorded online from a microphone placed in front of the device to be controlled.

Many algorithms based on Cepstral analysis or homomorphic analysis for automated speech recognition are used in research[1-3]. The hidden Markov model (HMM) is one of the most frequently used methods[4]. HMM uses a statistical modeling and libraries of words and grammar rules to select the highest probability outcome from a sequence of samples. The cepstral analysis supplanted the direct use of linear prediction analysis LP, derived from the hidden Markov modeling, other works has implemented a person identification system based on acoustic and visual features[5].

Speaker recognition or person identification is a process that automatically authenticates a personal identity based on his or her voice. Although speaker recognition includes diverse tasks that discriminate people in terms of their voices, most of studies focus on speaker identification and verification. a speaker recognition system often works in either of two operating modes: text-dependent and text-independent. By text-dependent, the same or known text is used for training and test. The algorithm presented in this study is based on text-dependent speaker recognition.

TIME AND SPECTRAL CROSS-CORRELATION

Correlation Function is used to obtain the similarity between the Pass-Phrase and the person speech. The cross-correlation function is given by:

Corresponding Author: Zoubir Hamici, Ph.D., Faculty of Electrical Engineering, Amman University, Amman University Post Office, P.O. Box, 129, Sweileh 19328, Jordan

$$r_{xy}(n) = \sum_{k=0}^{2.(N-1)} x(k).y(k-n) \qquad (1)$$

The cross-correlation function result depends on the recorded speech strength and the duration of recording. To perform cross-correlation independently of scale and duration the normalized cross-correlation is used. The recognition results are included in the interval[0-100].

$$\rho_{xy}(n) = \frac{\sum_{k=0}^{2.(N-1)} x(k).y(k-n)}{\sqrt{E_x.E_y}}.100 \qquad (2)$$

Where

$$r_{yy}(0) = \sum_{k=0}^{N-1} y(k).y(k) = \sum_{k=0}^{N-1} |y(k)|^2 = E_y \qquad (3)$$

And

$$r_{xx}(0) = \sum_{k=0}^{N-1} x(k).x(k) = \sum_{k=0}^{N-1} |x(k)|^2 = E_x \qquad (4)$$

r_{xx} and r_{yy} are the energies of *x(n)* and *y(n)* respectively.

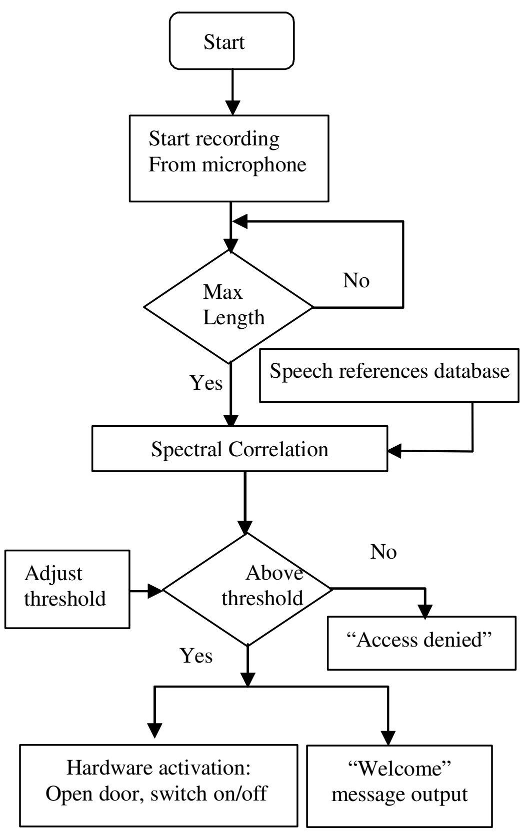


Fig. 1: Flowchart of the speech recognition system

The comparison of the signals in the time domain is not suitable for speech signals, because of the statistical behaviour of these signals.

$$Y_F(f) = \sum_{k=0}^{L-1} \frac{1}{L}.Y_M(f-k) \qquad (5)$$

This is a Moving Average filter (MA), with a Finite Impulse Response (FIR) system. L is chosen to be 10, YM is the magnitude spectrum and YF is the filtered Magnitude spectrum.

Spectral cross-correlation algorithm

Step 1: Compute FFTs of windowed *x(n)* and *y(n)*
Step 2: Eliminate the upper side of each spectrum.
Step 3: Smooth the spectrum using the MA filter.
Step 4: Resample the magnitude spectrum *X(f)* and *Y(f)* at each 4Hz (e.g.1/8 sample is taken)
Step 5: Computer the spectral cross-correlation S(f)
Step 6: Find the maximum of *S(0)*.
Step 7: if *S(0)* > Threshold then
Recognition procedure
Else
Recognition failure

When processing 2 seconds of speech digitized at 8KHz, this represents 16000 samples. The FFT is computed using a radix-2 on 16384 samples, this means that 8192 samples represents the actual spectrum and the other 8192 sample represents the image spectrum which is eliminated. The actual spectrum has a resolution of 0.5 Hz (e.g. the inverse of the signal duration), which is in reduced to 4 Hz. This means that the new spectrum contains only 1024 samples after re-sampling. Therefore the correlation function has a length of 2048 samples instead of 32768 samples in the time domain. The gain is then 16 in term of correlation size. This method is more appropriate for person identification or speaker-dependent speech recognition. The spectral analysis is based on calculating the cross-correlation between the FFT's of the two signals. $S_{XY}(f)$ is the normalized cross-correlation in the frequency domain, where |*X(f)*| and |*Y(f)*| are the magnitudes of *X(f)* and *Y(f)* respectively. The phase of the spectrum is ignored and has no significant interest. The Discrete Fourrier transform DFT is obtained by:

$$\tilde{X}(f) = \sum_{n=0}^{N-1} \tilde{x}(n).e^{-j.2.\pi.f.n} \qquad (6)$$

We can reduce sidelobe leakage by selecting windows that have low sidelobes. Hence, each signal is multiplied by a Blackman-Harris window before calculating the corresponding FFT (Fig. 1). This window gives a -92dB attenuation of the peak sidelobe. The N defines the length of the window.

$$\tilde{x}(n) = x(n).w_{BH}(n) \qquad (7)$$

Where

$$w_{BH}(n) = a_0 - a_1.\cos\frac{2.\pi.n}{N} + a_2.\cos\frac{4.\pi.n}{N} - a_3.\cos\frac{6.\pi.n}{N}$$

a_0=0.35875; a_1=0.48829; a_2=0.14128; a_3=0.01168

The actual spectrum is given by:

$$\tilde{X}(f) = X(f) * W_{BH}(f) \qquad (8)$$

Similarly

$$\tilde{Y}(f) = Y(f) * W_{BH}(f) \qquad (9)$$

The spectral correlation and the similarity criterion are given by:

$$S_{XY}(f) = \sum_{k=0}^{2.(N-1)} \left|\tilde{X}(k)\right| . \left|\tilde{Y}(k-f)\right| \qquad (10)$$

$$S(f) = \frac{S_{XY}(f)}{\sqrt{S_{XX}(0).S_{YY}(0)}} .100 \qquad (11)$$

Where

$$S_{XX}(0) = \sum_{k=0}^{N-1} \left|\tilde{X}(k)\right|^2 = E_x \qquad (12)$$

And $$S_{YY}(0) = \sum_{k=0}^{N-1} \left|\tilde{Y}(k)\right|^2 = E_y \qquad (13)$$

$S_{XX}(0)$ and $S_{YY}(0)$ are the energies of $x(n)$ and $y(n)$ respectively. S= $S(0)$ defines the similarity criterion.

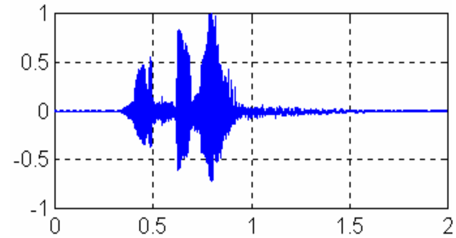
RESULTS AND DISCUSSION

We implemented a new recognition algorithm based on spectral cross-correlation. The first step was the programming of a cross-correlation between the reference signal and the pre-recorded (target) voice in the time domain Fig. 2. The second method is the use of a spectral correlation algorithm Fig. 3. The spectrum is estimated on 1024 samples. Table 1 gives the comparison of the similarity criterion result depending on the signal to noise ratio SNR.

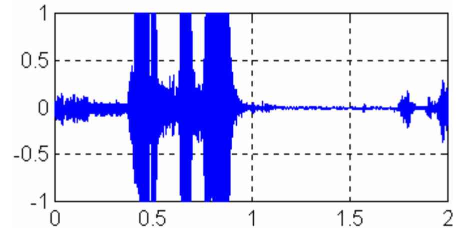
Table 1: Recognition experiments result

Algorithm	Test	SNR	S_{MAX}
Time Correlation using	Fig.2-d	20dB	20%
32000 samples (N=2x8KHz)	Fig.2-e	6dB	10%
	Fig.2-f	3dB	9%
Spectral Correlation using	Fig.3-c	20dB	82%
full spectrum:16384 samples	Fig.3-d	6dB	60%
	Fig.3-e	3dB	51%
Spectral Correlation (SC)	Fig.4-a	20dB	91%
with averaged 2048 samples	Fig.4-b	6dB	84%
	Fig.4-c	3dB	78%
SC with woman voice SC	Fig.5-a	20dB	77%
with Child voice	Fig.5-b	20dB	78%
Recognition failure voices:	Recognition Treshold=75%		
SC with woman voice SC	Fig.6-a	20dB	68%
with Child voice	Fig.6-b	20dB	49%

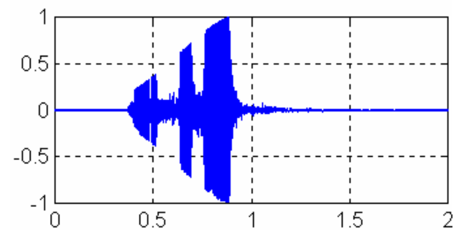
Figure 2 shows two signals, which are almost the same, but the time-domain cross-correlation gives a similarity of less than 20% due to the statistical behaviour of speech signals. The same signals are compared in term of spectral content using magnitude spectrum correlation, the result shows a similarity of 91%. Another advantage of using the spectral correlation is the computation time, which is the soul of a real-time process. As a comparison the time-domain cross-correlation of 16384 sample for each signal, takes 2 seconds, however the calculation of the two FFT and the spectral correlation is done in less than 0.1 seconds, therefore the second method is three times faster for 2 seconds of speech processing. The number of samples in the frequency domain is reduced by a factor of two when eliminating the image of the spectrum (i.e. the upper half part of the spectrum).



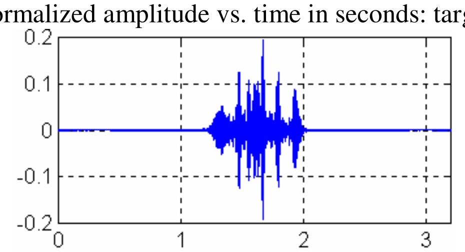
(a) Normalized amplitude vs. time in seconds: reference



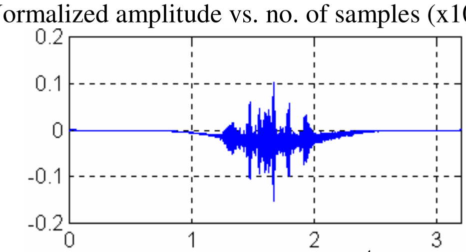
(b) Normalized amplitude vs. time in seconds: target



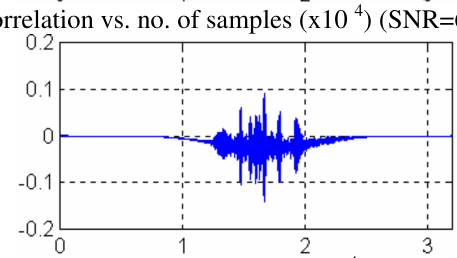
(c) Normalized amplitude vs. time in seconds: target



(d) Normalized amplitude vs. no. of samples (x10 ^{4})



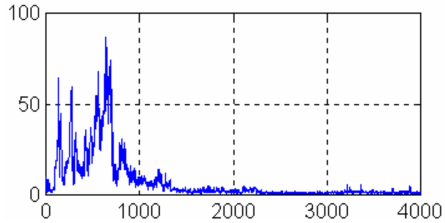
(e) Correlation vs. no. of samples (x10 ^{4}) (SNR=6dB)



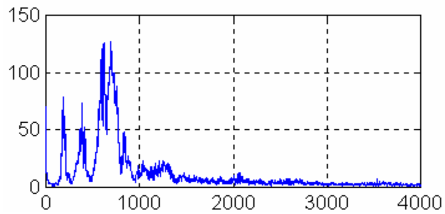
(f) Correlation vs. no. of samples (x10 ^{4}) (SNR=3dB)

Fig. 2: Time-domain cross-correlation: (a) reference voice (man), (b) target voice of the same man, (c) filtered target voice using Blackman-Harris window which eliminates the abrupt cutoff in the rectangular window, the voice contains background speech and

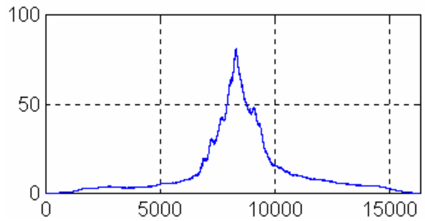
the SNR=20dB (d), (e), (f) cross-correlation function at SNR of 20dB, 6dB, 3dB respectively



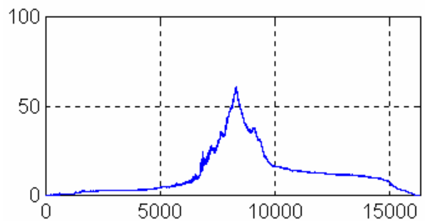
(a) Magnitude spectrum versus frequency (Hz)



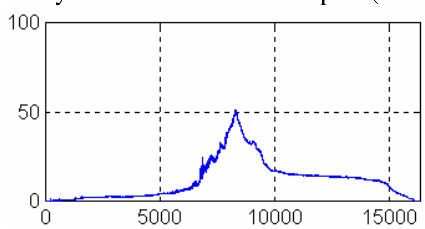
(b) Magnitude spectrum versus frequency (Hz)



(c) Similarity criterion vs. no. of samples (SNR=20 dB)



(d) Similarity criterion vs. no. of samples (SNR=6 dB)

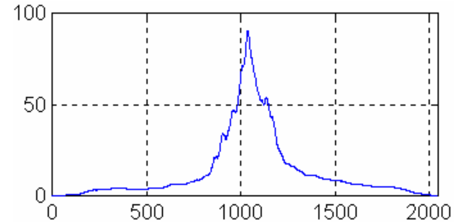


(e) Similarity criterion vs. no. of samples (SNR=3 dB)

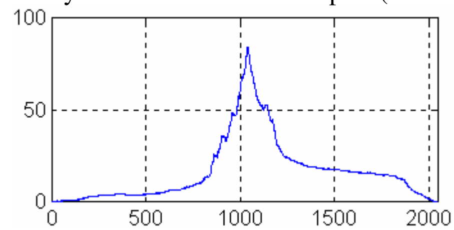
Fig. 3: Spectral contents cross-correlation: (a) reference magnitude spectrum, (b) target voice magnitude spectrum. (c), (d), (e) spectral correlation on 16384 samples with SNR 20dB, 6dB and 3dB respectively

The spectrum is smoothed by a moving average (MA) filter given by the difference equation (5), then the spectrum is resampled each 4 Hz. This results in a new frequency resolution instead of 0.5 Hz.

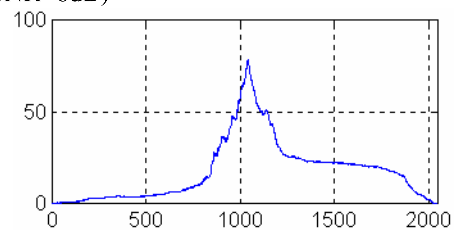
The filtered spectral content recognition shows robustness to noise. Another Main feature is that, the spectral correlation and we only need to compute one value of the spectral cross-correlation S(0). The overall results obtained show that time-domain correlation is not suitable for speech recognition.



(a) Similarity criterion vs. no. of samples (SNR=20dB)

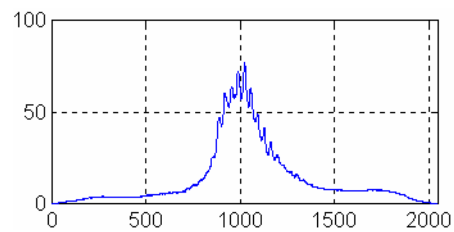


(b) Similarity criterion versus no. of samples (SNR=6dB)

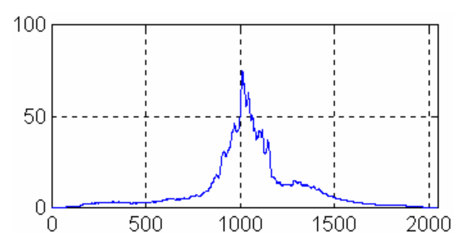


(c) Similarity criterion versus no. of samples (SNR=3dB)

Fig. 4: Spectral content cross-correlation using 2048 samples, of the same person: (a) SNR=20dB, (b) SNR=6dB, (c) SNR=3dB



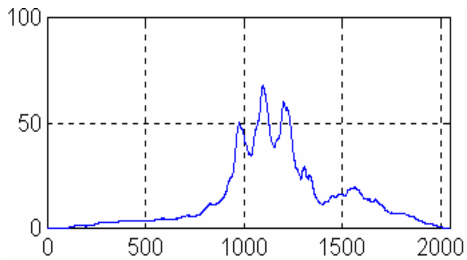
(a) Similarity criterion versus no. of samples (woman)



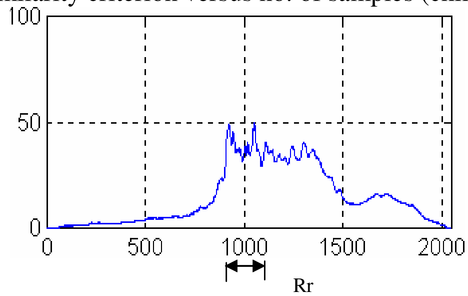
(b) Similarity criterion versus no. of samples (child)

Fig. 5: Spectral content cross-correlation using 2048 samples, SNR=20dB: (a) reference voice and a woman voice, (c) reference voice and a child voice for the same pass-phrase

Spectral cross-correlation creterion is an efficient too for recognizing the speech of a specific person and still capable of recognizing the same short-phrase or word pronounced by a different person.



(a) Similarity criterion versus no. of samples (child)



(b) Similarity criterion versus no. of samples (woman)

Fig. 6: Spectral content cross-correlation using 2048 samples on a different pass-phrase with SNR=20dB: (a), (b) recognition failure of child and woman voices for a different phrase respectively

(Fig. 4 and 5). A recognition failure test is done on a different voice short-phrase pronounced by a woman and a child Fig .6. The spectral recognition algorithm computation time is 0.80 seconds done on a personal computer with celeron processor running at 550 MHz CPU clock speed. This time includes the computation of two FFT's, magnitude spectrum, moving average filtering of the spectrum, cross-correlation and calculation of the similarity criterion. We can speedup the recognition algorithm by calculating the spectral correlation in a reduced range (Rr) of 5% (Fig. 6-b). since the similarity creterion is defined at S(0). Experements shows that only S(0) is necessary, hence a great reduction in the computation requirements.

CONCLUSION

Recently, speaker verification has been increasingly demanded for security in miscellaneous information systems[6-8]. This study introduced a modified cross-correlation computation algorithm applied for person identification used for a access control security system. The presented algorithm is used to compare the spectral content of a reference signal (short-sentence or word) with a pre-recorded signal. The results show a great improvement in the recognition process computation speed. This algorithm applied on short words or on segmented sentences, it is more appropriate for person identification or text-dependent speaker recognition.

REFERENCES

1. Aikawa, K. and K. Ishizuka, 2002. Noise-robust speech recognition using a new spectral estimation method phasor. ICASSP2002, pp: 397-400.
2. Ringger, E. and J. Allen, 1996. Error correction via a post-processor for continuous speech recognition. Proc. ICASSP-96, May 7-10, Atlanta.
3. Picone, J. *et al.*, 1993. Signal modeling techniques in speech recognition. IEEE Proc. Automatic Speech Recognition and Understanding Workshop, 81: 1215-1247.
4. Povlow, B. and S. Dunn, 1995. Texture classification using noncausal hidden markov models. IEEE Trans. Pattern Analysis and Machine Intelligence, 17: 1010-1014.
5. Bruneli, R. and D. Falavigna, 1995. Person identification using muliple cues. IEEE Trans. Pattern Analysis and Machine Intelligence, 17: 955-966.
6. Furui, S., Recent advances in speaker recognition. Pattern Recognition Lett., 18: 859 -872.
7. Furui, S., 1981. Cepstral analysis technique for automatic verification. IEEE Trans. Acoustics Speech Signal Process, 29: 254-272.
8. Chen, K., 2003. Towards better making a decision in speaker verification. Pattern Recognition, 36: 329 -346.