YET MORE ON SPEECH SPLICING

Robert E. Bogner and Radin B. Ikram

Department of Electrical and Electronic Engineering
The University of Adelaide.

ABSTRACT - Speech waveforms are being cut and spliced to effect time scale modifications without pitch distortion, for data compression or time scale change. Suitable instants at which to effect the cutting and splicing are found by use of multi-dimensional representations of the waveforms. Residual splicing errors are reduced by tapering the abutting sections.

1. INTRODUCTION

The nearly repetitive nature of the speech waveform offers possibilities of time scale modification without distortion of perceived pitch by the repetition or deletion of sections of the waveform that are a pitch period in duration. Gabor (1947) suggested the possibilities of data compression in this manner, following his studies of distortions produced by an optical playback device with a moving head. The latter, following ideas patented in the 1930's, produced significant distortion associated with the resultant cyclic mismatch. Fig. 1 shows such a system in principle, and fig. 2 shows the resultant discontinuity of the waveform produced by currently available devices.

The principle of time scale modification by splicing of suitable epochs illustrated in fig.1 (b) and (c) shows the desirability of ensuring that the abutted ends of the sections joined do match each other properly, to avoid a discontinuity.

A variety of timescale modifiers have followed, represented by (a) pitch-independent time-domain systems for example: Fairbanks 1954, Koch 1972, Schiffman 1974,

(b) pitch-dependent time domain systems, eg.: Jones 1971, Seo 1974, Suzuki 1976, Neuberg 1978, Malah 1979, Beddoes 1982,

(c) frequency-domain systems, eg: Flanagan and Golden 1966, and Portnoff 1981, the latter being pitch-dependent.

There are other compression schemes for efficient communication or storage, such as linear prediction vocoders, and adaptive transform vocoders. However, there are potential advantages in the combination of time scale compression with other waveform coding. Thus reduction of redundancy through reduction of repetitive sections may be combined with the reduction attainable through modelling of the waveform within a section.

The present work relates to the determination of appropriate instants on the speech waveform at which the wave may be cut and abutted to a similar section chosen from another epoch. For stretching of the timescale, ie when repetition of a section is required, the length of the section is determined such that the ends match suitably. We use the term "cut point"

to denote the latest point on a section of waveform to be retained, and "splice point" to denote the first point on the section of waveform to be joined after the cut point. In the case of a perfectly repetitive waveform with pitch period P such splicing points would occur at instants P apart; in fact any points P apart would be equally good.

2. CHOICE OF SPLICE POINT

2.1 Principle

At a splice point the signal should have a value close to that of the corresponding cut point. Correctness of the value is not sufficient to ensure unambiguous choice of splice point. Other aspects of the waveforms should match, for example the slopes and second derivatives might be suitable to ensure rejection of unsuitable candidate splice points. The splice point might then be chosen to minimise a measure of the dissimilarity of the cut point and the candidates for splice point, by examination of a distance measure or metric m,

$$m = d^2 = (x_C - x_S)^2 + (y_C - y_S)^2 + (z_C - z_S)^2 + \ldots \quad (1)$$

where d is the euclidean distance, and the subscripts C and S refer to cut and splice points respectively. For example, for a sinewave signal x a plot of its derivative dx/dt = y against x results in an ellipse, which becomes a circle at one frequency. It is desirable that each dimension should contribute equally and independently to the metric, and thus a selective function like the derivative is not optimum.

If the dimensions of the signal taken were:

$$y(n) = x(n-1); \quad z(n) = x(n-2); \ldots, \quad (2)$$

with a large number of components, then the choice of the minimum value of m corresponds with the choice of the peak of the autocorrelation function. Thus, we are in effect seeking to make a high efficiency autocorrelation ptich detector by choosing the efficient dimensions.

The Hilbert transform (HT) of a signal has power equal to that of the original signal, but is maximally uncorrelated with the signal. The HT of a sinewave is a cosinewave of the same amplitude, and thus the sinewave signal x and its HT y would always plot as a circle, and the x and y dimensions would contribute equally to the metric. Fig.3 shows these two dimensions of a fairly simple compound signal, and we see how the two equivalent waveform points plot to the same point in the xy plane. Fig. 4 shows the result for an example of a speech waveform. There are still possible ambiguities associated with crossovers in the diagram.

Introduction of another dimension z makes it possible to resolve many such potential ambiguities. Visualisation is aided by considering an untidy coil of garden hose lying on the ground. There are many crossovers. We can eliminate many of the points at which the hose touches itself by lifting the coils apart, in the z (altitude) direction. We have found that the addition of another two dimensions, derived from x by filtering as described below, provide sufficient discrimination from erroneous choices of splice points.

In essence the procedure is:

(1) Select a cut point fairly arbitrarily, and store the x, y, and z etc. coordinates.

(2) Examine the subsequent x, y and z values until a sufficiently similar set of the N is found, ie m is small enough. An example of a speech waveform and the corresponding metric behavior are shown in fig. 5.

(3) Proceed until a minimum of m is found, and tag this point as a splice point (in the research, we kept track of the set of such points by a file of their sample reference numbers, called "pointers").

(4) Record the splice point reference as a new cut point, and repeat the four steps for the next splice point.

2.2 Criteria and choice of the dimensions.

The desirable features of the filters for deriving the several dimensions y and z (etc) from the original signal are:

(1) the prime dimension should be the signal itself, as discontinuities in the waveform value are directly perceptible.

(2) each should contribute about equally to the metric;

(3) attention should be focussed on the region of the splice. Thus, the impulse response of each should be relatively brief and be substantially time-symmetric;

(4) each should be computed via an economical convolution.

From these considerations we settled on: the signal,x; an approximate 300 to 4500 Hz HT of order 15 for y; for z, an all pass filter, and for the fourth dimension, w, approximation to the first eigenvector of the covariance matrix of speech, of length 4. Impulse responses are shown in fig.6 for y and w.

2.3 Blending of abutments.

While earlier work showed promising results, occasional crackles were evident. We attributed them to the inevitable mismatches that must occur when the waveform is not periodic or has a period not equal to a multiple of the sampling interval. A scarf joint, made by linearly tapering contributions from the end of the waveform at the cut point, and from the beginning of the waveform at the splice point (fig. 7) was used to remove any sharp discontinuities. A graded overlap of 10 samples was found to be perceptually as effective as any greater length.

3. PRACTICAL ASPECTS

To ensure that a splice point was not chosen in the immediate vicinity of a cut, a guard algorithm was used; the search did not start until 4 ms after a cut. Then, m was continually compared with a threshold that expanded gradually, to ensure that even if the speech did not repeat, some

splice point would eventually be chosen.

For research purposes we needed to make a flexible and efficient system. The speech was digitised with 12-bit precision, and held in disc files. The additional dimensions y and z were formed by convolutions conditioned to retain time registration between x, y, and z, and held on additional files. The three files containing x, y, and z were then processed together to find suitable splice pointers. An arbitrary starting point was chosen, with x at some moderately small value in its range. Then the procedure described in Section ... was followed to determine suitable points for splicing or cutting. Such points were listed by their sample numbers in a file of "pointers" for future use.

The actual production of the compressed or expanded speech was then effected by selection of sections of the speech waveform from the x file, by reference to the file of pointers, and to a key that specified the operation required. The key was a 10 element code whose elements corresponded to the successive sections of the marked original speech, and determined the number of appearances of the corresponding section of the speech. For example, the key 1010101010 specifies a 2:1 compression; it indicates that section 1 is to be used once; section 2 deleted, section 3 used once, and so on. A 1:2 expansion is specified by key 2222222222 , and so on.

In a practical hardware implementation, the filtering, metric computation, and selection of splice points would be effected in real time without file keeping.

4. OBSERVATIONS AND CONCLUSIONS

Speech time scale modification by proper splicing is a simple, efficient means of compression or expansion. The resultant speech has little perceptible deterioration for 2:1:2 compression-expansion, but for larger expansions, fricatives become buzzy. Tapering has been found necessary to remove some residual crackles, but tapering beyond 5 samples from the splice point is not advantageous.

5. ACKNOWLEDGEMENTS

We are pleased to acknowledge contributions from Andrew Storm, John Asenstorfer, Mathew George, King Leong, Edward Neuberg, and Michael Beyrouti, and support from the Australian Radio Research Board (now Telecommunication and Electronics Research Board), and from The University of Adelaide,

6. REFERENCES

BEDDOES, M. P. and CHU, T. K. (1982) "Direct sample interpolation (DSI) speech synthesis: an interpolation technique for digital speech data compression and speech synthesis", IEEE Trans. Vol. ASSP-30 No. 1, pp.825-832.

BOGNER, R. E. (1983) "Glitchless speech splicing", IREECON International Convention, Sydney, pp. 540-542.

FAIRBANKS, W. L., et al.(1954) "Method for time or frequency compression-expansion of speech", IREE Trans, Professional Group on Audio, Vol. AU-2, pp. 7-12.

FLANAGAN, J. L., and GOLDEN, R. M. (1966) "Phase vocoder", Bell Sys. Tech. J., Vol. 45, pp. 1493-1509.

GABOR, D. (1947) "New possibiliies in speech transmission", J.IEE, Part III, November, pp. 1-33 (from a reprint of BTH Co.)

JONES, T. F. (1971), "Speeded speech", in "Aspects of Network and System Theory", Kalman R. E., and De Claris, N (Eds), Holt, Reinhart and Winston, pp. 527-532.

KOCH, R. F. (1972), private communication, and (1978) "Deltamodulation is alive and well and living in time compression", Audio Eng. Soc. 61st Convention, New York, prepr. No.1391 (H-6), No. 3-6. (Describes "Ambichron" product.)

MALAH, D. (1979), "Time-domain algorithm for harmonic bandwidth reduction and time scaling of speech signals", Trans IEEE, Vol. ASSP-27, No.2, pp.121-133.

NEUBERG, E. P. (1978) "Simple pitch-dependent algorithm for high quality speech rate changing", J. Ac. Soc. Am. Vol.63, No.2, pp.624-625.

PORTNOFF, M. R. (1981) "Time-scale modification of speech based on short-time Fourier analysis", IEEE Trans., Vol. ASSP-29, No. 3, pp. 374-390.

SCHIFFMAN, M. (1974) "Variable speech control", Electronics, Vol.47, No. 17, Aug. (The basis of one commercial device.)

SEO, I. (1974) "Speech Compression" in Duker, S "Time Compressed Speech", Scarecrow Press, pp. 581-523.

SUZUKI, J. (1976) "SPAC - Speech processing system by use of autocorrelation function", J. Radio Research Labs. (Japan) Vol.23 No.111, pp. 217-228.

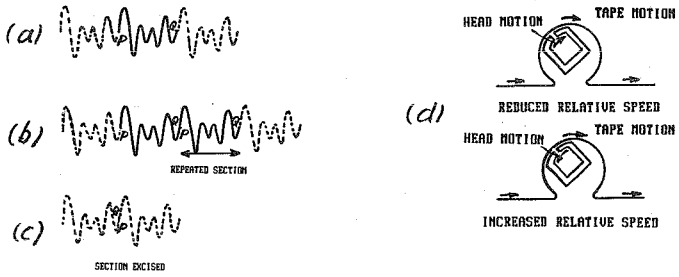

FIG. 1 Principle of time expansion and compression (a) original waveform; (b) time expanded by repetition; (c) compression by deletion; (d) early realisations of process.

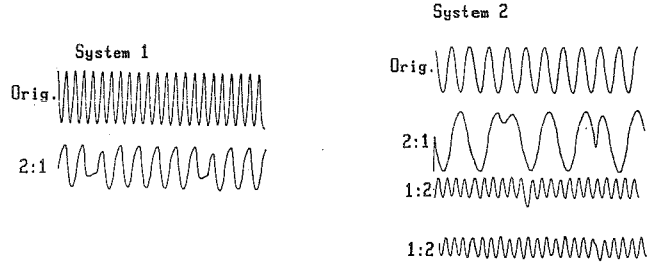

FIG.2 Examples of use of commercial devices.

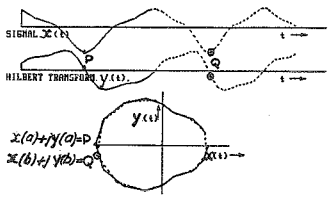

FIG.3 Compound signal x, and its HT, y, and xy plot.

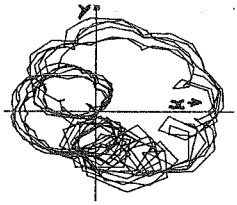

FIG. 4 xy plot for a speech signal, showing ambiguity.

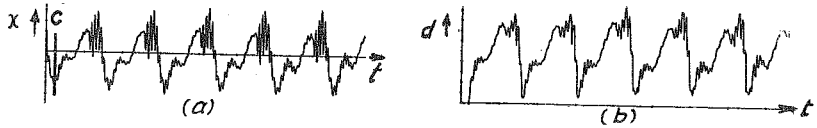

FIG. 5. (a) Speech signal x; (b) behavior of metric corresponding to cut at point C.

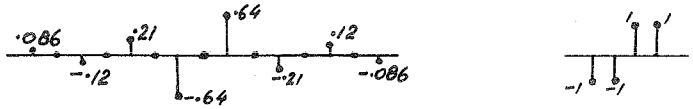

FIG. 6. Impulse responses of filters used to derive y and w.

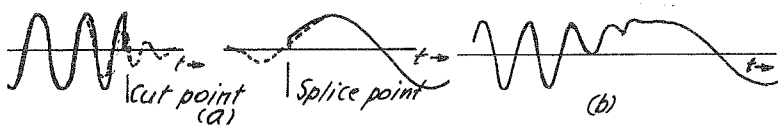

FIG. 7. Blending of abutments. (a) Cut end and splice end; tapered ends shown dotted; (b) Spliced.