Formants and Their Dynamics: Useful for Speaker Identification?

Erik Eriksson
Dept. Philosophy and Linguistics, Umeå University
erik.eriksson@ling.umu.se

Formant dynamics have been shown to exhibit speaker specificity in English by showing low intra-speaker variation compared to inter-speaker variation. By testing the intra-speaker variation for formant dynamics in Swedish the first step towards forensic significance of formant dynamics in Swedish is established. Findings were that although formant dynamics have low intra-speaker variation for read speech the variation increase in spontaneous speech. It is suggested that previous findings are not easily generalizable from read to spontaneous speech.

1 Introduction

This paper presents a pilot study investigating the variation of formant dynamics of one speaker of Swedish and changes to the dynamics in style shift and its impact on forensic phonetics.

The paper starts with a background in forensic phonetics and speech production followed by a definition of formant dynamics. Then the study is presented and the results are discussed in relation to similar studies on formant dynamics for English.

1.1 Forensic Phonetics and Speech Production

The upper vocal tract (the mouth and nasal cavities and the lips) can be said to filter the sound produced in the larynx (Fant, 1970). The filtering of the spectral content in the source signal creates peaks in the spectrum referred to as formants and commonly denoted $F_{1...n}$. The articulators (e.g. jaw, lips and tongue) correspond to different spectral areas. The first formant, F_1, corresponds to the mouths opening and closing (height); the higher the value of F_1 the more open the mouth. The second formant, F_2, corresponds to the tongue's position in front and backness terms. That is if the tongue is towards the back of the mouth the formant frequency will be low and vice versa. For F_3, the articulatory correspondence is related to the roundedness of the lips. If the lips produce a rounded sound, F_3 will be higher than if

an unrounded sound is articulated. The correlates for formants higher then the order three is unclear but have been argued to be related to the voice quality (Ladefoged, 1982).

Since speakers vary in their anatomy, the sound they produce will differ in spectral content. This assumption is used in forensic phonetics where measurable differences between speakers are contrasted with phonation variation within a single speaker. Key in forensic phonetics is to have features that exhibit low intra-speaker variability compared to high between speaker variability (Rodman et al., 2002).

Vowel quality shift with speech rate and stress (Lindblom, 1963). Vowels are reduced in formant space as a function of articulation rate and stress. The faster the articulation the closer to "schwa" the vowel will be pronounced. However, the vowel quality is also influenced by surrounding segments (co-articulation (Öhman, 1966) and contextual assimilation (Pitermann, 2000)). Contextual assimilation can inhibit vowel reduction towards schwa at higher speech rates (Pitermann, 2000). In the case of contextual assimilation, formant values of the affected vowel will approach those of the surrounding environment (Pitermann, 2000).

Speakers could have different strategies to handle stress and reduction of vowel quality (Pitermann, 2000; Tjaden and Weismer, 1998). Tjaden and Weismer (1998) investigated the formant trajectory of F_2 under different levels of speech rate and found high inter-speaker variability in frequency changes from onset to target frequency for this formant. Speech rate or style could therefore be significant in forensic identification, especially with features that are susceptible to change with speech rate.

Since speakers could use different strategies to handle reduction, style shifts etc. it could be argued that spectral changes that are spread over time, correlated with the movement of the articulators, can act as speaker identity cues. More specific the movement of formants have been argued to bear speaker specific details (MacDougall, 2004).

1.2 Formant Dynamics

Speech is not produced in distinct segments separated by silence. The articulators have to move between salient positions in order to produce different sounds.

As the articulators move to go from one sound to another, the formants will also change. These changes have been called formant dynamics (MacDougall, 2004), formant transitions (e.g. Johnson, 2003; Lieberman, 1988), formant trajectories (e.g. Ingram et al., 1996) or F-patterns (e.g. Elliott,

2001; Rose and Simmons, 1996) partly depending on application. The terminology adopted here will be formant dynamics.

Formant frequency values have been investigated for speaker specificity in a number of investigations (e.g. Elliott, 2001; Greisbach et al., 1995; Rodman et al., 2002; Rose, 1999, 2003; Rose and Clermont, 2001; Rose and Simmons, 1996). Traditionally, single formant measurements have been taken (usually midpoint of a vowel), or in concordance with single measurements, onset and offset of vowels or transitions.

Not restricted to vowels, formants can be measured in consonants. Rose (1999) used seven landmarks within an utterance (*hello*) to investigate intra-speaker variability for formant values. These seven landmarks were chosen both in vowels, consonants (liquid) and diphthongs.

Landmarks was also used by Elliott (2001) in her investigation of the utterance *okay*. Also here was seven landmarks identified throughout the whole utterance. Given that F_4 measurements can be unreliable due to bandpass filtering over the telephone network, the most reliable formants for speaker discrimination was found to be F_1 and F_2.

Greisbach et al. (1995) showed that, for German, formant values for F_1 and F_2 could successfully be used as speaker identification measures. They studied the identification rates of single measurements of formants compared with series of values extracted at equidistant points over the duration of the segment; either a vowel or a diphthong. They found that identification rates were increased when formant values sampled over the duration of the segment compared with midpoint extraction of formant values. Also, they found that using diphthongs increased identification rates to that of monothongs. Greisbach et al. (1995) reported visually more discrimination for F_3 (see Figure 4 in their paper).

Ingram et al. (1996) used formant trajectories from sonorant segments in their study. Formant extraction were here not made at equidistant points in time but rather continuously over the whole segment. Findings were that formant trajectories acted as good features for speaker discrimination in the phonetically controlled environment. An environment that was defined as having at least one transition between two different vowels with zero or more non-vowels in between as long as the segment has a clear formant structure.

The formant dynamics have been argued to be speaker specific (e.g. MacDougall, 2004) or at least function as a possible candidate for speaker discrimination (e.g. Rose and Simmons, 1996). It has been argued that the movement from one target sound to another creates different formant dynamics, since each speaker would use different strategies or make this transition at different speed (MacDougall, 2004).

Speaker alter their vowel quality in relation to surrounding environment and as a function of speech rate. If these changes are non-significant for a specific speaker, the impact of style shift would be reduced for the forensic phonetic society. MacDougall (2004) found that speech rate and stress had little impact on the speaker specificity of formant dynamics. However, Huntley Bahr and Pass (1996) showed that style shift will have an impact on listener ability to identify speakers.

Therefore an investigation of whether formant dynamics will change with style shift has been undertaken. This paper presents a pilot study using a single speaker and variance of formant dynamics values are evaluated visually as well as in table form, although not formally tested for significance. As speech rate increase in more non-formal speech situations (such as spontaneous vs. read speech) the vowels will decrease in quality following the reduction rules. However, MacDougall (2004) showed little impact of speech rate and stress on speaker discrimination which then poses the question if speaker variation in formant dynamics is still low in spontaneous speech compared to read speech? The third formant, which was argued to have higher speaker discrimination (Elliott, 2001; MacDougall, 2004; Rose and Simmons, 1996) will be more closely investigated as it is likely that it will show less variability between speakers in Swedish since it has a function of minimal discrimination between vowels.

2 Method

One male Swedish speaker was recorded in an sound-treated room. The speaker was first asked to read a text (ca. 3.5 mins) that was given several days prior to recording, in order to familiarize the speaker with the text (this to reduce the number of errors while reading). This text was then discussed with an experiment leader for several minutes. The discussion sessions lasted as long as possible and the experiment leader tried to prime for certain target words (e.g. /bjœ̞n/). This in order to get comparable material from both the read and the discussed part.

Formant values for the first four formants were extracted for four different segments from each recording. (MacDougall, 2004) used English diphthongs in her experiment. Since Swedish have a low frequency of diphthongs compared to English, segments of glide + vowel, specifically long and short front vowels with contrasting roundness, combinations were selected. This to test the higher variability of F_3 that (MacDougall, 2004) found in her data.

The isochunks selected were /jæ/, /jæː/, /jœ̞/ and /jɛ/. An isochunk

is defined as a segment that has the same underlying representation and is pronounced similarly from one speakers perspective. The segment is variable in length and may span linguistic borders, such as word or sentence borders, as long as the border does not introduce a pause into the sound. The segment must be realized more once for each speaker (Rodman et al., 2002). Further, the isochunks may be as short as one vowel and as long as needed, as long as there are sufficiently many segments in the speech material (however, see (Eriksson et al., 2004) for a discussion on vowel length for this method). The isochunks selected here follow this definition.

Formant extraction were done with a 18 order LPC stabilized covariance method with a frame window of 0.005 seconds and a window size of 0.049 seconds using a Hamming window type. Down-sampling was done to half the sampling rate, i.e. 8000 Hz and pre-emphasis factor of 1 was used. For formant tracking the software Wavesurfer[1] was used. The automatic tracking was manually checked and corrected if necessary for all segments.

Each isochunk were divided into ten evenly spread segments and formant values for each of these ten segments were used for data analysis. This enables time alignment between separate isochunks and follows the procedure by MacDougall (2004).

3 Results and Discussion

In Fig. (1) it can be seen that the variance of formant values differ between read and spontaneous speech. However, within each speech condition the formant variations seem to be comparable.

It can also be seen that variance differences are found between formant values for the different formants within an isochunk. The figure and the tables (1 and 2) will now be discussed for each isochunk separately.

For the vowel in /jæ/ all formants show relatively little variation for the read speech and slightly more for the spontaneous speech. Also, F_3 is slightly lowered in the spontaneous data.

The long version of the previous segment (/jæ:/) show similar results as the short version for the first three formants. The fourth formant, however, show increased speaker variation for the long segment. This finding is not found in the spontaneous data in which all formant variations are reduced compared to the shorter version of the segment. The reduced variation for the long version in the spontaneous speech is to be expected as shorter segments

[1] http://www.speech.kth.se/wavesurfer/

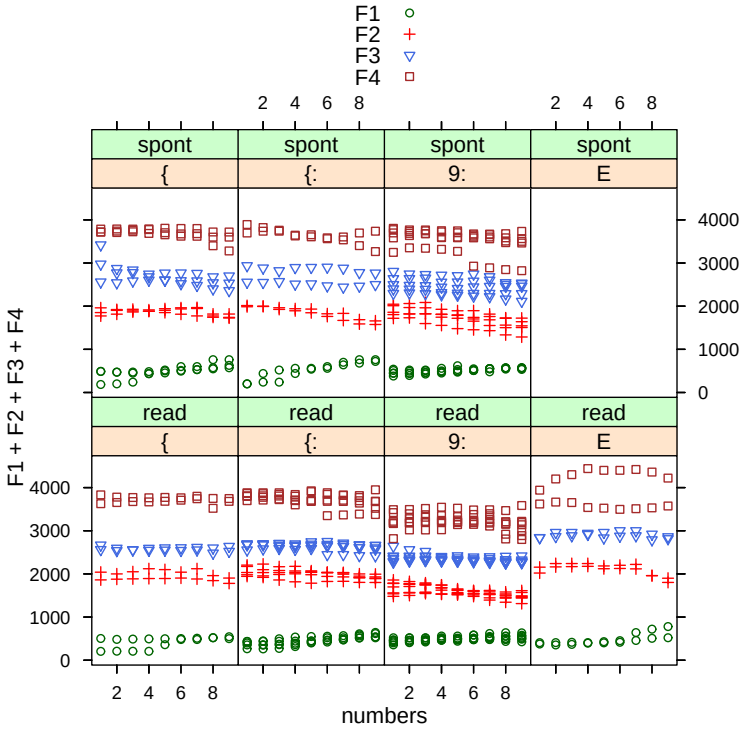

Figure 1: Plot displaying formant dynamics for all vowels stratified for speech style. "read" denotes read speach and "spont" denotes spontaneous speech. The plots are denoted by each isochunk's vowel using Swedish SAMPA; the isochunk /jɛ/ was not encountered in the spontaneous data. "Numbers" refer to the time separated segments.

will be reduced in quality and therefore reach its target value imperfectly (Lindblom, 1963).

For the segment /jœ̟/ the variation is low for the first three formants in the read speech data. This segment has a rounded vowel in it's final part which would explain the lowered variation of the third formant. However, variation for the fourth formant is high. For the spontaneous material material the variation is high for the upper three formants. This being a segment with a short vowel realization this can be attributed to the same explanation as for the previous segments discussed.

The data for the last segment, /jɛ/, suggest high speaker variability in F_4. However, this is likely to be an error in measurement, since the other formants exhibit low variation. For the spontaneous speech data no segment

Table 1: The variances of the formant values at each segment for all isochunks; data set is the read speech.

	jæ				jæː			
Seg	F1	F2	F3	F4	F1	F2	F3	F4
1	44417.0	15967.5	5467.1	20871.3	5021.9	13521.1	3444.4	5150.6
2	37326.1	7483.0	1523.4	7851.7	5661.4	15062.7	2352.1	4545.2
3	39649.2	13598.9	2142.4	4415.2	7180.9	12957.0	2289.8	4733.9
4	41270.7	25868.8	1103.8	4516.4	7167.2	17040.2	2226.5	8784.7
5	9412.4	21297.0	1361.3	3987.7	3505.4	13507.6	3234.3	14914.9
6	235.3	9309.2	1980.3	1521.5	2844.1	8265.7	12925.3	44123.4
7	429.8	28262.9	1792.1	1620.5	1729.3	6993.3	10756.0	35100.1
8	1.8	6597.2	6998.1	26215.9	1807.8	5157.8	10287.4	28163.0
9	963.9	7645.0	6622.5	2100.6	3021.8	4843.8	9798.8	45772.9

	jœ̟				jɛ			
Seg	F1	F2	F3	F4	F1	F2	F3	F4
1	2796.2	21116.5	15123.3	46670.7	487.7	9289.3	0.8	51299.4
2	1864.1	15748.2	8345.2	24046.7	1435.3	3008.4	4556.7	143256.7
3	1290.3	11790.9	5817.0	24913.1	1184.3	2507.9	2959.7	209393.68
4	1212.6	7694.0	1458.9	26150.8	76.5	2093.3	362.2	406535.4
5	1398.3	2993.4	2021.1	16188.1	379.0	2075.3	8126.5	384052.4
6	2191.4	2541.8	1444.6	14340.1	700.2	2749.2	7921.0	408855.7
7	2190.7	5919.2	2110.9	17783.5	16526.0	5371.8	6921.9	413285.9
8	4450.3	5888.1	3103.7	45659.6	22536.0	0	8667.3	344006.2
9	4537.9	9315.1	2967.1	67039.0	33598.7	4764.3	1037.2	209004.7

of this kind were encountered.

Rose (1999) showed low intra-speaker variation for the first three formants when uttering *hello*. His findings were based on both short- and long-term data (collected from recording separated by two weeks, and by a year). The findings in this paper are similar to his results even though recordings were made only once.

The data presented here also show an overall increase in intra-speaker variation for formant dynamics in spontaneous speech. This could mean that using read speech (as done by MacDougall (2004) for instance) would be inappropriate to find speaker specific acoustic cues. The speech would be too formalized to be generalizable to other speech situations.

The findings for the third formant are systematic with both MacDougall (2004) and Rose (1999). The formant exhibit very little speaker variation, at least for the read speech. However, the speaker discrimination ability of this particular formant cannot be argued in this paper as it only contains one speaker.

Table 2: Variances of formant values at each segment for the isochunks; this table presents the spontaneous data. The isochunk /jɛ/ is missing in this data set.

	jæ				jæː			
Seg	F1	F2	F3	F4	F1	F2	F3	F4
1	30515.9	8935.6	184942.2	1527.2	0	394.1	72921.4	19160.4
2	24503.3	3260.5	28768.7	1707.1	20000	21.3	55917.2	3511.1
3	16523.4	977.3	17563.6	979.2	39200	743.1	31565.9	240.7
4	724.6	311.0	4894.6	3169.5	7200	1119.7	73302.1	138.1
5	1164.2	2024.5	8868.4	5966.5	200	3612.4	72484.5	1141.6
6	3200.7	6327.6	16469.9	8429.0	800	1874.2	93596.1	199.8
7	1622.7	11246.8	19855.1	8419.7	1800	13137.9	95936.8	2216.0
8	13427.9	1394.8	19392.3	26091.3	3200	4828.7	49502.4	42963.5
9	9243.5	2430.9	28738.5	51411.2	800	3433.8	35454.5	111280.8

	jœ̟			
Seg	F1	F2	F3	F4
1	3826.7	19020.4	40028.1	51032.6
2	2410.2	17442.9	31855.5	27335.7
3	1380.7	33313.7	33327.6	29766.6
4	1675.3	18829.1	33914.0	29040.6
5	3514.3	24223.9	35225.6	34580.1
6	257.4	25595.7	43414.1	100253.5
7	1041.4	24442.6	38638.8	109663.1
8	140.2	26782.1	25648.7	108363.7
9	254.4	27326.9	32603.0	121326.6

4 Conclusions

This paper has presented Swedish material collected to replicate MacDougall (2004). Results indicate that the formant dynamics have little variance for read speech; also following the results by Rose (1999). Formant variance increased when using spontaneous speech which would implicate that acoustic cues showing good speaker discrimination in read speech should be carefully investigated before generalizing to spontaneous speech.

5 Acknowledgments

I would like to thank the cognitive science students Karolina Hammarbäck and Erik Jansson and RA Tomas Landgren for their help in processing data. I also thank my reviewers for their feedback.

References

Elliott, J. R.: 2001, Auditory and F-Pattern Variations in Australian OKAY: A Forensic Investigation, *Acoustics Australia* **29**(1), 37 – 41.

Eriksson, E. J., Cepeda, L., Rodman, R. D., McAllister, D., Bitzer, D. and Arroway, P.: 2004, Cross-language speaker recognition using spectral moments, *Proceedings FONETIK 2004, the XVIIth Swedish Phonetic Conference*, Stockholm, Sweden, pp. 76 – 79.

Fant, G.: 1970, *Acoustic Theory of Speech Production*, 2nd edn, Mouton & Co, The Hague.

Greisbach, R., Esser, O. and Weinstock, C.: 1995, Speaker Identification by Formant Extraction, *in* A. Braun and J.-P. Köster (eds), *Studies in Forensic Phonetics*, Wissenschaftlicher Verlag, Trier, pp. 49 – 55.

Huntley Bahr, R. and Pass, K. J.: 1996, The Influence of Style-Shifting on Voice Identification, *Forensic Linguistics* **3**(1), 24 – 38.

Ingram, J. C. L., Prandolini, R. and Ong, S.: 1996, Formant Trajectories as Indices of Phonetic Variation for Speaker Identification, *Forensic Linguistics* **3**, 129–145.

Johnson, K.: 2003, *Acoustic & Auditory Phonetics*, 2nd edn, Blackwell Publishing, Melbourne, Oxford, Berlin, Malden.

Ladefoged, P.: 1982, *A Course in Phonetics*, 2nd edn, Harcourt Brace Jovanovich, New York.

Lieberman, P.: 1988, *Speech Physiology, Speech perception, and Acoustic Phonetics*, Cambridge University Press.

Lindblom, B.: 1963, Spectrographic Study of Vowel Reduction, *Journal of the Acoustic Society of America* **35**(11), 1773 –1781.

MacDougall, K.: 2004, Speaker-specific Formant Dynamics: An experiment on Austrailan English /ai/, *Forensic Lingusitics: Speech, Language and the Law* **11**(1), 103 – 130.

Öhman, S. E.: 1966, Coarticulation in VCV Utterances: Spectrographic Measurements, *Journal of the Acoustic Society of America* **39**(1), 151 – 168.

Pitermann, M.: 2000, Effect of Speaking Rate and Contrastive Stress on Formant Dynamics and Vowel Perception, *Journal of the Acoustic Society of America* **107**(6), 3425 – 3437.

Rodman, R. D., McAllister, D., Bitzer, D., Cepeda, L. and Abbitt, P.: 2002, Forensic Speaker Identification based on Specral Moments, *Forensic Linguistics* **9**(1), 22 – 43.

Rose, P.: 1999, Long- and Short-Term Within-Speaker Differences in the Formants of Australian Hello, *Journal of the International Phonetic Association* **29**(1), 1 – 31.

Rose, P.: 2003, Strength of Forensic Speaker Identification evidence: Multispeaker Formant- and Cepstrum-Based Segmental Discrimination with a Bayesian Lilkelihood Ratio as Threshold, *Forensic Linguistics* **10**(2), 179 – 202.

Rose, P. and Clermont, F.: 2001, A Comparison of Two Acoustic Methods for Forensic Speaker Discrimination, *Acoustics Australia* **29**(1), 31 – 35.

Rose, P. and Simmons, A.: 1996, F-Pattern Variability in Disguise and Over the Telephone – Comparisons for Forensic Speaker Identification, *Proceedings of the Sixth Australian International Conference on Speech Science and Technology SST-96*, Adelaide, Australia.

Tjaden, K. and Weismer, G.: 1998, Speaking-Rate-Induced Variability in F2 Trajectories, *Journal of Speech, Language, and Hearing Research* **41**, 976 – 989.