Inter-Speaker Scaling of Poly-Segmental Formant Ensembles

Frantz Clermont
Department of Computer Science, Mathematics & Science,
The American University of Paris,
akustikfonetiks@yahoo.com.au

Abstract

A linear-scaling approach is described for handling inter-speaker variations. The approach is motivated (i) by the similarity commonly observed amongst the formant-patterns resulting from different speakers' productions of the same utterance, and (ii) by the fact that there are linear-scaling properties associated with similarity. In practical terms, linear transformations of the formant-patterns amongst different speakers are sought and interpreted as a set of scaling relations; the formant patterns are obtained from an ensemble of phonetically-varying segments. Using multi-speaker formant data on Australian English "hello", the ensemble scales are found to explain the bulk of inter-speaker differences. The approach is independent of segmental structure; it uses only linear regression as its main computational machinery.

1. Introduction

The present study draws its motivation from a familiar observation that has potential implications for handling inter-speaker variations. Indeed, despite the range of differences caused by organic features and articulatory habits (Nolan, 1983), there is a striking similarity amongst the formant-patterns resulting from different speakers' productions of the same utterance. In sympathy with Ohta and Fuchi's (1984) "constancy" interpretation, the similarity is thought to be a manifestation of different speakers tending to utilise similar vocal-tract configurations.

Thus, one promising implication of the similarity phenomenon is that, irrespective of the multiple causes of inter-speaker differences, there should be some hope for predictable regularity in formant-pattern variability from speaker to speaker. To characterise the regularity beyond the fine details of its components, the acoustic-phonetic segments selected from a given utterance are treated as a speaker-dependent ensemble.

If the scaling properties of similarity are then brought to bear on a multi-speaker family of ensembles for a fixed utterance, it is possible not only to quantify inter-speaker similarity but also invert the bulk of inter-speaker differences. These scaling effects are illustrated using multi-speaker formant data, which span a small subset of the phonetic space for Australian English, but which implicate a range of vocal-tract configurations.

2. Ensemble Similarity: The basic concept

The similarity phenomenon is conceptualised by way of Fig. 1. The terminology developed to unfold the concept will be our starting point.

Along the ordinate axis, each rectangle contains a set of dots which, for a given speaker and a fixed utterance, schematise the relative positions of a formant's frequencies obtained for a sequence of phonetic segments selected from that utterance. The unequal spacing between the dots simulates the acoustic-phonetic variation expected from segment to segment. Such a data set is defined as a "Poly-Segmental formant Ensemble" (a PSE or an ensemble in short).

The abscissa is a "speaker axis", along which each rectangle represents a different speaker. The constant positioning of the dots within the rectangles illustrates inter-speaker similarity for the phonetic sequence.

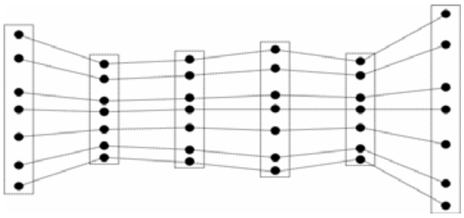

Figure 1: A systemic conceptualisation of inter-speaker similarity per formant. "Speaker axis" along the abscissa; "Poly-Segmental formant Ensemble" along the ordinate.

In the context of this work, the expression "speaker axis" therefore implies a re-organisation of multi-speaker data in terms of poly-segmental ensembles, which conform to the similarity behaviour motivated earlier. This is also schematised in Fig. 1, where the ensembles are all geometrically similar to each other and differ only in scale. In this sense, Fig. 1 portrays the case of current interest, under which PSEs would be linearly-scaled copies of each other.

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after abstract only review

3. Acoustic-Phonetic Data

Using the systemic approach exposed above, we set out to re-examine the acoustic-phonetic data presented in a previous study of the word "hello" (Rose, 1999).

In addition to being a frequent lexical item in spoken English, the word "hello" embodies a situational sensitivity that facilitates elicitation with spontaneous variability. Several situational tokens were thus produced (at one sitting) by 6 male speakers: DM (17 tokens), EM (3 tokens), JM (6 tokens), MD (12 tokens), PS (4 tokens) and RS (7 tokens). They all are native speakers of Australian English with accents ranging from general to slightly broad.

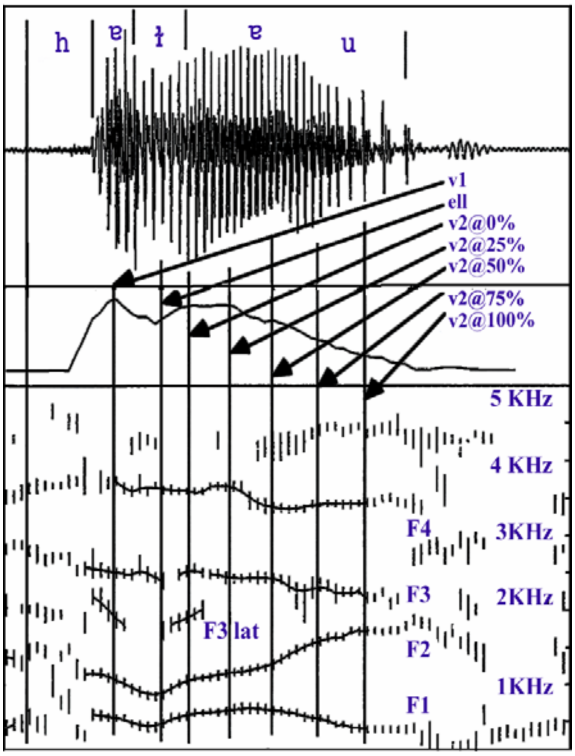

Figure 2: Acoustic-phonetic analysis of spoken "hello" [a reproduction of Rose's (1999: 9) Figure 1]. *Top panel*: Acoustic signal. *Middle panel*: Energy contour. *Bottom panel*: Linear-prediction "polegram" & selected F-patterns at 7 segments (see arrows). Segment labels: phonetic (on top of waveform); operational (at right of arrows).

The acoustic-phonetic structure for the word "hello" is adopted from Rose (1999), and consists of 7 segments (/v1, ell, v2@0%, v2@25%, v2@50%, v2@75%, v2@100%/) which span a small subset of the phonetic space, but which include a range of vocal-tract configurations – one at the initial monophthongal target /v1/, one in the middle of the lateral consonant /ell/, and five at equidistant instants of the final diphthongal gesture /v2/. For each segment and for every token, the 4 lowest formant-frequencies (F1, F2, F3 and F4) were extracted using linear-prediction analysis.

Per speaker and per formant-frequency, a poly-segmental ensemble is defined numerically as the set of token-averaged values obtained for each of the 7 segments. In sum, there are 6 speaker-dependent ensembles, for which scaling relations are sought. In the next section we proceed with details and illustrations of the scaling technique employed for the investigation of F1-, F2- and F3-ensembles.

4. Ensemble Scaling Technique

The scaling technique employed is based on Broad and Clermont's (2002) analogous development for characterising the frame-to-frame similarity of co-articulation effects on vowel formant ensembles (VFEs). Under the first-order assumption of linearity, the same technique is applicable to poly-segmental formant ensembles (PSEs), provided the data at hand exhibit a certain consistency in ensemble similarity from speaker to speaker.

In Section 4.1 it is shown that the scaling technique affords a preliminary diagnostic for lack of consistency. In Section 4.2 the technique is completely unfolded.

4.1. Pre-Scaling Diagnostics

A basic aspect of the scaling technique is the use of the speaker-averaged PSE (or the mean PSE), as a reference ensemble with respect to which individual PSEs are to be scaled. It stands to reason that the mean ensemble should be desirable for its representative behaviour and its statistical robustness. However, it is its objective role that is paramount in seeking a relative measure of ensemble-to-ensemble similarity. This quest can be pursued more confidently if, indeed, there is evidence of consistent similarity in the data at hand.

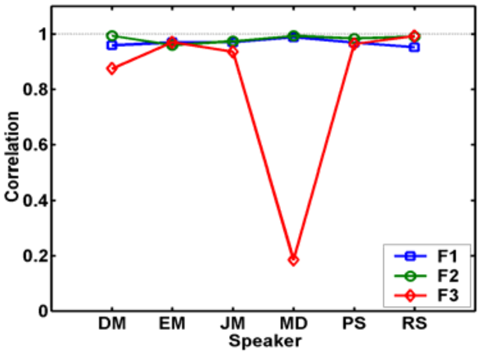

Figure 3: Profile of correlations between 6 speaker-dependent PSEs and the mean PSE (6-speaker average). The very weak correlation of **0.18** for **MD**'s **F3**-ensemble indicates his departure from similarity in F3.

One approach to detecting departure from similarity is to look at the strength of correlation between individual PSEs and the mean PSE. Fig. 3 displays such correlations for the 6 speakers and for the 3 formants.

Accepted after abstract only review

Whilst there is a very strong indication of similarity amongst all speakers' F1- and F2-ensembles, there is strong evidence against the inclusion of the F3-ensemble for speaker MD. Rather than include him only for F1 and F2, we chose to retain the 5 speakers for whom all ensembles are consistently similar, thus avoiding a compounding factor in the evaluation of the scaling technique as a tool for expressing similarity.

Indeed, the correlations re-calculated (see Table 1) for the non-problematic, 5-speaker set remain quite strong with even a slight improvement for speaker EM's F2-ensemble and speaker DM's F3-ensemble. It is with this 5-speaker set of PSEs that the scaling technique is fully exposed in the next section.

Table 1: Correlations between 5 speaker-dependent PSEs, and 2 mean PSEs: one excluding speaker **MD** (values at left of parentheses), and the other including speaker **MD** (values in parentheses).

	F1	F2	F3
DM	0.96 (*0.96*)	0.99 (*0.99*)	**0.89** (***0.87***)
EM	0.96 (*0.96*)	**0.97** (***0.96***)	0.97 (*0.97*)
JM	0.98 (*0.97*)	0.97 (*0.97*)	0.93 (*0.93*)
PS	0.97 (*0.97*)	0.98 (*0.98*)	0.96 (*0.96*)
RS	0.95 (*0.95*)	0.99 (*0.99*)	0.99 (*0.99*)

4.2. Ensemble Scaling via Linear Regression

The strong correlations reported above have confirmed the existence of consistent similarity amongst 5 out of the 6 speakers' ensembles examined, thereby paving the way for the scaling implementation itself. However, the procedure used for this purpose is more directly motivated by first taking a glimpse at actual ensemble data as shown in Fig. 4.

4.2.1. *A glimpse at Poly-Segmental Ensemble data for F2*

On the "speaker axis" of Fig. 4 are juxtaposed the F2-ensembles obtained from the 5-speakers' data. Perhaps the first observation to be made is that the ensembles are translated with respect to one another. While this may be a useful factor of differentiation amongst speakers, it is inconsequential to scaling. Instead, the crucial factor of similarity is the ensemble-to-ensemble regularity in relative position and spacing of the segments' formants. Although the ensembles shown in Fig. 4 do not appear to be exactly linearly-scaled copies of each other, there is a sufficiently noticeable trend to warrant the next step leading to scaling relations.

4.2.2. *Linear-Regression Procedure*

The scaling procedure consists of linear-regression fits of each speaker's PSE translated by its mean against the mean of all translated PSEs. This is illustrated in Fig. 5 for speaker DM, where the slope of the fitted line is an estimate of the scaling factor, justly referred to as an ensemble scale that describes a proportion with respect to the mean ensemble. The scales thus obtained for both DM and the other 4 speakers are also shown in Fig. 4 at the bottom of the rectangles.

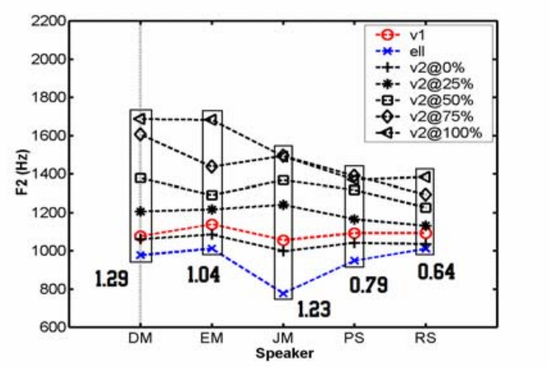

Figure 4: Five-speaker family of F2-ensembles. Ensemble scales are shown at bottom of rectangles. Fig. 5 illustrates how the scale for DM's ensemble was obtained.

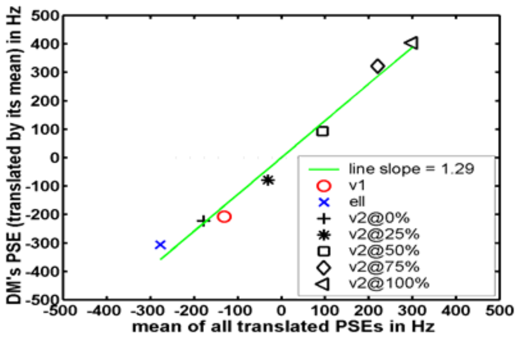

Figure 5: Linear-regression fit through DM's PSE against the mean PSE. Scale estimate = line slope.

4.2.3. *Linear-Regression Fitness*

The linear regression also yields a measure of goodness-of-fit expressed as the Root-Mean-Squared (RMS) deviations of the fitted lines from the ensemble data. Table 2 gives such measures with numerical values that are tolerable, and ranges for F1 ([17-28]), F2 ([17-57]) and F3 ([23-63]) that lie comfortably within the range of perceptual difference limens.

Table 2: RMS deviations (Hz) of fitted lines.

	F1	F2	F3
DM	25	34	42
EM	28	55	63
JM	17	57	47
PS	22	33	36
RS	24	17	23

Beyond the procedural steps described above, a closer examination of the ensemble scales is desirable to gain deeper insights into their properties and their potency.

5. Ensemble Scales

We now return to the similarity proposition, which motivated the procedure described above for deriving scaling relations. The numerical profiles of such relations are examined in this section, and shown to give concrete insights into scaling behaviours amongst our 5 speakers. In particular, the question of uniformity across formants is found to be a compounding factor that will lead to a procedural refinement of the scaling technique.

5.1. Uniformity: Observations

The ensemble scales shown in Fig. 6 are based on the original (token-averaged) PSEs and, for this reason, they will also be referred to as raw scales. The most striking observation is that DM and EM stand out with F3-scales that are at odds with the patterns for the other speakers. In addition to this apparent aberration, there are very weak correlations between F1- and F2-scales (0.02) and between F1- and F3-scales (-0.17). The raw scales clearly exhibit a strong non-uniformity that goes against the notion of similarity. A deeper investigation is warranted and undertaken in the next section.

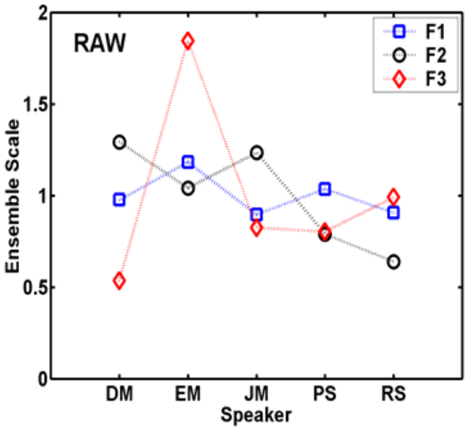

Figure 6: Per-formant profile of **RAW** scales for the 5 speakers. "**RAW**" signifies that the scales shown are based simply on the original (token-averaged) PSEs.

5.2. Uniformity: Insights from Vocal-Tract Length

Thus far, the scaling technique has yielded insights that might have been obscured if it had simply encompassed all formants in the first place. Nor does it need to as a tool for expressing similarity. The non-uniformity is therefore investigated by independently evaluating the formant ensembles before and after ensemble scaling.

5.2.1. *Vocal-Tract Length (L4)*

To understand the possible causes of the non-uniformity observed earlier, we first appeal to a measure proposed by Paige and Zue (1970) for estimating vocal-tract length. The measure has the desirable property of implicating all formants, up to F4 in our case and hence referred to as L4. The left panel of Fig. 7 displays, speaker by speaker, the raw L4 as a function of phonetic segment. The pattern of variations is continuous as the speakers' gestures progress from segment to segment with differing degrees of lip rounding, conceivably with concomitant adjustments of larynx height. Whilst the overall pattern is globally "similar" from speaker to speaker, it is quite different amongst the 5 speakers in absolute terms. The intriguing question then arises – Is the non-uniformity manifest in the raw scales partly induced by differences in vocal-tract length patterns?

5.2.2. *Inverse Scaling*

To answer the question raised above, it is critical to be able to examine inter-speaker variation left after ensemble scaling. This is readily achieved by using the reciprocals of the raw scales for inverse scaling the ensembles formant by formant. The L4 measure is then re-applied to the inversely scaled formants, yielding the new pattern shown on the right panel of Fig. 7.

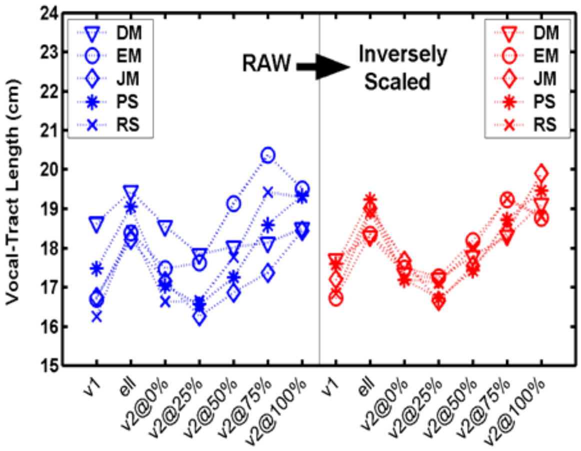

Figure 7: Profile of Vocal-Tract Lengths (L4s) based on F1, F2, F3 and F4. ***Left panel***: L4s based on raw formant ensembles. ***Right Panel***: L4s based on inversely-scaled formant ensembles.

The new pattern is revealing in several ways. The spread in L4 amongst speakers is now much smaller, thus causing a typical behaviour to emerge from segment to segment. This result clearly indicates that the scaling technique has captured significant similarity amongst the 5 speaker's ensembles. However, it is also evident that the residual pattern exhibits a certain lack of similarity from speaker to speaker, which suggests that pre-normalisation of the raw ensembles by L4 might render them more similar and hence more consistent with the scaling technique itself.

Pre-normalisation is attempted in the next section, where the results presented yield a more definite perspective on the ensemble scaling approach.

6. Ensemble Scales and Pre-Normalisation

The argument put forward in the previous section has brought into focus the fact that the scaling technique assumes no knowledge of inter-formant relationships and, therefore, it should not be able to handle the speaker-to-speaker differences in vocal-tract length patterns observed in Fig. 7. Our aim here is to secure a fairer outcome of the scaling technique by pre-normalising the raw PSEs.

In Section 6.1 we describe the pre-normalisation procedure and, in Section 6.2, we cross-examine the resulting scales given in Fig. 8 with the raw scales shown in Fig. 6. We will also return to vocal-tract length by way of Fig. 9, which illustrates the effects of pre-normalisation and inverse scaling on the pattern from speaker to speaker. Finally, the two stages put in place will take us to Section 6.3, where the raw data and the inversely-scaled data (L4-normalised) are contrasted in the planes spanned by F1 and F2, and by F2 and F3.

6.1. Normalisation by Vocal-Tract Length

The technique used for normalisation by vocal-tract length is inspired from Wakita's (1977) approach to automatic identification of 9 American English vowels uttered by 14 men and 12 women. It is relevant to note that, in accord with the reasoning unfolded in Section 5, Wakita argues that his approach is "not unreasonable as a first step toward inter-speaker normalisation in consideration of the structural similarity of the human vocal organs from individual to individual" (p. 184).

By analogy with Wakita's procedure, therefore, the ratio of raw L4s to their average is adopted as a normalisation factor. For each of the 7 phonetic segments, there are 5 such ratios corresponding to the 5 speakers, which are then used to normalise all the formants for that segment.

6.2. Uniformity Revisited

By applying the scaling technique to PSEs based on L4-normalised formants, a more meaningful picture emerges from the new ensemble scales shown in Fig. 8.

The aberrant behaviour observed earlier for DM's and EM's scales has now disappeared and, as a result, the scales follow a much more consistent pattern across all speakers. DM's and EM's ensembles are relatively larger by comparison with the other 3 speakers' ensembles, and the downward trend from left to right of Fig. 8 is also consistent for the 3 formants. The inter-formant correlations between scales have expectedly grown stronger: from 0.02 to 0.68 between F1- and F2-scales, from 0.76 to 0.89 between F1- and F3-scales, and from -0.17 to 0.93 between F2- and F3-scales.

The emergent perspective is indeed clearer. The bulk of the variations manifest in the 5-speakers' formant data appears to be caused by inter-speaker differences in vocal-tract length through the 7 phonetic segments representing the word "hello". This is confirmed in Fig. 9, where the residual variation in the new L4s is extremely small. Collectively, these results show a believable tendency towards uniformity, thereby lending support to the similarity proposition.

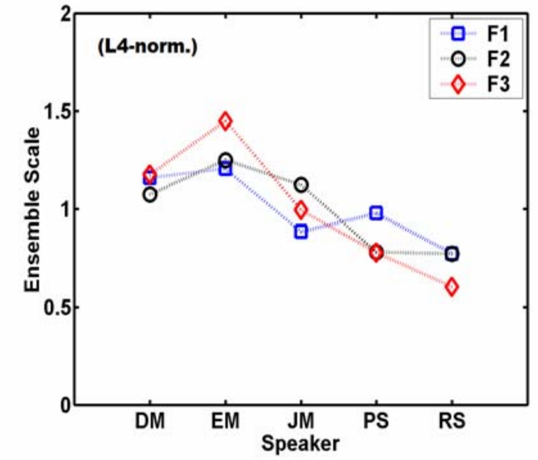

Figure 8: Per-formant profiles of the 5-speakers' ensemble scales that result from pre-normalising all the raw PSEs by vocal-tract length (L4).

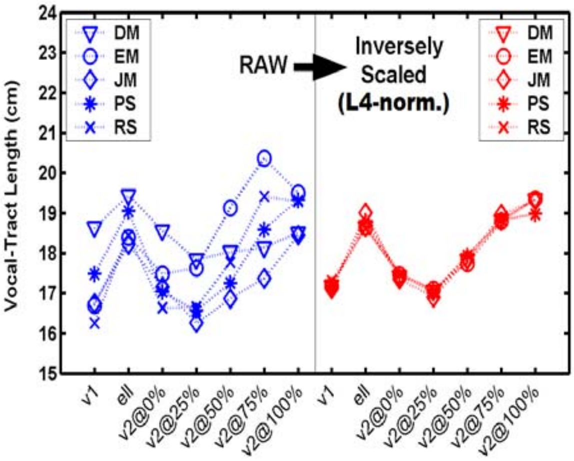

Figure 9: Profile of Vocal-Tract Lengths (L4s) based on F1, F2, F3 and F4. ***Left panel***: L4s based on raw formant ensembles. ***Right Panel***: L4s based on L4-normalised and then inversely-scaled formant ensembles.

6.3. Reduction of Inter-Speaker Variation

The two-stage process described above has yielded two measures – ensemble scales and vocal-tract length estimates, which have been instrumental in unfolding the similarity properties of our 5-speakers' formant data. It is still a question as to how much of the inter-speaker variation is indeed accounted for by these measures.

Figs 10 (F1-F2 plane) and 11 (F2-F3 plane) uncover the significant effects of both measures when they are applied in tandem through the two-stage process. The reduction in speaker variations is clearly substantial across all 7 segments in both planes. Fig. 12 gives a quantitative summary of the reduction in terms of RMS values, which are brought down to the expected level of inter-token variation for F1, F2 and F3.

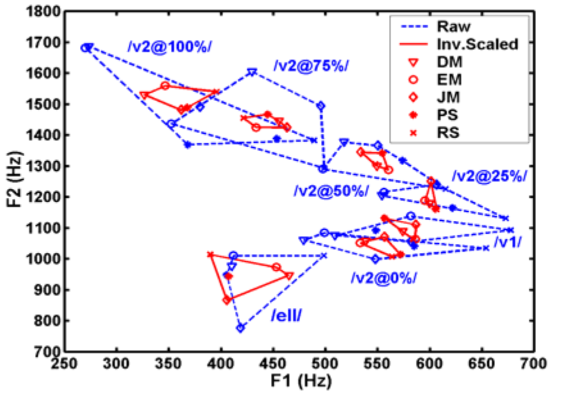

Figure 10: F1-F2 plane for 7 segments in “hello” from 5 male speakers of Australian English (DM, EM, JM, PS, RS). Raw data in blue (dashed lines); data reduced by inverse ensemble scaling (**L4-norm.)** in red (solid lines).

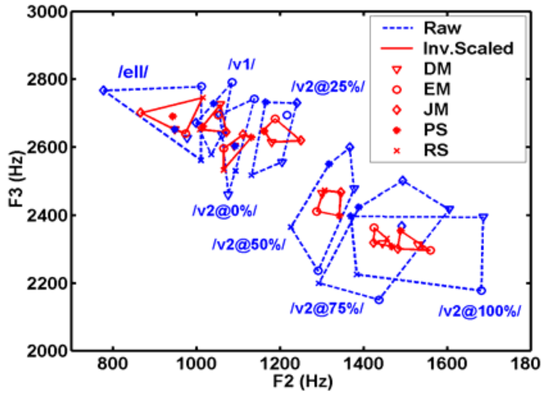

Figure 11: F3-F2 plane with Fig.10’s labeling convention.

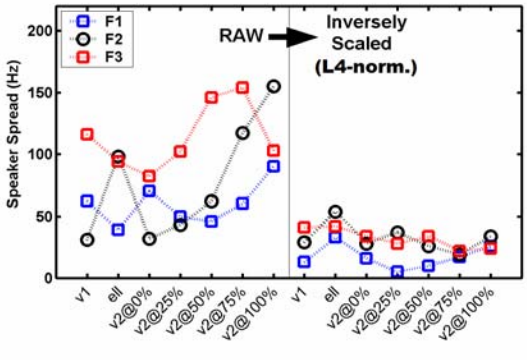

Figure 12: Speaker Spread (RMS). ***Left Panel***: raw data; ***Right Panel***: L4-normalised and inversely scaled data.

7. Summary

We have presented a new approach for handling inter-speaker variations. The approach is first motivated by similarity as an underlying phenomenon that transcends the multiple sources of speaker differences and, therefore, unlocks the predictable regularity in formant-pattern variability from speaker to speaker. It is also motivated by the fact that there are scaling properties associated with similarity, which provide a pathway for describing the regularity in practical terms.

Beyond these considerations there is a systemic philosophy that permeates the methodology developed to capture the underlying similarity. Indeed, the poly-segmental ensemble has been instrumental in uncovering speaker-specific properties of the “hello” data, which otherwise would have been obscured by looking at one phonetic segment at a time. In this sense, our notion of a poly-segmental ensemble meshes well with Laver’s (1980) poly-segmental definition of a setting.

The approach is still in its infancy, as it would need to be further evaluated with a range of segmental structures and speakers. Nevertheless, the results reported here have revealed the potentiality of the approach as well as the effectiveness of the techniques used to implement it.

8. Acknowledgements

I gratefully thank Dr David J. Broad for many inspiring discussions. I extend sincere thanks to Dr Phil Rose for giving me access to his “hello” data.

9. References

Broad, D.J. and Clermont, F. (2002). *Linear scaling of vowel-formant ensembles (VFEs) in consonantal contexts*. Speech Communication 37: 175-195.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Ohta, K. and Fuchi, H. (1984). *Vowel constancy on antimetrical vocal tract shapes between males and females*. Progress Report on Speech Research, Bulletin of the Electrotechnical Laboratory, 48: 17-21.

Paige, A. and Zue, V.W. (1970). *Calculation of vocal-tract length*. IEEE Trans. Audio, Electroacoustics, 18: 268-270.

Rose, P. (1999). *Differences and distinguishability in the acoustic characteristics of HELLO in voices of similar-sounding speakers”: A forensic phonetic investigation*. Australian Review of Applied Linguistics 22(1): 1-42.

Wakita, H. (1977). *Normalization of vowels by vocal-tract length and its application to vowel identification*. IEEE Trans. Acoust., Speech and Sig. Proc., 25(2): 183-192.