FREQUENCY-AXIS WARPING TO IMPROVE AUTOMATIC WORD RECOGNITION

EDWARD P. NEUBURG

Department of Defense
9800 Savage Rd., Ft. Meade, MD 20755

ABSTRACT

Frequency normalization of talkers remains a problem in word recognition, especially where new talkers cannot be asked to provide samples (of their vowels, for example) in advance. Several methods were investigated; for each, parameters were derived by calculating their effect on formant histograms derived from casual speech. Methods tried were

a) uniform multiplication of frequencies ("stretching" the vocal tract);
b) "stretching" each formant region by a different amount;
c) combined shift and stretch (affine mapping);
d) different affine mappings for different formants (this includes warping each formant as a function of its range);
e) warping each formant non-linearly as a function of its distribution.

Experiments show that parameters derived from casual speech improve vowel recognition markedly, and that method e) appears strongest.

Introduction

The use of spectral-temporal pattern matching for automatic word recognition has become commonplace. In the single-speaker, fixed-channel, isolated word situation, spectral-temporal amplitude pattern matching gives operationally (and commercially) reliable recognition. In this simple scenario the only adjustment that seems to be necessary to align stored template with incoming utterance is "distortion" in the time direction to undo differences in speaking rate, and overall gain adjustment.

The multi-speaker non-fixed-channel situation introduces a host of (unresolved) problems, one of which is the subject of this paper; frequency axis distortion. No "distortion" has occurred, of course--it is simply that for a given speech sound the user of the device may not have concentrations of spectral energy in the same places as did the speaker of the template. Of course his spectra are likely to differ from the template in many other ways as well; however the problem considered here is how to move his broad spectral peaks to apprxoimate those of the template talker. It should be noted that if this can be done it should also improve the channel-normalization process known as "blind deconvolution".[1]

Procedure

In doing frequency warping for a "not uncooperative speaker" (one who will speak clearly but cannot be asked to train the device) one cannot use schemes that depend on having the talker utter certain sounds or certain words; whatever information is used must be taken from the speech itself. The information used in this study was derived as follows: in a corpus of the talker's speech, measure voicing and power for each centisecond segment, and discard weak or unvoiced segments. On the remaining segments do a 12-coefficient autocorrelation LPC, find the roots of the LPC polynomial, and convert them to frequency and bandwidth. Discard segments that do not have exactly one "formant" of reasonable bandwidth in each of the F1, F2 and F3 regions. What is left is 500 to 1500 segments per minute of speech: most of the talker's vowel segments (including transitions), many semivowels, some nasals, and a very few voiced fricatives. For each talker four frequency counts, or histograms, were made from these data: an F1 count, an F2 count, an F3 count, and a joint F1-F2 count.

The aim of the exercise was to produce a mapping of the frequency axis onto itself which, when applied to all the speech of a user, makes the stored templates most effective in recognizing his words. (There is no intended implication that a fixed overall frequency distortion is best.) Mappings considered were:

a) linear transformation (multiply all frequencies by the same factor);

b) affine transformation (add a constant to all frequencies, and then do a linear transformation);

c) three different linear transformations, of the F1, F2, F3 regions of the frequency axis;

d) three different affine transformations;

e) non-linear (actually piecewise-linear) transformations of each of the formant regions. This method is discussed more fully below.

Scoring Parameters

To find the optimum parameters for a warp, one must have a way of evaluating ("scoring") warps. This was done by comparing the talkers' F1-F2 histograms (because they are estimates of F1-F2 distributions). First, compute the F1-F2 histogram for talker B. Warp A's formants, then compute his F1-F2 histogram. Normalize both histograms to have (say) 2000 entries. The "score" of the warp is the square root of the sum of the squares of the differences between corresponding cells. This is like regarding the histograms as surfaces over the F1-F2 plane, and regarding the score as the distance between the two surfaces. A small score is "good".

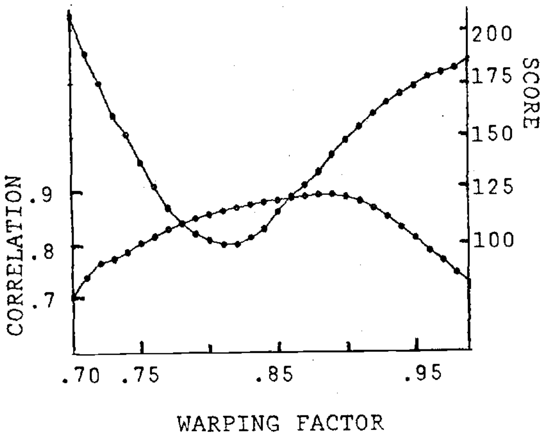

Figure 1. Upper curve shows score of linear warp (ordinate on right) vs warping factor. Lower curve shows spectral correlation (ordinate on left) vs warping factor. Curves show that score is a good predictor of effectiveness of warp.

Deriving Parameters

Parameters for each warping technique were derived for several talker-pairs.

a) Linear: Many multiples (warps) of A's formant frequencies were scored against B's. (This experiment is equivalent to Wakita's vocal tract length normalization.[2]) Figure 1 (upper curve) shows a typical plot of scores of warping factors (showing a clear minimum).

b) Affine: one can again try all reasonable (two-parameter) warps. The best typically has a score of about 85.

c) Three linear warps: since only F1 and F2 are used in the scoring, this is really a test of two linear warps. Typical best score is 86.

d) Three affine warps. (Again, to evaluate the idea, two-affine warps were scored.) Two classes of warp were tried.

1) Although an affine mapping is defined as a stretch and a shift, it is completely determined by specifying the images of two points. If we choose those points to be the ends of the range of a formant, the affine warp is equivalent to the "normalization" suggested by Gerstman [3] in which F1 (say) is expressed as a fraction of the talker's F1 range. Now, the range of F1 can at best only be estimated from a small sample. Some statisticians use the semirange, the set of values between the 25th and 75th percentiles of F1. Extending this notion, the set from the k-th percentile to the (100-k)-th was tried as the "range" of F1 and F2, for every k from 0 to 50. The score for the best of these was the evaluation score for what might be called range matching. A typical best score was (a rather poor) 120.

2) In addition, all (unrestricted) affine warps were tried for F1 and F2. Typical best score was 82.

e) Piecewise linear warp: the restricted affine warp maps A's F1 range into B's F1 range. An extension of range-matching is mapping A's F1 distribution into B's F1 distribution. To approximate this, slice A's F1 histogram into (say) 10 equal areas (with 11 boundary points) and do the same for B. Now map A's 11 boundary points into B's boundary points, with linear interpolation in-between. Do the same for F2 and F3. This amounts to 11 contiguous affine transformations for each formant. There is no "best" such equal-area mapping; there is just one such for each formant. A typical score for such a mapping is 68.

Table 1 sums up the evaluation of the five warping schemes; equal-area mapping is a clear winner.

Table 1. Summary of scores for various types of frequency-warping functions.

Type of Warp	Score (typical)
Linear Transformation	86
Affine Transformation	85
3 Linear Transformations	86
3 Affine (range matching)	120
3 Affine (unrestricted)	82
Piecewise linear	68

Testing

Linear and piecewise linear warping were tested on some marked speech. Warping factors were obtained as above on one minute of speech of two talkers. Then individual vowel nuclei of those talkers were located (by listening) and logamplitude spectra made. Spectra of the same phoneme from A's speech and from B's were compared by computing the correlation coefficient (which goes from -1 to +1, +1 being a "perfect" match). Then A's spectrum was warped linearly by the various factors, and compared again with B's spectrum. Typical results are shown in Figure 1 (bottom curve). The predicted best factor (minimum of upper curve) is very close to being actually best (maximum of lower curve). If several vowel results for these talkers are averaged one finds that the average unwarped correlation is .55, the average correlation using predicted best factor is .75. Using piecewise-linear warping the average correlation is .77. The average best correlation possible is 81.

The average correlation over wrong vowel matches is .01 for unwarped and -.01 for warped (not a significant change). Thus false matches should not increase under warping.

Conclusions

1) Spectral-temporal amplitude pattern matching (at least of vowels) can be markedly improved by frequency-axis warping.

2) Warping parameters can be satisfactorily extracted from casual speech.

3) There is some indication that equal-area warping is superior to linear or affine warping.

REFERENCES

[1] T. G. Stockham, T. M. Cannon, and R. B. Ingbretsen, "Blind deconvolution through digital signal processing," PROC. IEEE, Vol. 63, No. 4, pp. 678-692, April 1975.

[2] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," IEEE TRANS. ACOUST., SPEECH, SIGNAL PROCESSING, Vol. ASSP-25, pp. 183-192, April 1977.

[3] L. J. Gerstman, "Classification of Self-normalized Vowels," IEEE TRANS. AUDIO AND ELECTROACOUSTICS, Vol. AU-16, No. 1, pp. 78-80, March 1968.