Measuring the Dynamic Encoding of Speaker Identity and Dialect in Prosodic Parameters

Michael Barlow †

and Michael Wagner‡

†University of NSW/ADFA, AUSTRALIA

‡University of Canberra, AUSTRALIA

ABSTRACT

This paper describes a methodology, and the results stemming from it, for analysing the dynamic encoding of speaker identity and dialect in prosodic parameters. A method based on employing properties of the well known Dynamic Time Warping (DTW) algorithm's path of best match allows the separation of purely dynamic from static properties of acoustic parameters and hence their evaluation as to dynamic encoding of speaker characteristics.

Nineteen adult speakers of Australian English were recorded uttering a set of four sentences on five separate occasions over a period of at least one week. The prosodic parameters F_0, short-time energy, zero crossing rate and voicing were extracted for all data and analysed as to their dynamic encoding of speaker identity and dialect. Discriminate analysis (for speaker identity) and correlation analysis (for speaker dialect) analysis showed higher dynamic encoding of identity (75%) and dialect (0.58) than static encoding (55% and 0.45 respectively). Normalisation of all parameters into the range 0—1 reduced discriminate and correlation scores to 70% and 0.54 respectively.

Contrasting the warp path parameters with the more conventionally employed DTW distance showed that the warp path parameters better measured speaker identity (72% versus 54%) and speaker dialect (0.56 versus 0.31) encoding. Individual analysis of the prosodic parameters shows a far higher encoding of identity and dialect in F_0, though all four parameters encode dialect and identity.

1. INTRODUCTION

One of the well recognised divisions in the form of speaker characteristic encoding in utterances is that between static (time invariant) and dynamic (time varying) [4].

While many researchers have pointed out this division, most work has targeted static, segmental level encoding and chiefly that of speaker identity. One of the reasons for this lack of attention to dynamic, suprasegmental (prosodic) encoding is the difficulty in measuring, quantifying, and meaningfully contrasting both the dynamic parameters and other speaker characteristics.

With a methodology that allows the separation of the inter-woven dynamic and static encoding in any acoustic parameter it becomes possible to evaluate the degree and form of speaker characteristic encoding in prosodic parameters—something often illustrated by perceptually motivated experiments (e.g., [7]), but not normally quantified adequately.

This paper describes a methodology employing the well-known Dynamic Time Warping algorithm to measure both the local and global temporal differences between two acoustic parameter contours. Whereas the DTW distance computes a single, scalar average difference between two contours the *warp path*: the calculated temporal path of best match captures micro and macro temporal differences between the two contours.

Properties of the warp path, rather than the DTW distance, were used to measure the dynamic encoding of speaker identity and dialect in the acoustic parameters.

2. SPEECH DATA

A database of nineteen (19) adult speakers (12 male and 7 female) of Australian English was recorded. The speakers uttered a set of four (4) sentences on no less than five (5) occasions each, over a period of not let less than one week. Table 1 lists the four sentences employed.

Text
"We were away a year ago."
"I cannot remember it."
"How do you know?
"We are firm."

Table 1: Sentences employed in the study and recorded by nineteen adult speakers of Australian English on no less than five separate occasions each.

Although clearly a simplification, the dialects of Australian English, as described in [2] were mapped onto a linear numerical scale. A linguist listened to all utterances and assigned each speaker a dialect score ranging from zero (for cultivated dialect), through the dialect spectrum continuum to ten (for broad dialect) [2].

2.1. Parameter Extraction

Speaker utterances were recorded in a quiet environment. The recordings were then low-pass filtered at 7.6kHz before 12-bit quantisation at a sampling rate of 16kHz. The recordings were then hand-segmented to detect sentence start and end-points.

For each sentence four prosodic parameters were extracted with 25ms frames as follows:

- Energy – Log Mean Squared Amplitude.
- Zero Crossing Rate
- Fundamental Frequency (F_0) – extracted using a time domain parallel pitch detector with a 25ms frame, and a 10ms frame shift. Unvoiced frames were eliminated
- Voicing – Voiced/Unvoiced values were extracted for a frame size of 25ms, with a 10ms frame shift based on the output of the pitch tracker. Voiced frames were assigned the value 1 and unvoiced frames assigned the value 0, creating a square-wave representation.

Derived acoustic parameter series were post-processed with a median-5, followed by a mean-3 filter [6] in order to eliminate spurious values.

3. METHODOLOGY

In order to both quantify the temporal variability of acoustic parameter values and also allow meaningful evaluation of speaker characteristic encoding a number of new methods were designed for the experiments.

3.1 DTW Mechanism & Parameters

As conventionally applied the DTW algorithm allows the time-alignment of two vectors such that a distance may be calculated between the two aligned vectors. Fundamental to this approach is the calculation of a warp path: an alignment of the two vectors which is constrained by conditions of monotony, continuity and limited divergence, and which, in essence, stretches portions of each vector (through repetition of values) so that their alignment is optimal (mean distance between aligned pairs along the path is minimal). In conventional DTW the warp path is a by-product of the distance calculation and not generally used further.

Clearly, however, the warp path encodes details of the *relative* temporal differences between the two vectors in question; in other words their relative dynamics. To quantify such information means to create a powerful tool for examining dynamic encoding.

While there are any number of variants on the DTW algorithm, a very simple approach allowing only horizontal, vertical, and diagonal transitions (without skipping) was chosen in order to facilitate the extraction of warp path properties.

Examining a warp path, as illustrated in figure 1 and employing the formulation above, certain key properties of the warp path become clear. A *transition* is a movement along the warp path from one point to the next. Transitions may be horizontal, diagonal, or vertical. A series of transitions, all in the same direction and bounded by either a transition in another direction, or the end of the warp path may be considered an *excursion*. For instance, in figure 1, there are a total of eight diagonal *transitions*, which comprised four separate diagonal *excursions*, the longest being of length three (in the middle of the path).

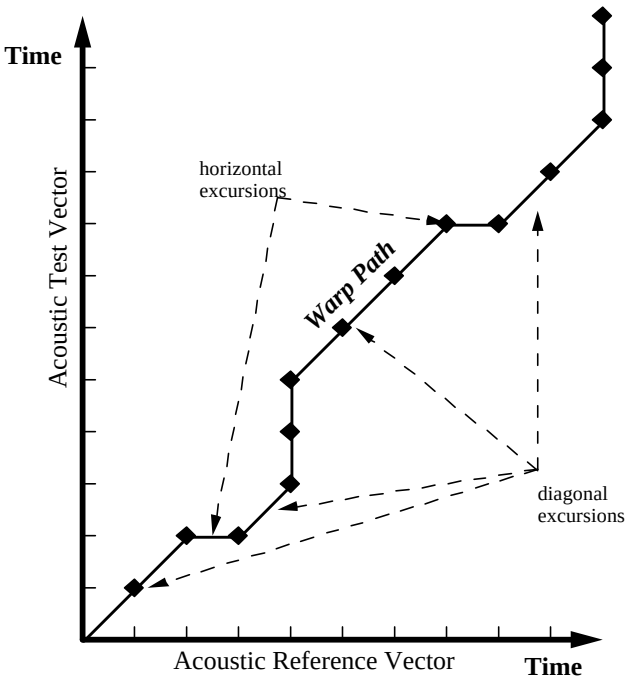

Figure 1: Illustration of the basic DTW (Dynamic Time Warping) paradigm in which a warp path (path of best fit) is calculated between a test and reference acoustic vector.

Intuitively, these transitions and excursions indicate the `goodness of fit', and hence temporal alignment of the two acoustic parameters in question. Diagonal transitions indicate the contours are well aligned at that point, whereas horizontal and vertical indicate misalignment. Excursions show regions of alignment (diagonal) or misalignment (vertical or horizontal). The number of excursions indicates the number of temporal adjustments required for a best match, while the length of excursions the degree of alignment (diagonal) or misalignment (vertical or horizontal) in that region.

Nine properties of the warp path were extracted relating to the concept to transitions and excursions. For each of the three directions of transition (horizontal, diagonal, and vertical) the following three properties were measured: the total number of transitions in that direction; the number of excursions in that direction; the length of the longest excursion in that direction. All values were normalised by dividing them by the total length of the warp path (yielding a value between 0 and 1).

Two additional measures were devised in an attempt to quantify the macro alignment of the two acoustic parameters. The first simply measured the ratio of the warp path length (K) to the longer of the two acoustic parameters (values close to 1 indicate little "non-productive" warping). A second measure (δ), which calculated the mean distance of the warp path ($w_k = [i_k,j_k]$) from the theoretically optimal path of the straight line joining the start ([1,1]) and end-points([M,N]), was also created.

$$\delta = \frac{1}{K}\sum_{k=1}^{K} | i_k - j_k \frac{\min(M, N)}{\max(M, N)} |$$

$$x_i' = \frac{x_i}{\max(\tilde{x}) - \min(\tilde{x})}$$

3.2 Speaker Identity Experiments

Experiments to determine the degree of identity encoding were conducted by calculating intra-speaker and inter-speaker scores for individual sentence, repetition and acoustic parameter couplings using the procedure described above.

In order that the influence of other speaker characteristics was minimised inter-speaker comparisons were only made between speakers of the same gender.

Discriminant analysis [5], the degree of separation between inter-speaker score and intra-speaker score distributions, was applied. This yielded a single percentile figure that measured the overlap between the two distributions:- 100% being no overlap and 0% implying no separability. It is worth noting that this figure is an absolute lower-bound on the performance of a recognition system that employed the parameterisation and weightings.

3.3 Speaker Dialect Experiments

Experiments to determine the degree of dialect encoding were conducted by computing scores between instances of individual acoustic parameters for each sentence and repetition pairings.

In order that the influence of other speaker characteristics was minimised only inter-speaker and intra-gender scores were calculated (i.e., always between different speakers but of the same gender).

Characteristic	Dynamic	Static	Combined
Identity	74.6%	54.8%	75.2%
Dialect	0.563	0.452	0.584

Table 2: Speaker identity and dialect encoding results contrasting dynamic and static encoding.

To allow the evaluation of dialect encoding, a dialect difference score was calculated for each comparison as the absolute difference between the dialect values assigned the speakers of that utterance. For instance, all comparisons between a speaker with a dialect score of seven, and another with a score of 3 would be regarded as having a dialect difference of four. These dialect differences were then correlated with the values extracted from the DTW based comparison method.

4. RESULTS

Table 2, and figures 2 and 3 show the basic results for speaker identity and dialect when measures of the dynamic encoding are contrasted with static encoding measures.

In order to ensure that purely dynamic (temporal) properties of the acoustic contours were being employed the parameters were *normalised* into the range 0—1 via the following formulation:

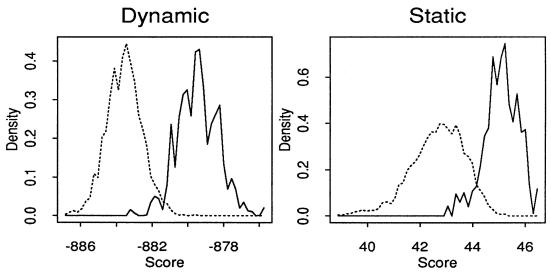

Figure 2: Speaker identity discriminate analysis results showing a rate of 74.6% for dynamically encoded information and 54.8% for statically encoded information. The intra-speaker score distribution is represented by the broken-line, while the inter-speaker distribution is represented by the solid line.

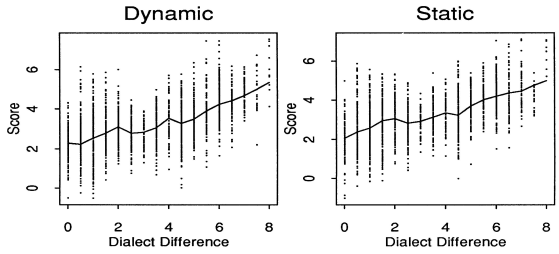

Figure 3: Speaker dialect correlation analysis results showing a correlation of 0.563 for dynamically encoded information and 0.452 for statically encoded information. Dialect difference between contrasted speakers is plotted along the horizontal axis.

Such an approach removes all absolute static information about the parameters (e.g., absolute pitch level). Table 3 presents the results of the normalisation when purely dynamic properties are examined.

Characteristic	Un-Normalised	Normalised
Identity	74.6%	69.6%
Dialect	0.563	0.524

Table 3: Speaker identity and dialect encoding results showing degree of dynamic encoding as measured when all acoustic parameters are normalised into the range 0—1.

Table 4 shows the results of the analysis of the individual acoustic parameters:- energy, fundamental frequency, voicing and zero crossing rate, as to their individual encodings of speaker identity and dialect.

In order to evaluate the utility of the warp path parameters when compared with the more conventional DTW distance approach, speaker identity and dialect encoding experiments were conducted which directly contrasted the two measures. Table 5 shows the results of those experiments.

Acoustic Parameter	Identity Encoding	Dialect Encoding
Energy	58.2%	0.359
F_0	64.6%	0.364
Voicing	47.6%	0.261
Zero Crossings	51.5%	0.306

Table 4: Contrast of the four prosodic parameters as to degree of speaker identity and dialect encoding.

5. DISCUSSION

Examining the results presented in the previous section it is clear that both identity and dialect show high degrees of encoding in the prosodic parameters. This is well known for identity but somewhat surprising for dialect: the three Australian idiolects are defined by their allophonic variance [2]. As can be seen from figure 3 the dialect correlation with the measured parameters is a general trend, and one that does not apply in all instances.

Characteristic	DTW Distance	Warp Path	Combined
Identity	54.2%	72.0%	74.6%
Dialect	0.301	0.450	0.563

Table 5: Contrast of the conventional DTW distance with measures derived from the DTW warp path as to utility for extracting speaker identity and dialect encoding.

Examining the form of encoding it is clear from table 2 that both identity and dialect are more strongly encoded in the dynamic, temporal properties of the prosodic parameters than their static properties.

This result is made even clearer by the results shown in table 3. Even with all parameter contours being normalised to lie within the range 0—1 high levels of encoding of both identity and dialect are shown. Clearly, with the absolute static information of the parameters eliminated by normalisation the methodology is left measuring only temporal properties of the parameters, and suffers little drop in discrimination (identity) or correlation (dialect) rate. Indeed, contrasting the results of table 2 with those of table 3 it may be seen that even after normalisation more information is carried in the dynamic properties of the prosodic parameters than in their pre-normalisation, static properties.

Analysing the results for individual prosodic parameters it is clear that all encode some degree of information concerning speaker identity and dialect. Clearly, fundamental frequency shows the highest levels of encoding; though still considerably less than that when all four parameters are combined (showing the utility of using multiple parameters).

When the methodology itself is examined, as illustrated in table 5, an interesting result becomes clear. The warp path encodes, or more accurately is capable of extracting significantly higher levels of both identity and dialect encoding, than the DTW distance itself. This has implications for recognition systems employing dynamic-time-warping, as well as potentially offering a new *'lease on life'* for an algorithm that is little used in today's recognition engines. Indeed, past investigations [3] have shown the significant reduction in error rate achievable by a DTW based recognition system that incorporates properties of the warp path.

Though not reported on here, further experiments [1] examined other parameters of the experimental design. The individual warp path parameters were all found to be of some utility in extracting encoded speaker characteristics; though, when considered alone, only two showed a performance equivalent to the DTW distance, namely the number of horizontal transitions, and the δ-distance. However, as seen above, a weighted sum of those parameters appears (at least for the current problem) to be significantly superior to the conventional distance.

Similarly, speaker gender discrimination experiments [1] were carried out. As expected, static measures of fundamental frequency were the strongest discriminators. However all four prosodic parameters encoded some gender specific information and even after normalisation of parameters into the range 0—1, speaker gender could be discriminated at over 77%.

6. REFERENCES

1. Barlow, M., *Prosodic Acoustic Correlates of Speaker Characteristics*, PhD Thesis, University of NSW, 1991.
2. Bernard, J.R. L-B, *Some measurement of some sounds of Australian English*, PhD Thesis, Sydney University, 1967.
3. Booth, I., Barlow, M., and Watson, B, "Enhancements to DTW and VQ decision algorithms for speaker recognition", *Speech Communication 13:* 427-433, 1993.
4. Furui, S. "Comparison of speaker recognition methods using statistical features and dynamic features", *IEEE Trans. ASSP-24(3):* 342-350, 1981.
5. Hays, W.L., and Winkler, R.L., *Statistics: Probability, Inference and Decision*, Holt, Rinehart and Wi5ston Inc., New York, 1971.
6. Hess, W., *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.
7. Williams, C.E., and Stevens K.E., "Emotions and speech: some acoustic correlates", *J. Acoust. Soc. Am. 52(4),* 1238-1250, 1972.