Chapter 5

Naturalness in Synthesised speech: the Formant Transformation way

5.1 Introduction

A formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are several formants, each at a different frequency, roughly one in each 1000Hz band for average men and in every 1100Hz for average women. They can be seen as peaks in the speech spectrum or in a wideband spectrogram as dark bands; the darker the band, the stronger is the formant. At any one point in time, there may be any number of formants, but for speech the most informative are the first three, appropriately referred to as F1, F2, and F3. The energy in a formant comes from the sound source. In the case of a voiced sound, it is the periodic vibration of the vocal folds, producing a series of harmonic tones. With vowels, the frequencies of the formants determine which vowel sound we hear and, in general, are responsible for the differences in quality among different periodic sounds. Therefore changing the formants of one vowel to another would make it sound like the other. For example, the typical values of formants F1, F2 and F3 for vowel /u/ are F1=440Hz, F2=1020Hz, F3=2240Hz. If these formant values are changed to say, F1=390, F2=1990 and F3=2550, corresponding to vowel /i/, the /u/ will sound like an /i/. The F2 formant corresponds to vowel backness (the higher the value, the more front is the vowel) and the F1 formant corresponds to vowel height or vowel openness (the higher the value, the more open is the vowel articulation). F3 corresponds to lip rounding. In other words, values of F1, F2, F3 indicates how a particular vowel sound is produced or articulated. The most common way of representing the vowel space of a language or dialect is the F2 versus F1 plot.

Formants, considered to be responsible for differences in vowel quality, also represent regional variations in the vowel/diphthong sounds of a language. In this work, we are considering the AIR and NAL varieties of Assamese, and three approaches based on Gaussian Mixture Models (GMMs) are used to develop mapping functions to map the most informative formants, F1, F2 & F3 of vowels/diphthongs of AIR to NAL variety. The first is based on a single GMM for vowel/diphthong formants in training data. The second, maps the formants at four equidistant temporal points of vowel/diphthong duration. The third approach trains separate GMMs for the formants of each vowel/diphthong. In objective evaluation, all three approaches bring the vowel/diphthong formants of the source variety closer to the target variety. It is observed that the third approach outperforms the previous two. The current study is limited to the six vowels /ax/ (as in /axmrit/- 'nectar', IPA:/ɔ/), /aa/ (as in /aam/- 'mango', IPA:/a/), /i/ (as in /itaa/- 'brick', IPA:/i/), /u/ (as in /uraa /- 'fly', IPA:/u/), /e/ (as in /etaa/- 'one', IPA:/e/) and /o/ (as in /mor/- 'mine', IPA:/ʊ/) and a number of commonly used diphthongs such as /eaa/,/uaa/, /iaa/, /eu/, /iu/, /oi/, /ou/ etc.

The rest of the chapter is organised as follows. Section 5.2 presents a brief review of literature related to Formants and Formant Transformation, Section 5.3 presents the motivation for this work, and Section 5.4 describes the experimental framework used for developing the three approaches towards the transformation of formants. Experimental results are presented in Section 5.5 and finally in Section 5.6 conclusions and plans for future work are presented.

5.2 Literature Review

A number of studies relating formant frequencies to dialect variation have been reported in the literature. Fox and Ewa [40] report the dynamic nature of spectral change in the vowels of three varieties of American English spoken in Western North Carolina, in Central Ohio and in Southern Wisconsin. Hagiwara [48] reports considerable variation in the vowel spaces of contemporaneous regional variants of American English. Likewise Clopper et al. [28] carried out an acoustic analysis of the vowel systems of six regional varieties of American English, results of which reveal consistent regional variation with respect to production of vowels. Tamimi [3] in his study on the Jordanian and Moroccan dialects of Arabic, show that use of dynamic cues, i.e., formant slopes obtained from a linear regression analysis, improves the correct classification rates of about 5% for Moroccan Arabic and 13% for Jordanian Arabic. Williams and Escudero [153] compare the first two formants of eleven nominal monophthongs and five nominal diphthongs in Standard Southern British English and a Northern English dialect. Results of the study indicate that most cross-dialectal differences are characterised by the formant trajectory means (represented by zeroth DCT coefficients). Differences in the first DCT coefficients, which represent magnitude and direction of formant trajectory changes, are more for diphthongs. Labov [80] reports that the correlation between social factors and vowels is almost entirely concentrated in F2, while for cognitive differentiation of vowels F1 is more important. Grama's [1] findings run counter to Labov's claims and report that social meaning can lie in both F1 and F2. His findings are supported further by Teutenberg and Watson [140] who also infer that F1, F2 formants have a significant contribution to vowel quality and points on the F1-F2 plane often represent the pronunciation of a speaker. Therefore they attempt to modify the source accent to match the target accent by mapping the vowel space of source to target. Although vowel quality alone is not sufficient for accent modification, they state that even a simple transformation can yield a significant shift in the perceived accent. Since a dialect is almost always associated with an accent, transformation of vowel space, in terms of formant frequencies, from source dialect to target dialect is very likely to bring the source dialect closer to the target dialect.

Formant Transformation is a popular topic mainly associated with Voice Conversion (VC) which aims to change the source speaker's voice to sound like that of the target speaker. Formants are commonly used to represent the features of the vocal tract system and formant transformation is used to transform the vocal tract system from source to target speaker. A number of techniques are used for transforming formants. The method using artificial neural networks [97] captures the non linear relation between source and target formants. Rentzos et al. [120] model the statistical distributions of formants by a two-dimensional HMM. Two methods are explored for mapping the formants of a source speaker to those of a target speaker. The first method is based on an adaptive formant-tracking warping of the frequency response of the Linear Prediction (LP) model and the second

[1] `http://www.ling.hawaii.edu/research/WorkingPapers/wp-Grama.pdf`

method is based on the rotation of the poles of the LP model of speech. Both methods transform all spectral parameters of the resonance at formants of the source speaker towards those of the target speaker. Bohm and Nemeth [16] present a method based on the LP model to track and modify formants in speech signals which enables the modification of speech timbre and voice quality.

5.3 Motivation

In **Chapter 3**, we have observed that the position of vowels such as /e/, /o/, /u/, /ax/ and /aa/ in the F1-F2 plane is significantly different for the two varieties of Assamese considered for this study, i.e., AIR and NAL. Furthermore the diphthongs in the AIR variety are perceptually more prominent than their NAL counterparts. The dynamic formant plots of diphthongs in the two varieties indicate incomplete movement from the first vowel in a Nalbaria diphthong to the second, making it difficult to perceive the diphthong. Therefore if the formants of the vowels/diphthongs of a language/dialect can be transformed to match that of another language/dialect, the vowel/diphthong sounds of the source variety would sound more like their counterparts in the target variety. From our review of literature regarding formants, their association with dialectal variation and their transformation, it is observed that though a number of works report on formants and dialectal variation, most transformation works relate to the transformation of the vocal tract characteristics from one speaker to another. To the best of our knowledge, FT has not been used in any work related to the synthesis of dialects, i.e., for incorporating naturalness to synthesised speech. This work therefore capitalizes on these observations and attempts to transform the vowel/diphthong formants of one dialectal variety to another. Three approaches based on GMMs[2] are taken to transform the formants F1, F2 and F3 of the source variety (AIR) to that of the target variety (NAL) using GMM based mapping functions.

5.4 Experimental Framework

This section presents a detailed description of how the speech corpus, required for training and developing the mapping functions, is built. It gives an insight into the

[2] https://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics-GMM.pdf

methodology used for developing the mapping function to be used for transforming the vowel formants from one variety to another.

5.4.1 Building the Speech Corpus

A set of 60 text prompts (TP-FT) of short sentences in the Nalbaria variety is prepared to include at least 10-20 occurrences of the vowels/diphthongs of the Assamese language. The database previously described in Section 4.5.2.4 for voice conversion experiments, has been updated with utterances containing additional samples of frequently used vowels and diphthongs. These are recorded from the speaker (SPK), fluent in both the varieties of Assamese, at a sampling rate of 48kHz with 16 bit resolution in a sound proof room using a Zoom H4Next recorder. The set of recorded wav files are then transcribed by a person and cross-checked by another, both well-versed in transcription. The set of recorded sentences is our target set (T) and TR represents the set of phonetic transcriptions of T. 50 utterances from T are used for training the system, while the remaining 10 are used for testing. The set of source utterances (S) is generated from a HMM-based TTS (TU-TTS) which we have developed for the AIR/standard variety, with TR as input. TU-TTS is trained with speech data in the AIR variety of Assamese from SPK with pronunciation and syllabification rules pertaining to the standard variety of Assamese. Therefore S containing Nalbaria vocabulary is of AIR variety. S is generated from a TTS and not directly recorded from the speaker since our aim is to incorporate dialectal features into synthesised speech. Therefore what we essentially have is (i) a source set of utterances which we consider as the AIR variety since these have been generated from a TTS for AIR although with Nalbaria text, and (ii) a target set of utterances in Nalbaria, generated from a Nalbaria speaker. Both sets of speech data are down sampled to 16kHz to reduce memory requirements for file storage without loss of quality, and then cleaned to remove unwanted pauses, mispronunciations and noise.

TP-FT: Set of 60 text prompts in Nalbaria.

SPK: A speaker who speaks both the varieties of Assamese (AIR and NAL) fluently.

T: Set of utterances in the NAL variety (target) recorded from SPK using TP-FT.

TR: Set of phonetic transcriptions of T

TU-TTS: A standard Assamese TTS trained with standard Assamese speech data from speaker SPK

S: Set of utterances in AIR variety (source) generated using TR as input, from TU-TTS

Both source and target sets of utterances, i.e., S and T are annotated with vowel/diphthong labels and stored as textgrid files using the PRAAT tool [15] for phonetic analysis. A sample textgrid containing annotations at different tiers is shown in Figure 5.1. A PRAAT script is written to extract formant frequencies F1, F2 and F3 from all vowel /diphthong segments in S and T at four equidistant temporal locations corresponding to the 20%, 40%, 60% and 80% points of vowel/diphthong segment duration and store the results in an Excel sheet as shown in Figure 5.2. The formants are extracted in this manner to eliminate the effects of adjacent consonants on formant transitions. Moreover the dynamic nature of vowels and diphthongs can be captured by measuring the formants at such multiple points instead of measuring formants at the nucleus. Using the popular Burg algorithm [29], PRAAT extracts the specified number of formant frequencies at the specified time instants. However PRAAT sometimes gives erroneous results during formant extraction, therefore the formants are manually corrected for better accuracy in modelling.

5.4.2 Methodology

5.4.2.1 Analysis of Vowel Space

A statistical analysis is carried out on the vowel/diphthong formants F1, F2 and F3 extracted from source and target vowel/diphthong segments /ax, /aa/,/i/, /u/, /e/, /oi/, /o/, /ou/, /ui/, /axi/, /eu/, /eaa/ and /aai/. It is observed that there is not much change in F3 values in the vowels of AIR and Nalbaria. This indicates that F3 does not contribute much for the differences in vowel quality in the considered variants and therefore its transformation will not be of much significance. Results in terms of means and standard deviations (for F1 and F2) are presented in Figure 5.3. These results show that there is considerable difference in F1 and F2 mean for most

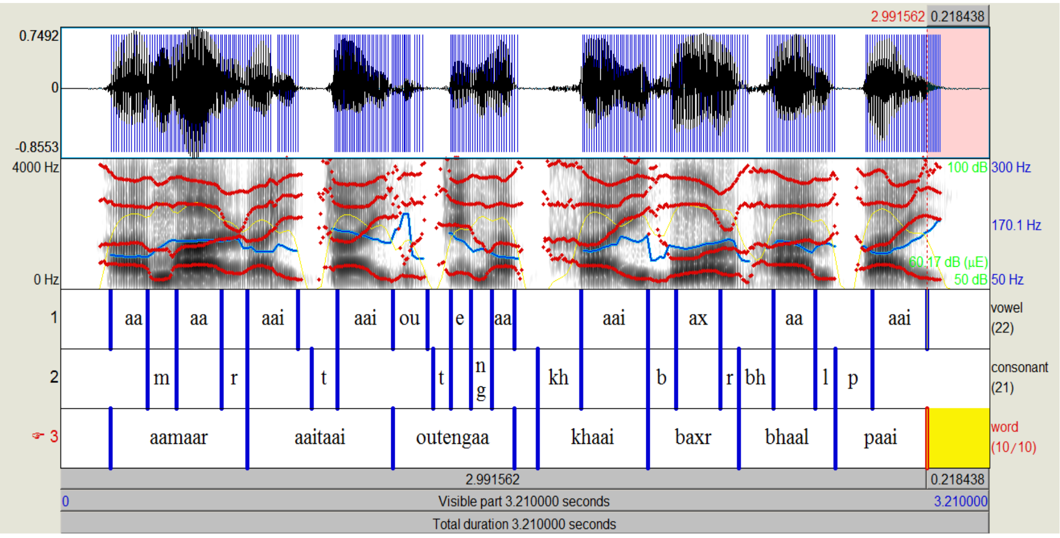
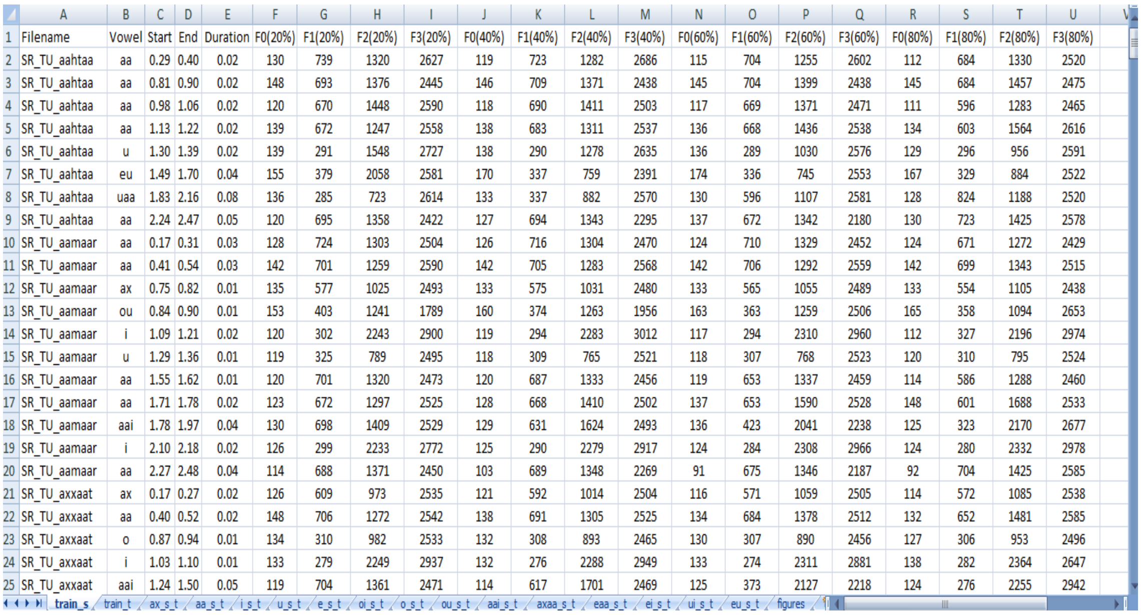


Figure 5.1. A sample textgrid in PRAAT showing annotations in three tiers vowel, consonant and word.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Filename	Vowel	Start	End	Duration	F0(20%)	F1(20%)	F2(20%)	F3(20%)	F0(40%)	F1(40%)	F2(40%)	F3(40%)	F0(60%)	F1(60%)	F2(60%)	F3(60%)	F0(80%)	F1(80%)	F2(80%)	F3(80%)
2	SR_TU_aahtaa	aa	0.29	0.40	0.02	130	739	1320	2627	119	723	1282	2686	115	704	1255	2602	112	684	1330	2520
3	SR_TU_aahtaa	aa	0.81	0.90	0.02	148	693	1376	2445	146	709	1371	2438	145	704	1399	2438	145	684	1457	2475
4	SR_TU_aahtaa	aa	0.98	1.06	0.02	120	670	1448	2590	118	690	1411	2503	117	669	1371	2471	111	596	1283	2465
5	SR_TU_aahtaa	aa	1.13	1.22	0.02	139	672	1247	2558	138	683	1311	2537	136	668	1436	2538	134	603	1564	2616
6	SR_TU_aahtaa	u	1.30	1.39	0.02	139	291	1548	2727	138	290	1278	2635	136	289	1030	2576	129	296	956	2591
7	SR_TU_aahtaa	eu	1.49	1.70	0.04	155	379	2058	2581	170	337	759	2391	174	336	745	2553	167	329	884	2522
8	SR_TU_aahtaa	uaa	1.83	2.16	0.08	136	285	723	2614	133	337	882	2570	130	596	1107	2581	128	824	1188	2520
9	SR_TU_aahtaa	aa	2.24	2.47	0.05	120	695	1358	2422	127	694	1343	2295	137	672	1342	2180	130	723	1425	2578
10	SR_TU_aamaar	aa	0.17	0.31	0.03	128	724	1303	2504	126	716	1304	2470	124	710	1329	2452	124	671	1272	2429
11	SR_TU_aamaar	aa	0.41	0.54	0.03	142	701	1259	2590	142	705	1283	2568	142	706	1292	2559	142	699	1343	2515
12	SR_TU_aamaar	ax	0.75	0.82	0.01	135	577	1025	2493	133	575	1031	2480	133	565	1055	2489	133	554	1105	2438
13	SR_TU_aamaar	ou	0.84	0.90	0.01	153	403	1241	1789	160	374	1263	1956	163	363	1259	2506	165	358	1094	2653
14	SR_TU_aamaar	i	1.09	1.21	0.02	120	302	2243	2900	119	294	2283	3012	117	294	2310	2960	112	327	2196	2974
15	SR_TU_aamaar	u	1.29	1.36	0.01	119	325	789	2495	118	309	765	2521	118	307	768	2523	120	310	795	2524
16	SR_TU_aamaar	aa	1.55	1.62	0.01	120	701	1320	2473	120	687	1333	2456	119	653	1337	2459	114	586	1288	2460
17	SR_TU_aamaar	aa	1.71	1.78	0.02	123	672	1297	2525	128	668	1410	2502	137	653	1590	2528	148	601	1688	2533
18	SR_TU_aamaar	aai	1.78	1.97	0.04	130	698	1409	2529	129	631	1624	2493	136	423	2041	2238	125	323	2170	2677
19	SR_TU_aamaar	i	2.10	2.18	0.02	126	299	2233	2772	125	290	2279	2917	124	284	2308	2966	124	280	2332	2978
20	SR_TU_aamaar	aa	2.27	2.48	0.04	114	688	1371	2450	103	689	1348	2269	91	675	1346	2187	92	704	1425	2585
21	SR_TU_axxaat	ax	0.17	0.27	0.02	126	609	973	2535	121	592	1014	2504	116	571	1059	2505	114	572	1085	2538
22	SR_TU_axxaat	aa	0.40	0.52	0.02	148	706	1272	2542	138	691	1305	2525	134	684	1378	2512	132	652	1481	2585
23	SR_TU_axxaat	o	0.87	0.94	0.01	134	310	982	2533	132	308	893	2465	130	307	890	2456	127	306	953	2496
24	SR_TU_axxaat	i	1.03	1.10	0.01	133	279	2249	2937	132	276	2288	2949	133	274	2311	2881	138	282	2364	2647
25	SR_TU_axxaat	aai	1.24	1.50	0.05	119	704	1361	2471	114	617	1701	2469	125	373	2127	2218	124	276	2255	2942

Figure 5.2. A screenshot of the Excel sheet showing source formant data at 20%, 40%, 60% and 80% of vowel/diphthong duration.

of the vowels/diphthongs. It is also seen that F1, F2 values of target vowels deviates more from their means compared to the F1, F2 values of source vowels, while F2 values of the source diphthongs deviates more from their means than F2 values of target diphthongs. In short, the vowel space for source is different from that of target and can be a candidate for transformation. Our study attempts to carry out this transformation of formants using GMMs.

For comparison of vowel spaces, formant data is extracted from speech data

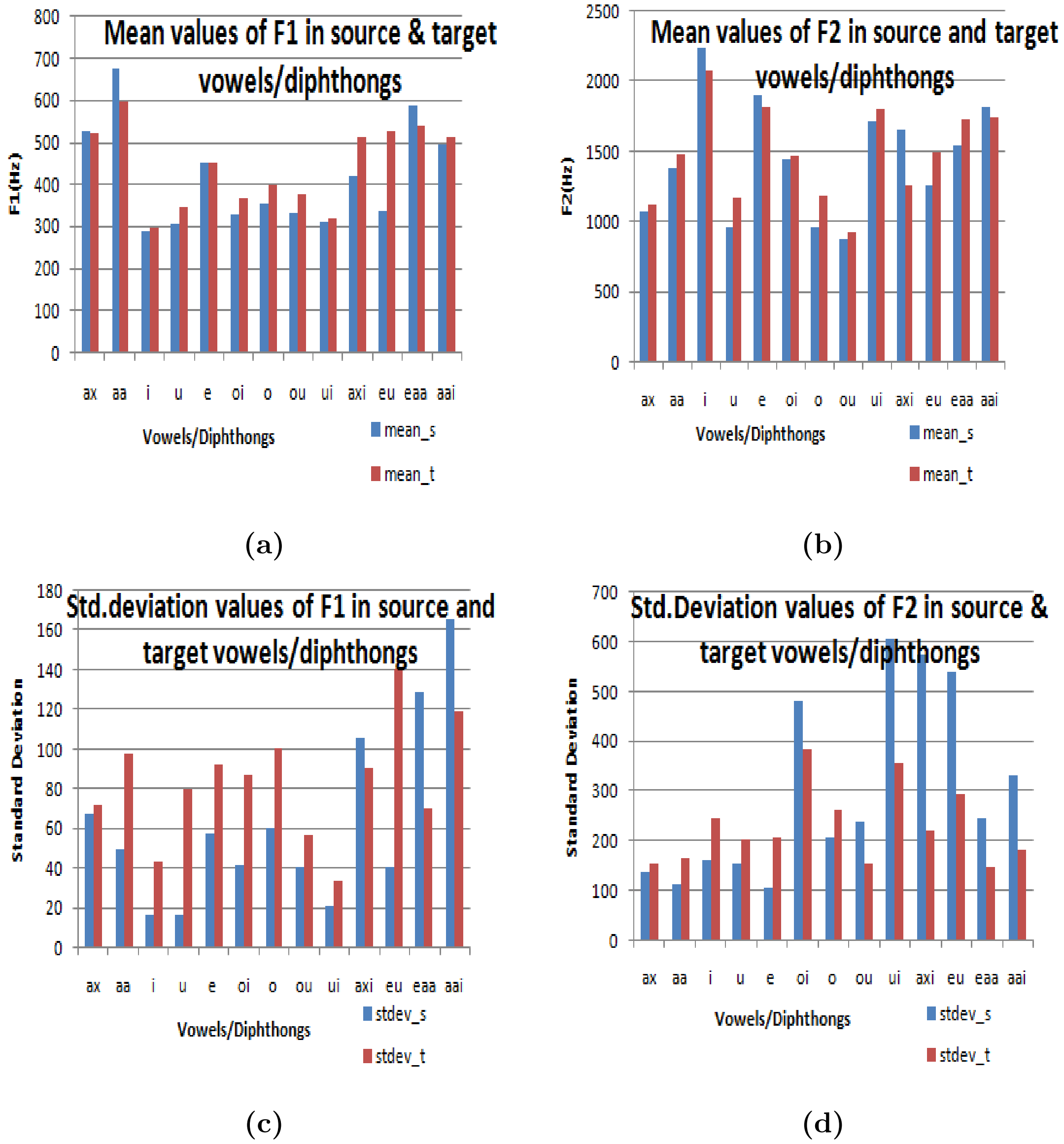
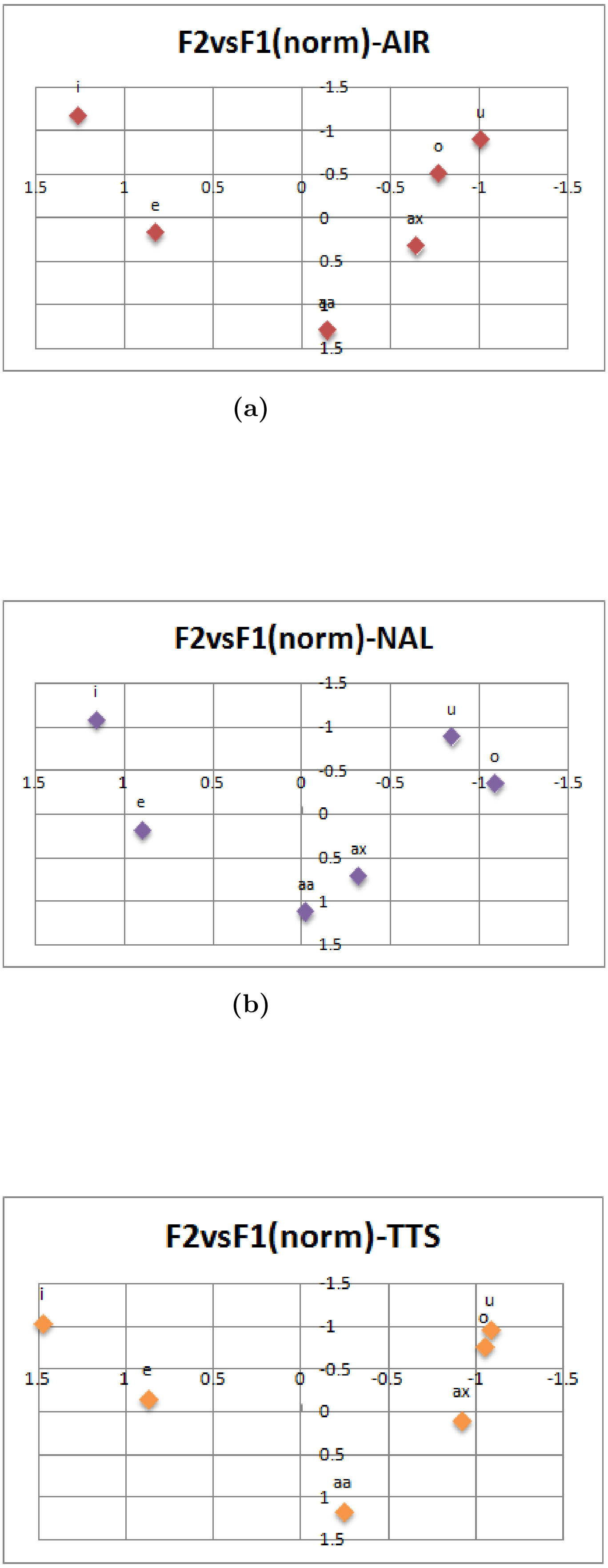


Figure 5.3. Mean & Std.Deviation of F1,F2 formant frequencies in Source and Target Vowels/Diphthongs

collected from two speakers in AIR, two in Nalbaria and also from speech generated from two Text-to-Speech systems, IITG-HTS (developed at IIT Guwahati) and TU-TTS (developed at Tezpur University). In order to permit accurate cross-speaker comparisons of vowels space layout, the formant data is normalised using the Lobanov measure [38] of normalisation. When extrinsic normalisation is applied to acoustic vowel data, the differences due to speakers can often be substantially reduced and Lobanov normalisation is a very basic and effective extrinsic normalisation technique [3] which works well to eliminate physiologically caused differences in formant values. In this method formant values are normalised by subtracting a

[3] https://www.phonetik.uni-muenchen.de/~jmh/research/pasc010808/old/ch3.pdf

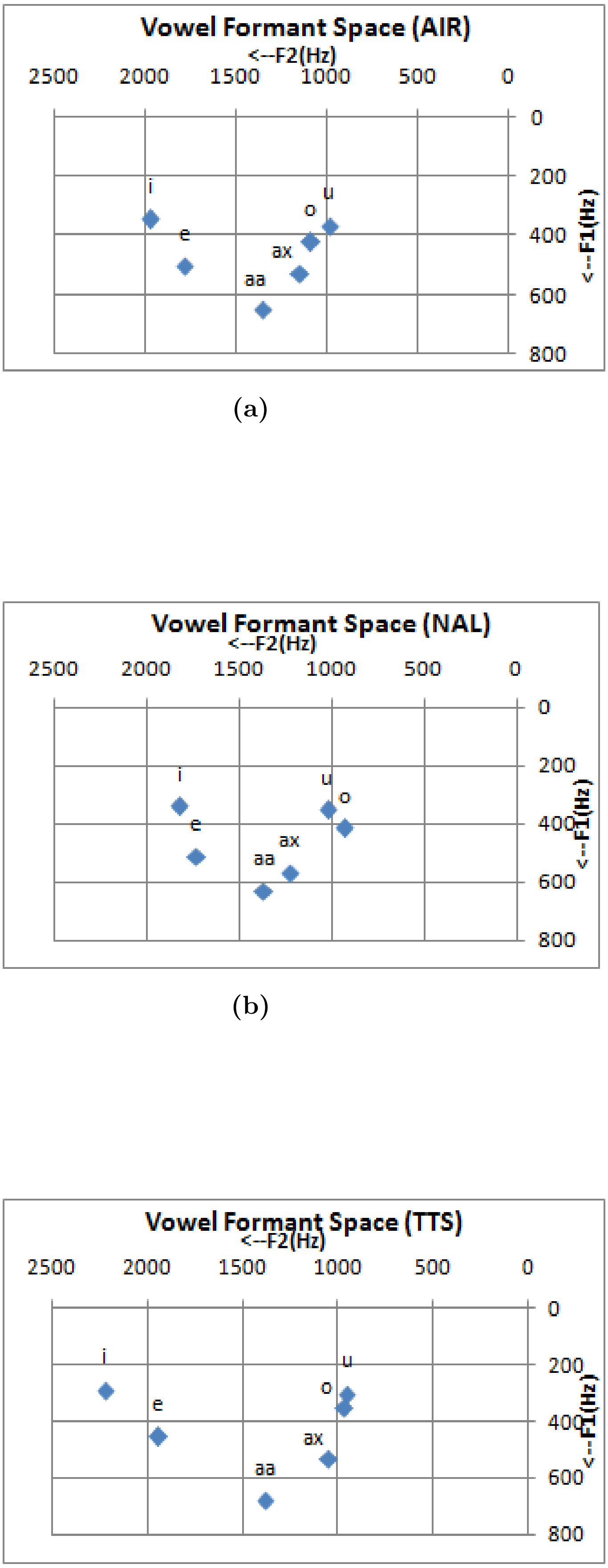


(a)

(b)

(c)

Figure 5.4. Normalised F2 vs F1 plot for (a) AIR, (b) NAL, (c) TTS



(a)

(b)

(c)

Figure 5.5. F2 vs F1 plot (unnormalised) for (a) AIR, (b) NAL, (c) TTS

speaker's mean formant value (μ_i) across all vowel tokens and then dividing by the standard deviation for the formant across all vowels (σ_i) of formant order 'i' for that speaker using Equation 5.1. However, normalization may eliminate dialectal differences as well and therefore if no significant differences exist among the speakers in the lengths of their vocal tracts, comparison of vowel space areas without normalisation would be a better choice [39].

The normalised vowel spaces for AIR (VS-AIR), Nalbaria (VS-NAL) and Nalbaria synthesised from standard AIR TTSs, (VS-TTS), are plotted and presented in Figure 5.4a, Figure 5.4b and Figure 5.4c. Since all our speakers are adult males we do not expect significant differences in the length of their vocal tracts and therefore the unnormalised vowel spaces for AIR, Nalbaria and Nalbaria synthesised from standard AIR TTSs, are also plotted for comparison and presented in Figure 5.5a, Figure 5.5b and Figure 5.5c. It can be observed that VS-TTS is similar to VS-AIR in both the cases. This is because, although the input to TU-TTS is a set of transcriptions of Nalbaria utterances, the TTS has been trained with the phonetics, pronunciation and syllabification rules pertaining to the AIR variety and therefore the output of TU-TTS is of the AIR variety, which also implies that the vowels generated by TU-TTS are similar to the vowels in AIR in terms of formant frequencies.

We also calculate the areas of VS-AIR, VS-NAL and VS-TTS. Vowel space area (VSA) refers to the two-dimensional area bounded by lines connecting F1, F2 coordinates of vowels. Typically VSA is computed by making static measurements of the F1,F2 values for each of corner or cardinal vowels at the the mid-point, for several productions of each vowel. A cardinal vowel refers to a vowel sound produced when the tongue is in an extreme position, i.e., either front or back, high or low [4]. Therefore the area of usually the quadrilateral or trapezoid, formed by the corner vowels is computed using the mean F1,F2 value for each of the vowels at the extreme points of articulation. From the vowel spaces presented in Figure 5.4a, Figure 5.4b and Figure 5.4c, it is seen that the corner vowels (four in number) for VS-AIR and VS-TTS are /u/, /ax/, /aa/ and /i/ while the corner vowels (five in number) for VS-NAL are /u/, /o/, /aa/, /e/, and /i/. We have written a MATLAB program to compute the area of VS-AIR, VS-NAL and VS-TTS using F1, F2 values of these corner vowels. VSA for TTS (=3.06) is the largest, followed by the VSA for NAL (=2.99), and VSA for AIR is the smallest (=2.73). Since frequencies of the first and second formants roughly relate to the size and shape of

[4] https://en.wikipedia.org/wiki/Cardinal_vowels

the cavities created by the opening of the jaw, i.e., vowel height (F1) and tongue position (F2), the VSA gives an indication of how displaced the articulators are while producing the vowels. In general, studies have shown that VSA is larger in speech that is clearer and more intelligible than speech associated with smaller VSAs since a larger VSA corresponds to greater articulatory excursions and more distinct acoustic-articulatory vowel targets [124]. From this comparison we can therefore infer that vowels in the NAL variety of Assamese are more pronounced than those of AIR but those of TTS are the most distinct.

$$F_i^N = (F_i - \mu_i)/\sigma_i \tag{5.1}$$

To evaluate how far or close the vowels are in the three vowel spaces, i.e., VS-AIR, VS-NAL and VS-TTS, we measure the Euclidean distance between like vowels in the three vowel spaces using Equation 5.2 where ‘v1’ represents the first variety and ‘v2’ represents the second variety. The vowel distance is measured with respect to their positions in the vowel spaces, in terms of formant frequencies F1 and F2. In other words we measure the distance between /u/ in VS-AIR and /u/ VS-NAL, and also the distance between /u/ in VS-TTS and /u/ in VS-NAL. Likewise for the other five vowels. The vowel distances are shown in Figure 5.6 in order to get a better visualisation of how distant the vowels in AIR are from their corresponding counterparts in NAL, and how distant the vowels in NAL are from their corresponding counterparts generated by the TTS for the standard variety of Assamese. It is observed from Figure 5.6 that the vowels in the two varieties, i.e., AIR and NAL are different in terms of their positions in their respective vowel spaces. Furthermore, the distance between vowels in NAL and the TU-TTS generated vowels is much higher.

$$D = \sqrt{(F1_{v1} - F1_{v2})^2 + (F2_{v1} - F2_{v2})^2} \tag{5.2}$$

Another noticeable difference between source and target formant contours is with respect to the formant contours of the diphthongs. It can be seen in Figure 5.7 that the formant contours in the source diphthongs are discontinuous, clearly showing the component vowels, i.e., the primary and secondary vowels, in the diphthongs, while in the target diphthongs the formant contours are continuous.

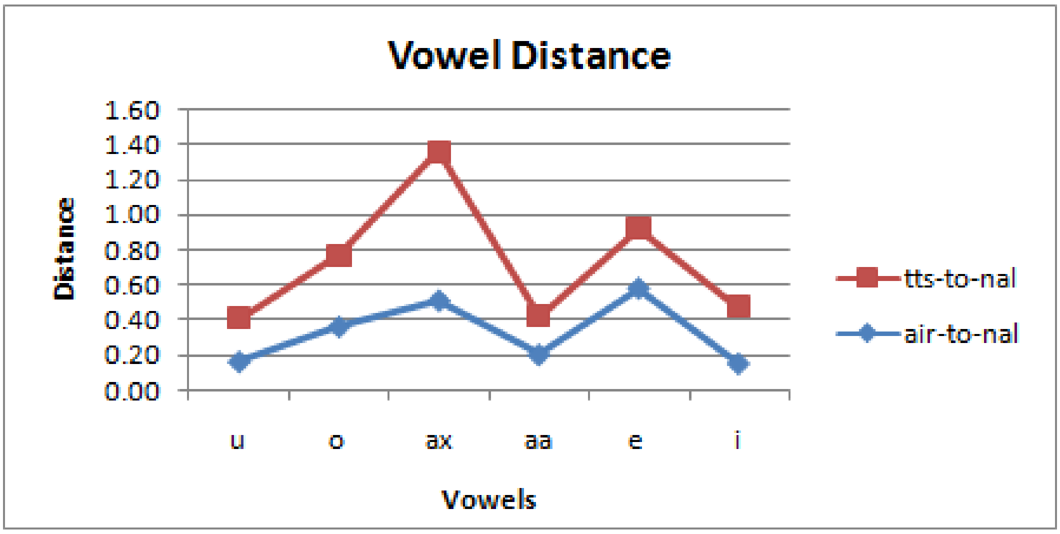


Figure 5.6. Distance between vowels in (a) VS-AIR and VS-NAL, (b) VS-TTS and VS-NAL

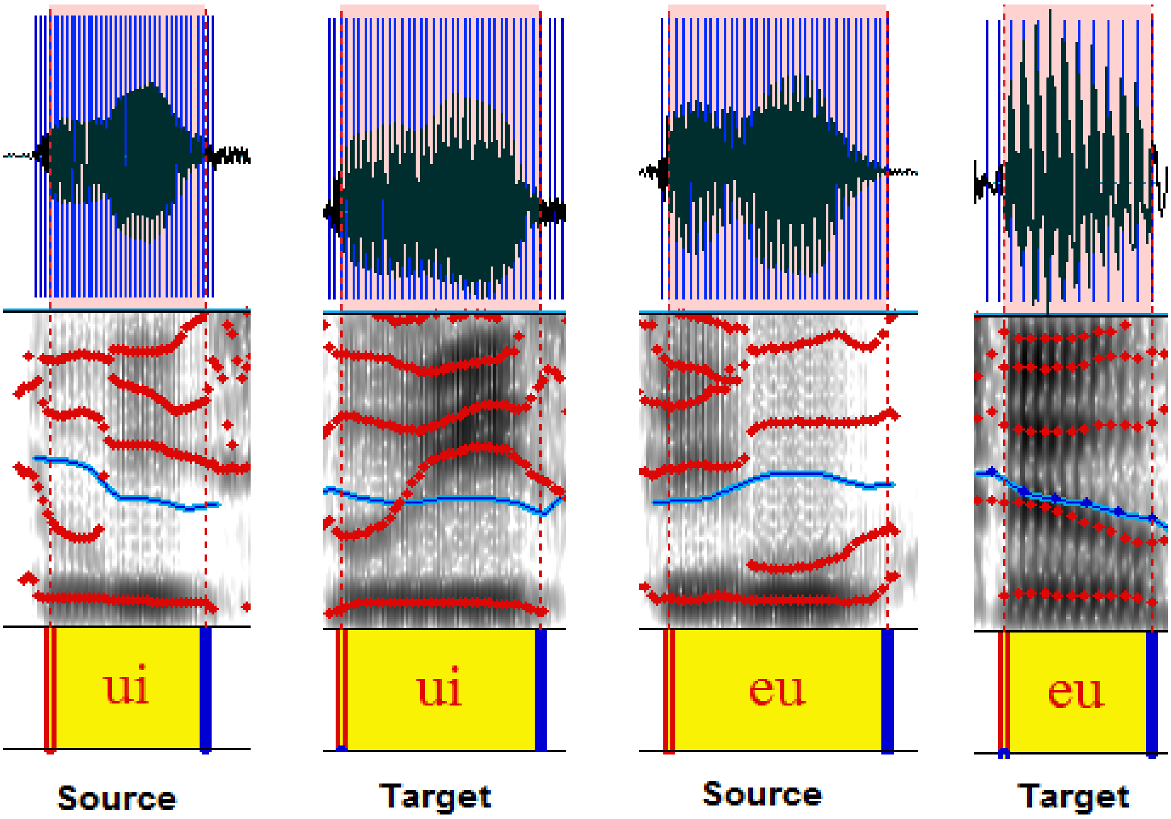


Figure 5.7. Diphthongs /ui/ and /eu/ in Source and Target

5.4.2.2 Fitting a GMM to formant data

Formant data consisting of F1, F2 and F3 formant frequencies of vowels and diphthongs, from both source and target speech data is stored in an excel file. A

MATLAB script is written to read F1-F2 data from the excel file, plot scatter plots and cluster source and target formant data separately using GMM clustering. Results of the clustering are presented in Figure 5.8 and Figure 5.9. It is visually observed from the marked clusters that the formant data for the different vowels for the source, i.e. the AIR variety, is well separated and distinguishable, while the clusters are overlapping in target formant data. This would imply that soft clustering of the data would be more appropriate instead of giving the data points a hard assignment to exactly one cluster. GMM is a soft clustering method whereby a data point can be assigned to more than one cluster by assigning probabilities of belonging to each of the clusters. This makes GMM a good option for modelling our formant data. We also try to fit a separate GMM to each of source and target formant data with equal number of mixtures using the *'fitgmdist'* function of MATLAB. Results of this fit with number of mixtures set to 12, can be observed by plotting the formant data over the fitted GMM contours. This is done using the *'gscatter'* function for plotting the data and the *'ezcontour'* function for plotting the contours. The resulting plots are presented in Figure 5.10.

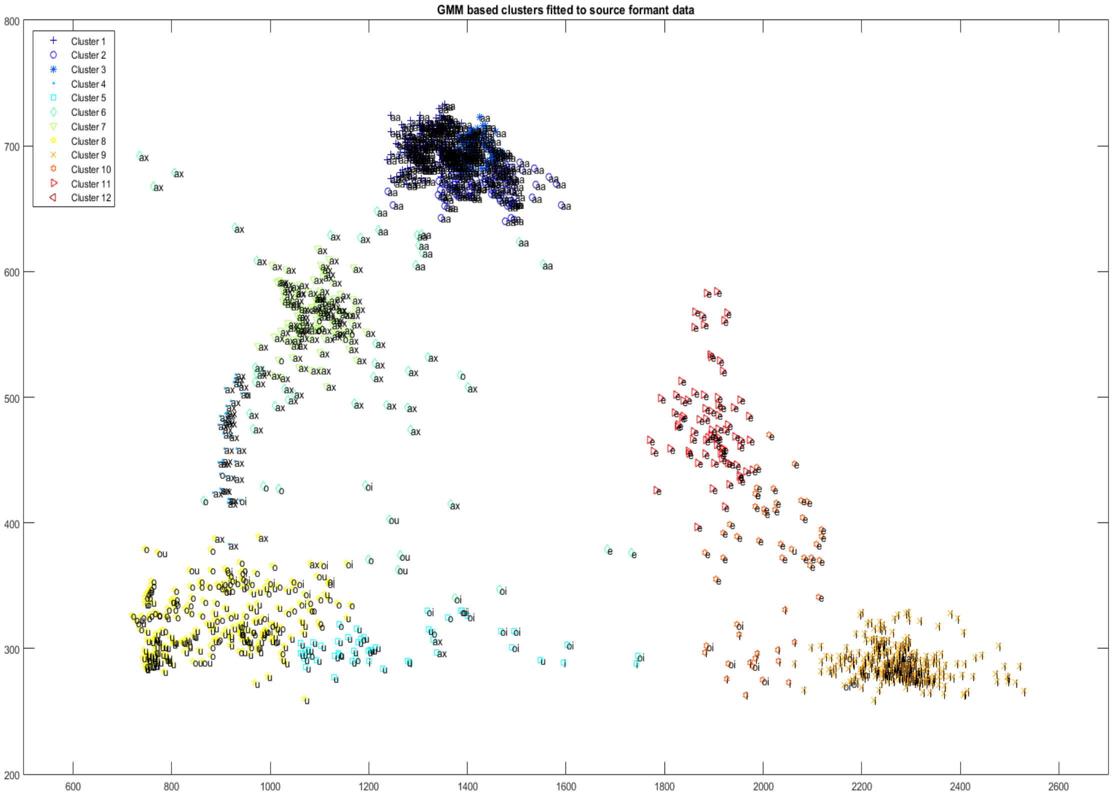


Figure 5.8. GMM clusters in Source Vowel Formant Data

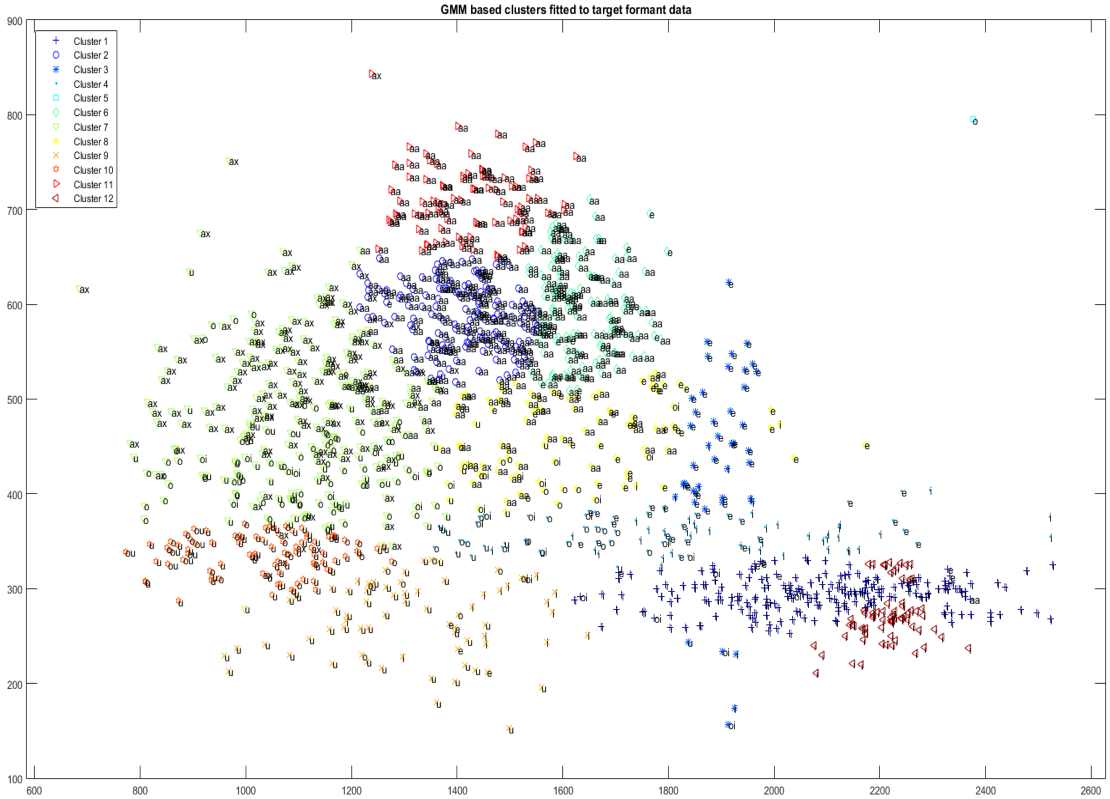


Figure 5.9. GMM clusters in Target Vowel Formant Data

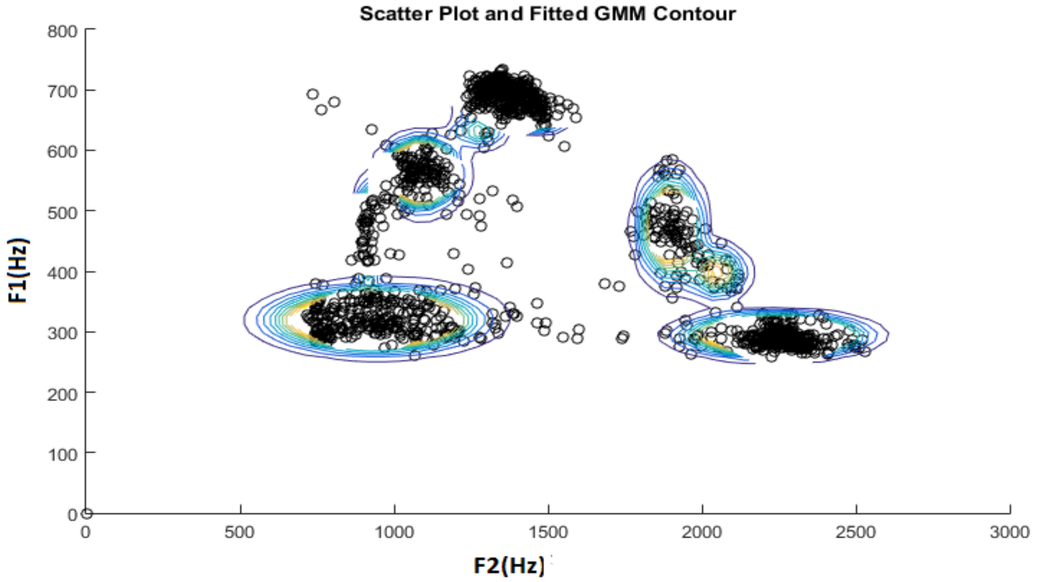
5.4.2.3 Development of the Mapping Function

In a GMM based transformation, initially proposed for VC by Stylianou et al. [135], the system learns by fitting a GMM model to the augmented source and target feature vectors. During training, a GMM is adopted to model the distribution of the paired feature sequence z_t, representing the joint feature vector of source speech vector x_t and target speech vector y_t at frame t. The Expectation Maximization (EM) algorithm[5] trains the GMM with the joint source and target vectors which are already aligned since the formants are extracted at four equidistant points in every vowel/diphthong segment.

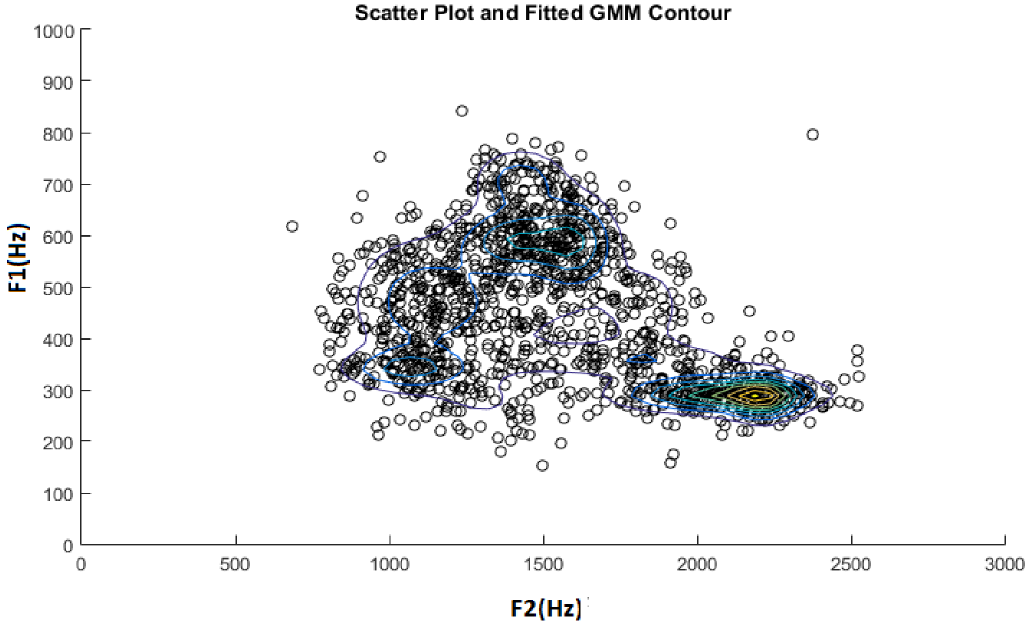
The mapping function as described by Toda et al. [142], converting the formant frequencies, of source to target speech data is given by

$$F(x_t) = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) E_{m,t}^{(y)} \tag{5.3}$$

[5] http://www.cse.iitm.ac.in/~vplab/courses/DVP/PDF/gmm.pdf



(a)



(b)

Figure 5.10. Scatter Plot with GMM contours fitted to (a) Source Vowel Formant Data and (b) Target Vowel Formant Data

where

$$P(m|x_t, \lambda^{(z)}) = \frac{w_m N(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} w_n N(x_t; \mu_n^{(x)}, \Sigma_n^{(xx)})} \tag{5.4}$$

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} (x_t - \mu_m^{(x)}) \tag{5.5}$$

$N(x; \mu, \Sigma)$ denotes normal distribution with mean vector μ and covariance matrix Σ. The total number of mixture components is M and the weight of the m^{th} mixture component is w_m. $\mu_m^{(x)}$ and $\mu_m^{(y)}$ are the mean vectors, $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yy)}$ are the covariance matrices, $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$ are the cross-covariance matrices of the m^{th} mixture component of source and target feature vectors.

F1, F2, F3 formants are extracted from all the vowels/diphthongs of source and target speech data, at 20%, 40%, 60% and 80% of vowel/diphthong duration to form the three-dimensional feature vector. This gives four samples each for every occurrence of a vowel/diphthong segment. So if the number of vowels/diphthongs is 'm' and each vowel/diphthong occurs 'n' times, the number of rows in the feature matrix will be 'm x n x 4' and the number of columns will be 3. The augmented feature vector consists of F1, F2, F3 values of vowel/diphthong segments in both source and target speech data. The three approaches, based on GMMs, used to carry out the transformation of formants, are implemented in MATLAB. The mapping functions are developed to transform F1, F2 and F3 formants from one variety of Assamese to another. However since F3 transformation is not significant, transformation results presented in Section 5, are confined to F1, F2 only. Following subsections describe the three approaches in brief.

Approach 1: A single GMM for F1, F2, F3 formants The first three formant frequencies, F1, F2 and F3, extracted from all the vowels/diphthongs of source and target speech data, at 20%, 40%, 60% and 80% of vowel/diphthong duration, form the feature vector. The augmented feature vector consists of F1, F2 and F3 values of vowel/diphthong segments in both source and target training speech data and a GMM is trained to model this data. The EM algorithm is then used to determine the model parameters which are used to develop the mapping function to map the formants from the source to the target.

Approach 2: Separate GMMs for F1, F2, F3 at equidistant temporal points of vowel /diphthong segments The second approach uses four separate mapping functions for the formants at the four temporal points 20%, 40%, 60% and 80% of vowel/diphthong duration. This is done assuming that formants at a particular time point may exhibit similar behavioral characteristics. So, F1, F2 and F3 extracted at 20% of vowel/diphthong duration, form one feature vector, those extracted at 40% form another feature vector, and so on. This gives us four feature vectors to train the four mapping functions. In order to estimate the parameters of the mapping functions, the probability distributions of the joint vectors $z_i = [x_i^T, y_i^T]$ are represented by separate GMMs. x_i and y_i are vectors containing the source and target formant frequencies F1, F2, F3 at i=20%, 40%, 60% and 80% of vowel/diphthong duration and 'T' represents the transposition of the vector. Each mapping function maps the formants at the specified time instant. This means we have a separate mapping function to map the formants at 20% of vowel/diphthong duration, another to map the formants at 40% of vowel/diphthong duration, and so on.

Approach 3: Separate GMM for each vowel/diphthong segment In this approach, label files consisting of vowel/diphthong labels are assigned to the utterances. These label files are used to cluster extracted formants from training speech data into separate vowel/diphthong groups. A separate GMM is built for each vowel/diphthong and a mapping function is developed for each of the vowels and diphthongs. The number of mixture components (M) for each vowel/diphthong GMM is selected after experimenting with different values for best results. During transformation, the label file associated with the test feature vector specifying the vowel/diphthong identity, is used to select the appropriate mapping function for formant transformation. Transformed formant contours for the vowels/diphthongs in the test utterances are passed through a Moving Average (MA) filter to get smooth trajectory.

5.5 Results and Evaluation

5.5.1 Objective Evaluation using RMSE

The mapping functions are tested with different values of M, to transform formants F1, F2 of the vowels/diphthongs in a set of 10 test utterances. For the first approach best results are obtained with M=16, for the second with M=8, while for the third, M is set to different values (4, 8, 16) for individual vowels/diphthongs. The transformation is evaluated objectively in terms of root mean square error (RMSE) between test and target formants and between mapped and target formants. Transformation results for the test utterance "baamuntui saagaaltuk puihbaaklegi kinsil"

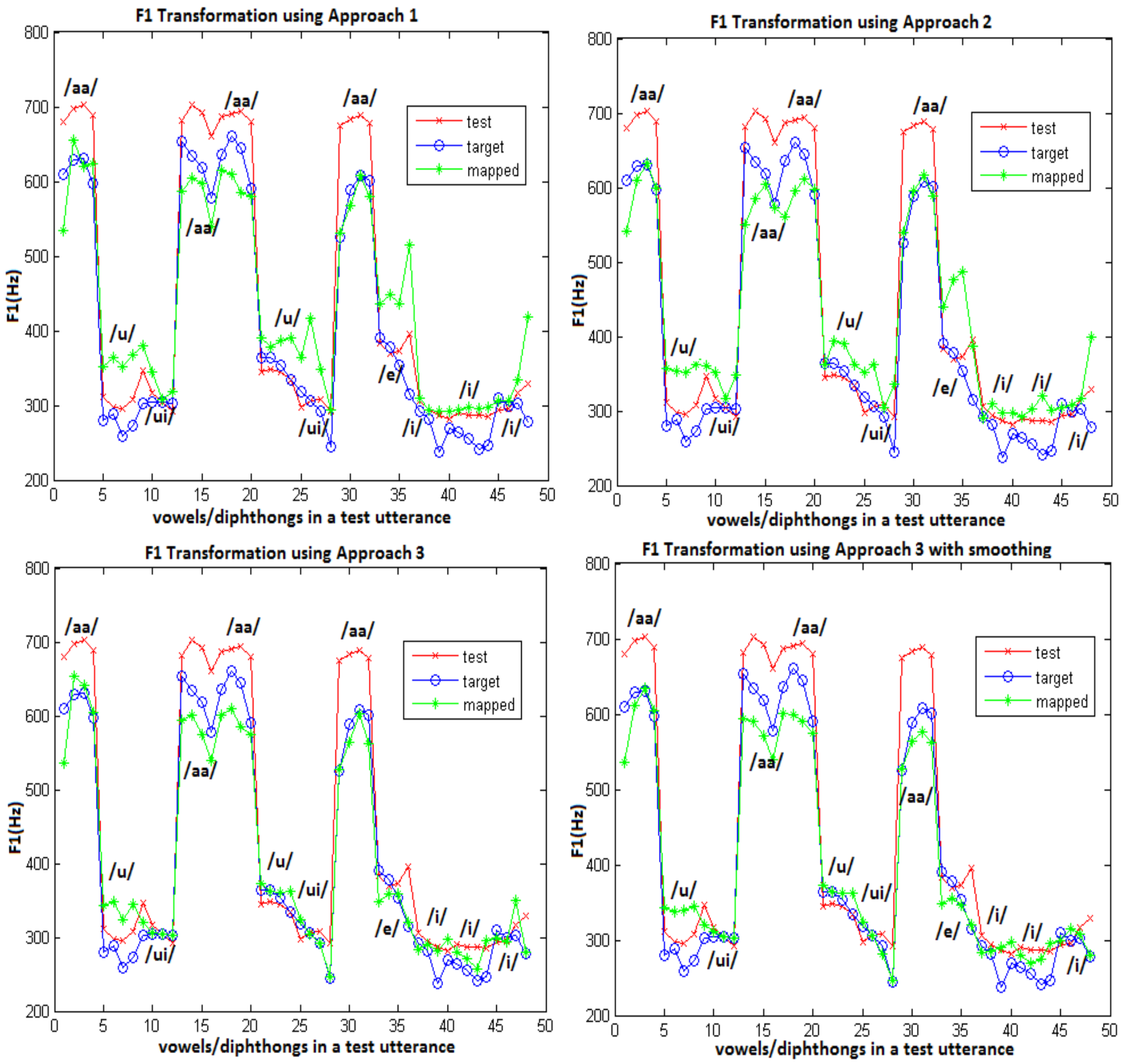


Figure 5.11. Transformation of F1 formant frequencies

having the vowels/diphthongs /aa/, /u/, /ui/, /aa/, /aa/, /u/, /ui/, /aa/, /e/, /i/, /i/, /i/ in order of occurrence, using the three approaches are presented in Figure 5.11 for F1 transformation and Figure 5.12 for F2 transformation. The first three figures present the results of the transformation using the three approaches. The fourth subfigure presents the results of the third approach after smoothing the contours with a MA filter.

In order to provide a quantitative representation of the transformation results, percentage improvement of RMSE values before and after transformation, using the three approaches, are calculated for all vowel/diphthong tokens in the test set of utterances. Figure 5.13 presents a comparison of the results using Approach 1 (single GMM for entire formant training data) with different number of mixture components. Figure 5.14 presents a comparison of the results using the three

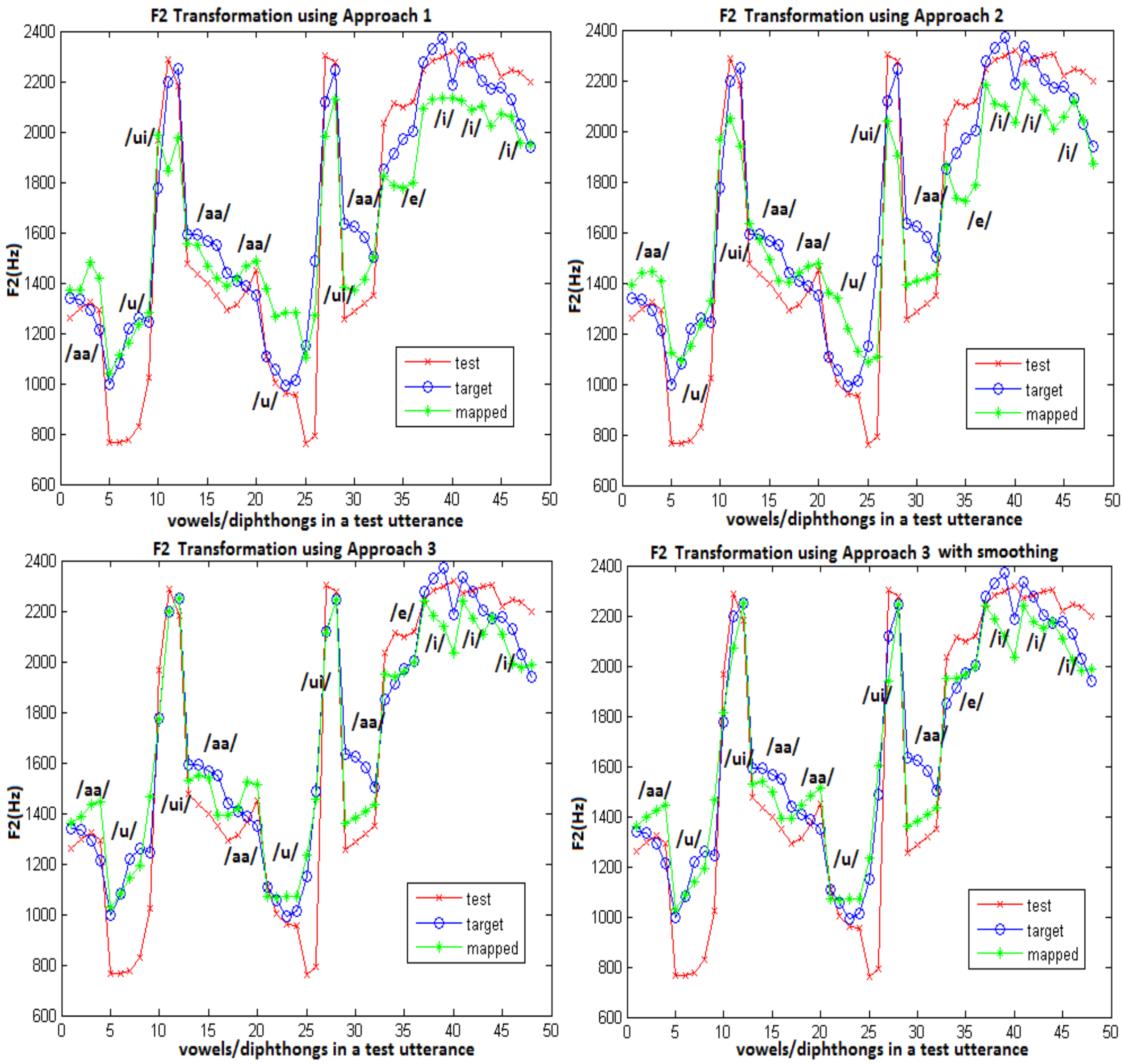


Figure 5.12. Transformation of F2 formant frequencies

approaches.

Furthermore, formants F1, F2 extracted from the vowels occurring in the set of test utterances and in the corresponding set of target utterances are plotted in Figure 5.15 and Figure 5.16. The transformed vowel formants are plotted in Figure 5.17. The transformed vowel space is observed to be much closer to the target vowel space than the source vowel space. The distance between like vowels in the source and target vowel spaces before and after the transformation are plotted in Figure 5.18 for a better visualisation of transformation results. It is observed that the transformation has brought the source vowels closer to the target vowels in the vowel space except for the vowel /o/.

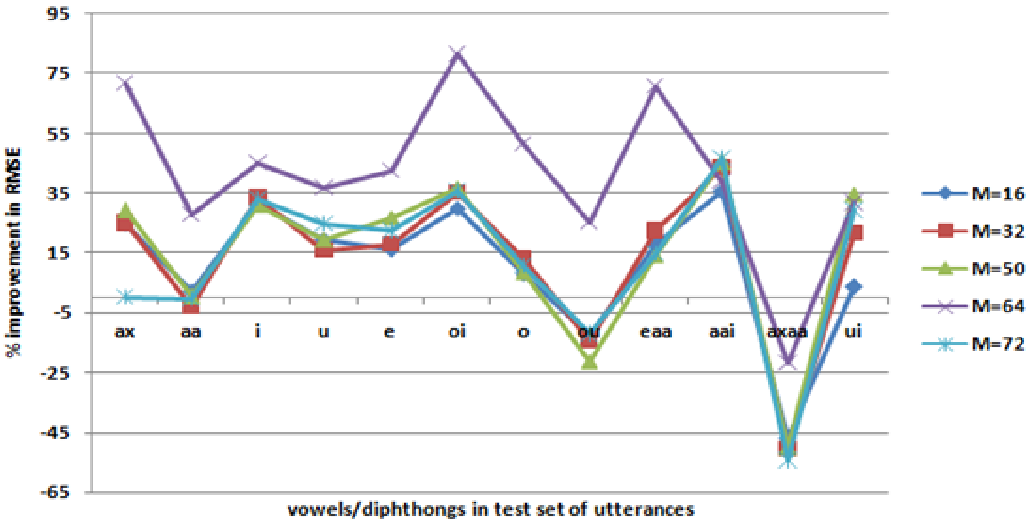


Figure 5.13. Plot showing %improvement in RMSE values after transformation using Approach 1

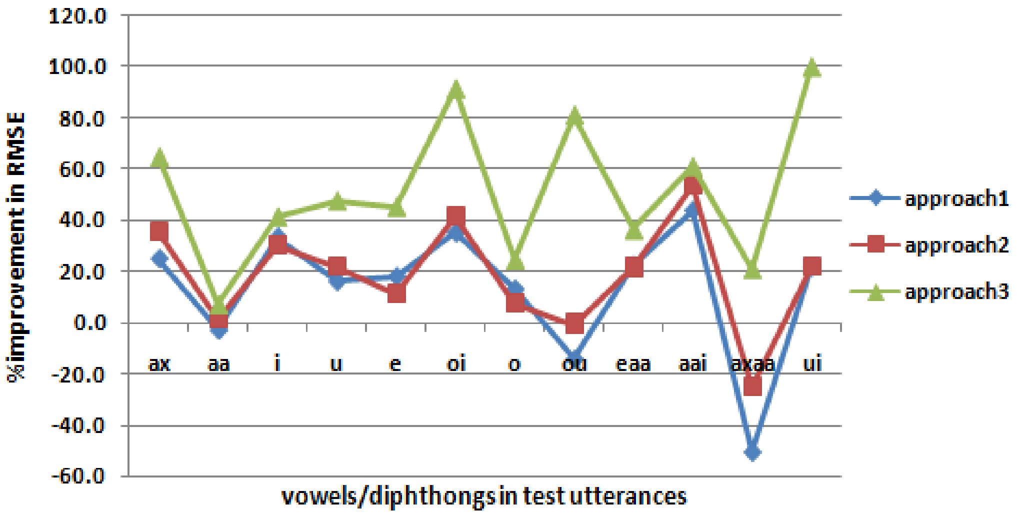


Figure 5.14. Plot showing %improvement in RMSE values after transformation using the 3 approaches

5.5.2 Comparison with the ANN based method of Formant Transformation

Not many works have been reported in the field of Formant Transformation. Probably the earliest is that of Narendranath et al. [97] where ANNs are used to carry out the transformation of formants from source to target. The authors in this work transform the formants of a source speaker to match that of a target speaker for the purpose of converting voice quality. Rentzos et al. [120] report another work where the statistical distributions of formants are modelled by a two-dimensional Hidden

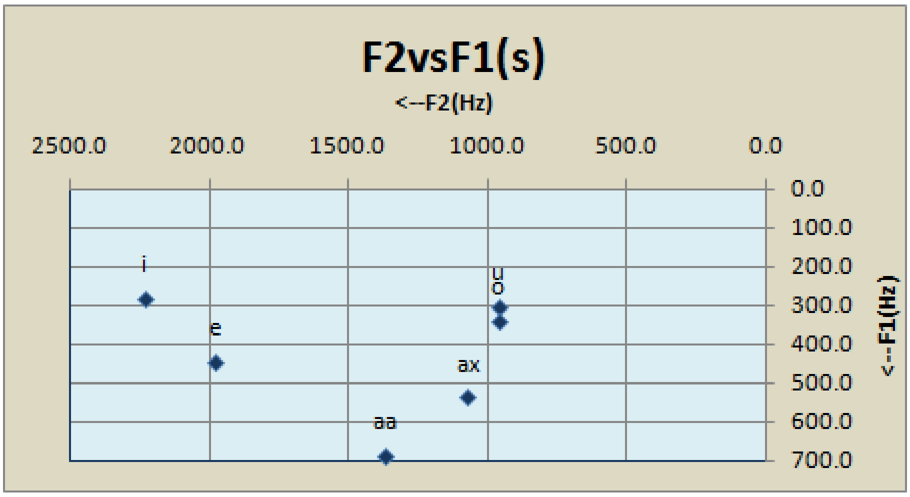


Figure 5.15. F2 vs F1 for Source(s)

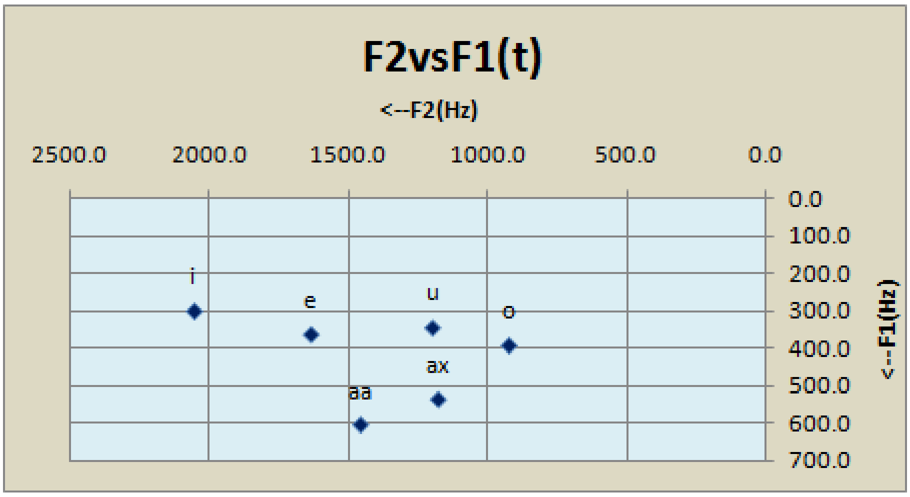


Figure 5.16. F2 vs F1 for Target(t)

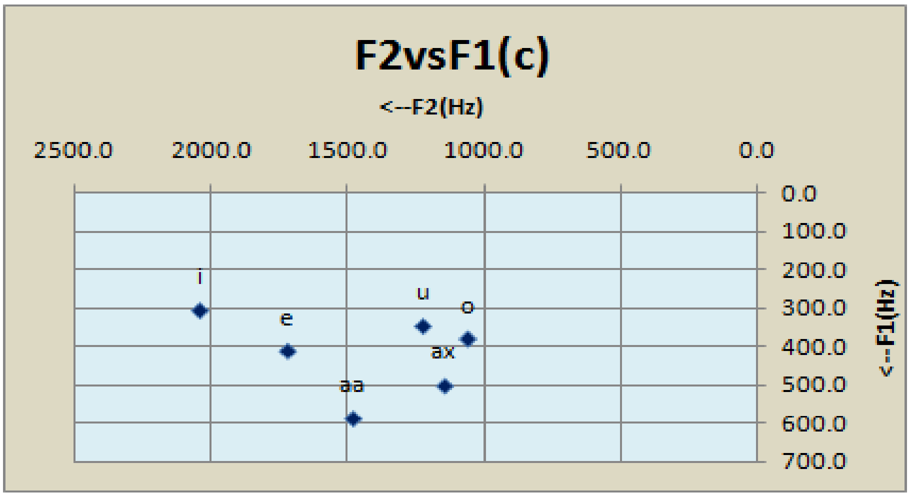


Figure 5.17. F2 vs F1 for Converted(c)

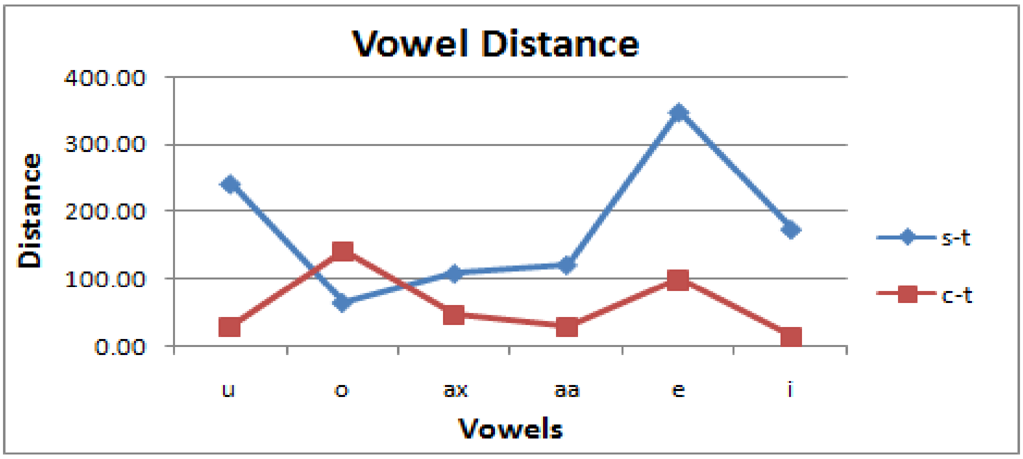


Figure 5.18. Vowel distance after Formant Transformation using the GMM-based Mapping Function

Markov Model (HMM) spanning time and frequency dimensions. Two methods are explored for mapping the source speaker formants to target speaker formants. The first method uses adaptive formant-tracking warping of the frequency response of the LP model. The second method is based on the rotation of poles of the LP model of the speech signal. Both methods are reported to transform all spectral parameters of the resonances at formants of the source speaker towards those of the target speaker. We have taken a different approach for transforming the formants. Our approach models the joint source and target formant data by a GMM or GMMs. As a preliminary comparison, we have compared our results with that of the ANN based method of formant transformation by Narendranath et al. The aim of our work is to explore a different approach for transforming the formants from source to target data and apply the transformation function so developed, to transform the vowel formant space of one variety of Assamese to another. The larger goal is to make the vowels/diphthongs of one variety of Assamese (AIR) sound more like their counterparts in the other variety (NAL).

The formant data, F1, F2 and F3 values are extracted frame wise for every 25ms frame with a 10ms overlap, from the vowels/diphthongs in both source and target sets of utterances. Since the duration of vowels/diphthongs are not equal in the source and target data sets, source and target formant data are aligned using DTW. The warped and joint data is now used by neural networks to learn a mapping function which can transform spectral features from source to target. Various architectures in terms of number of hidden layers, number of neurons in each layer, etc, are tested before settling for the best architecture. Experimentally the best transformation results with a set of ten test utterances, are obtained with two hidden layers having sixty neurons each using the tangent sigmoid transfer function, together with the input and output layer having three neurons each corresponding to F1, F2 and F3. The output layer uses a linear transfer function and the network is trained by a training function which uses the BFGS quasi-Newton method. The ANN architecture is shown in Figure 5.19.

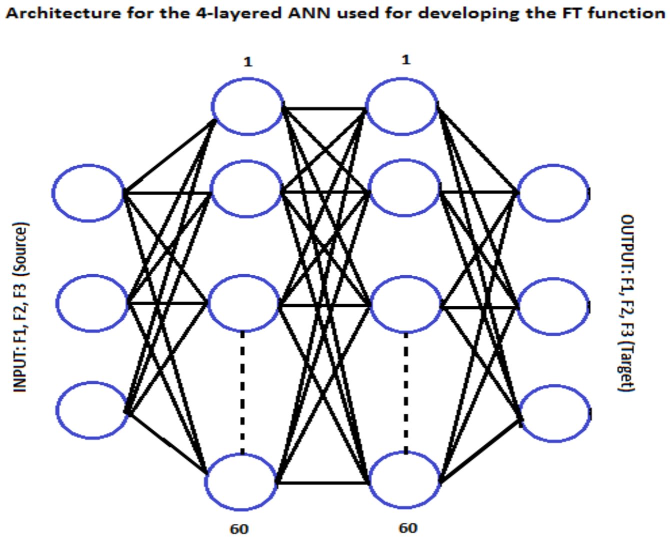

Architecture for the 4-layered ANN used for developing the FT function
1
1
INPUT: F1, F2, F3 (Source)
OUTPUT: F1, F2, F3 (Target)
60
60

Figure 5.19. ANN architecture for formant transformation with 2 hidden layers having 60 neurons each

In order to compare the results so obtained, with the results of our GMM-based approach, the first step is to generate the vowel/diphthong formant contours from the transformed formants at the four temporal points, i.e., at 20%, 40%, 60% and 80% of vowel/diphthong duration. This is done using the cubic method of interpolation where the formants at 20%, 40%, 60% and 80%points are interpolated using cubic splines. The number of points to be used for interpolation is given by the product of the duration of the vowel/diphthong and the sampling rate of the speech wave file. The duration is determined from the label files associated with each utterance containing the vowel/diphthong labels together with the duration. These transformed and regenerated formant contours for the vowels/diphthongs in the test set of utterances are then compared to that transformed by the ANN-based method using the RMSE measure. The average RMSE for each of the vowels/diphthongs categories in the test set of utterances, are then calculated for both the methods and the results are presented graphically in Figure 5.20.

Observation 1: The RMSE values after transforming the formants from source to target, using the ANN-based method and the GMM-based method, for the

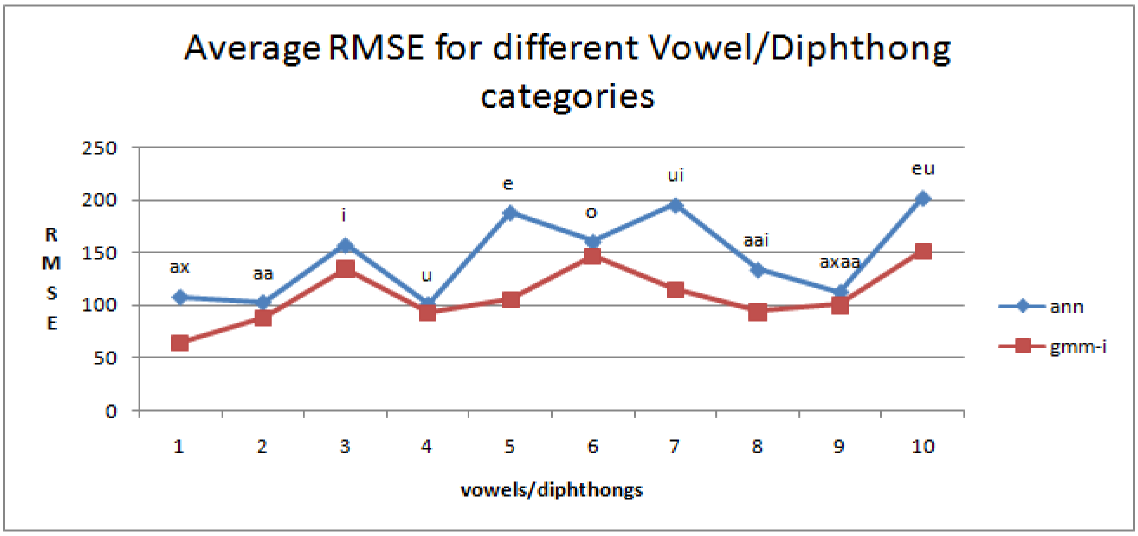


Figure 5.20. RMSE values after transformation using (i) ANN-based & (ii) GMM-based methods

vowel/diphthong segments in three of the test utterances are also presented in the Figures 5.21, 5.22 and 5.23. For most of the vowel/ diphthong segments in these test utterances, the RMSE values after the transformation using the GMM-based method is comparatively reduced.

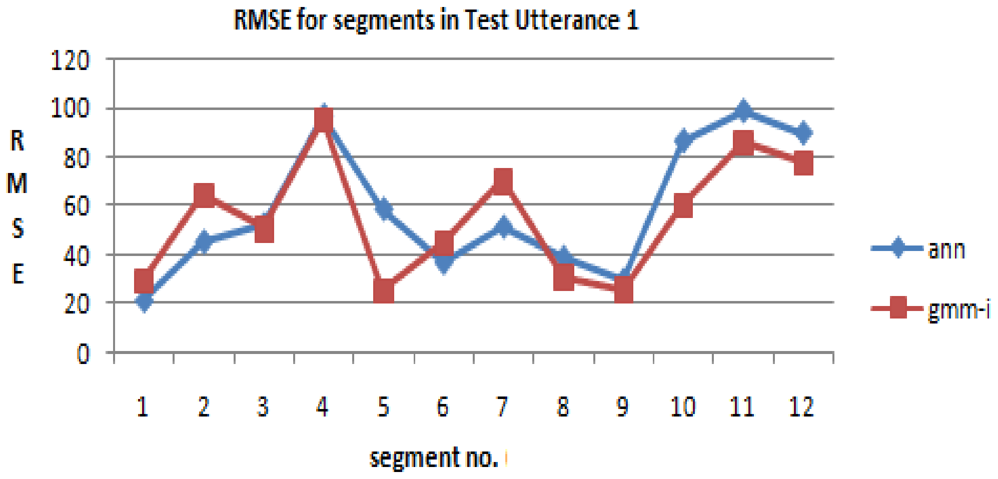


Figure 5.21. RMSE values after transformation using the (i) ANN-based & (ii) GMM-based methods for Utterance 1

Observation 2: The formants F1 & F2, of the vowel/ diphthong segments in a test utterance, after transformation using the ANN-based method and our GMM-based method, are presented in Figure 5.24 and Figure 5.25. Our method performs better, specially for the diphthongs. In this test utterance, the diphthong /ui/

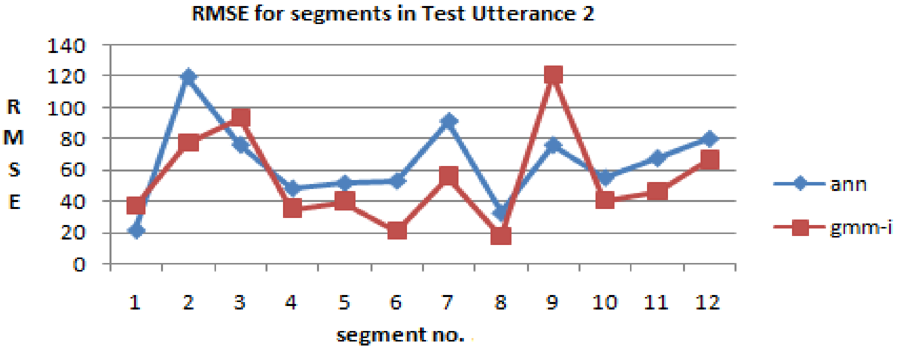


Figure 5.22. RMSE values after transformation using the (i) ANN-based & (ii) GMM-based methods for Utterance 2

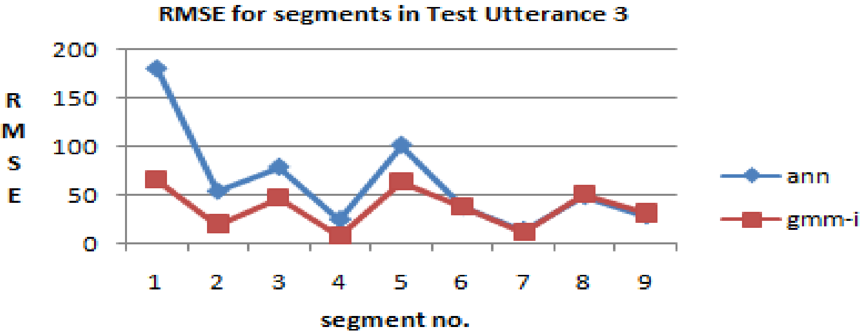


Figure 5.23. RMSE values after transformation using the (i) ANN-based & (ii) GMM-based methods for Utterance 3

occurs twice and is highlighted in green circles. The discontinuity in the formants of the source diphthongs as the formant trajectory moves from the primary vowel to the secondary vowel, is made continuous by the GMM-based transformation function.

5.5.3 Graphical representation of results using SSANOVA

SSANOVA or Smoothing Spline Analysis of Variance is a test that determines whether there are significant differences between curves that are fitted to data sets being compared. For our work we have used the R package *gss* [47] for nonparametric statistical modelling which is a suite of functions implementing smoothing spline ANOVA models. We use SSANOVA for comparing the vowel/diphthong formant contours in two datasets; the first consisting of formant contours extracted from

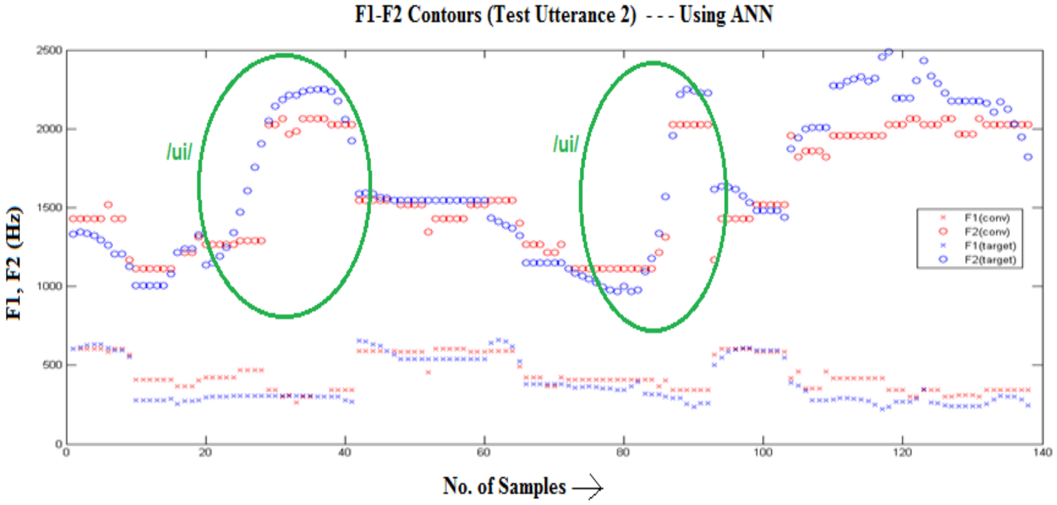


Figure 5.24. Transformation results for a test utterance (utt2) using the ANN-based method

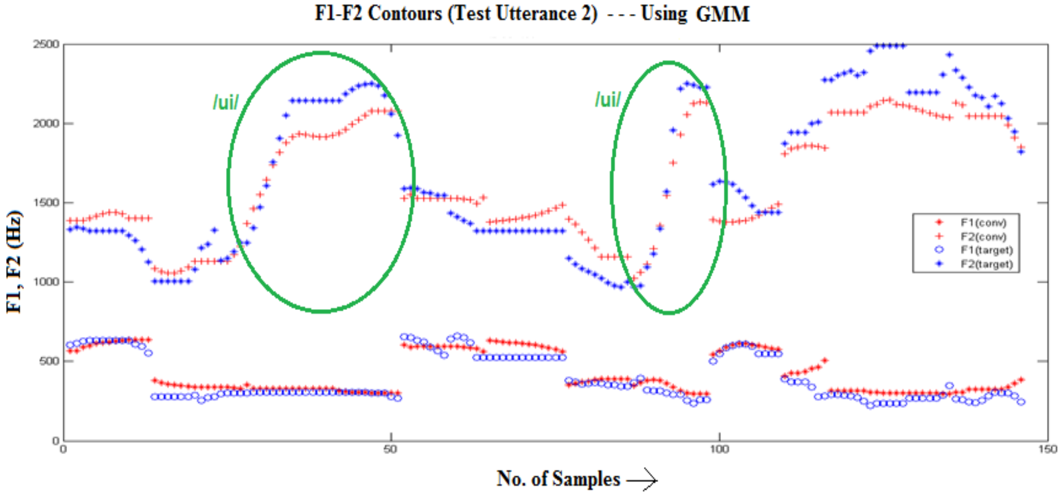


Figure 5.25. Transformation results for a test utterance (utt2) using the GMM-based method

the vowels/diphthongs generated by the TU-TTS (source1) and those extracted from the set of utterances recorded from the Nalbaria speaker (target), the second consisting of formant contours generated from the transformed formants (source2) and those extracted from the set of utterances recorded from the Nalbaria speaker (target). The steps for carrying out the SSANOVA test are listed below:

(i) For source1 and target vowels/diphthongs, formants are extracted frame wise (10ms frame).

(ii) For the transformed formants (source2) at 20%, 40%, 60% and 80% of vowel/

diphthong duration, the formant contours are regenerated by interpolating accordingly.

(iii) Formant data (F1, F2, F3) from the vowels/diphthongs in source1 and target formant contours are provided as input (in the required formant) to the *'compareformants'* function of the *'gss'* package, the output of which are plots with spline estimations and 95% confidence intervals.

(iv) Formant data from the vowels/diphthongs in source2 (converted) and target formant contours are provided as input to the *'compareformants'* function.

Results from the SSANOVA test are presented in Figures 5.27 : 5.35. Each figure contains two sets of three contour lines. The three contour lines show the smoothing spline fit for each of the three formants. Each set represents the F1, F2, F3 contours for either source and target contours or converted and target contours. The pair of dotted lines indicate 95% Bayesian confidence intervals. An overlap of the contours (source and target or converted and target) indicates that the difference between the contours are not significant. It is observed from the smoothing splines also that the transformation has brought the source formant contours closer to the target formant contours. Best results are obtained for the vowels /ax/, /aa/, /i/, /u/, /aai/, /ui/ and /oi/.

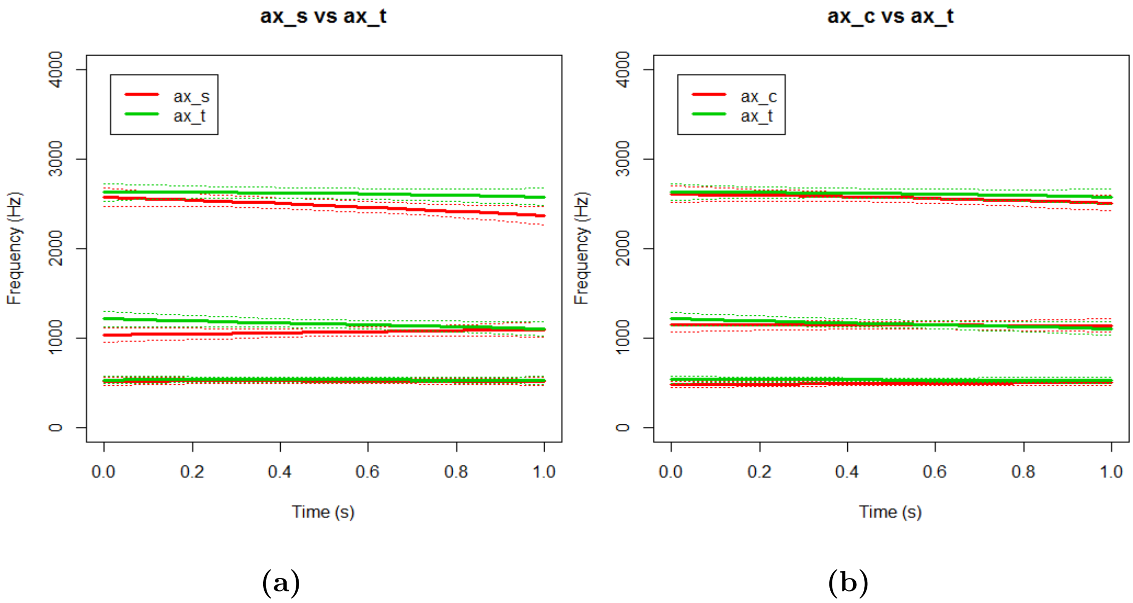
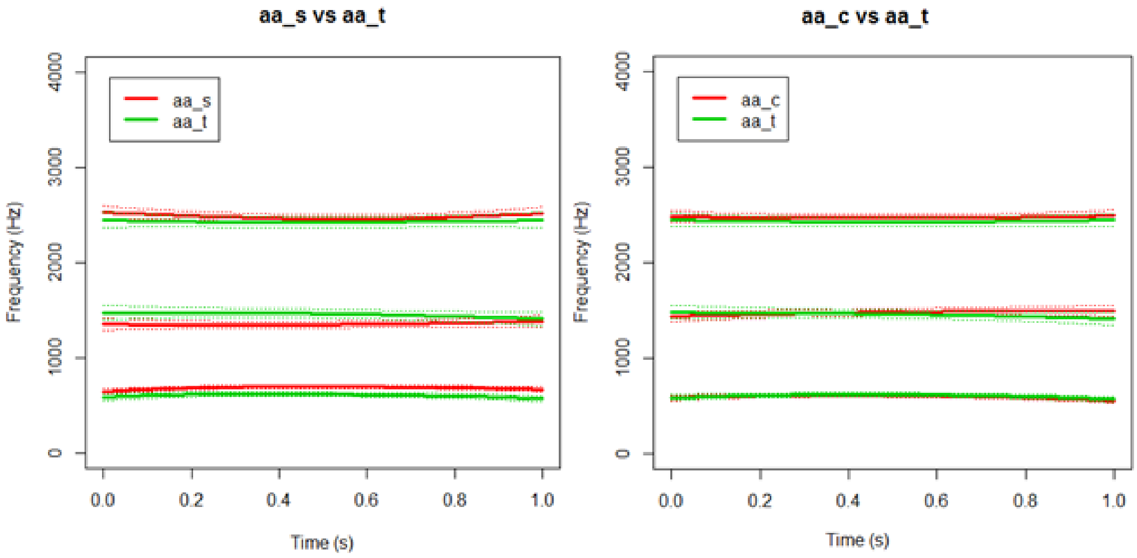
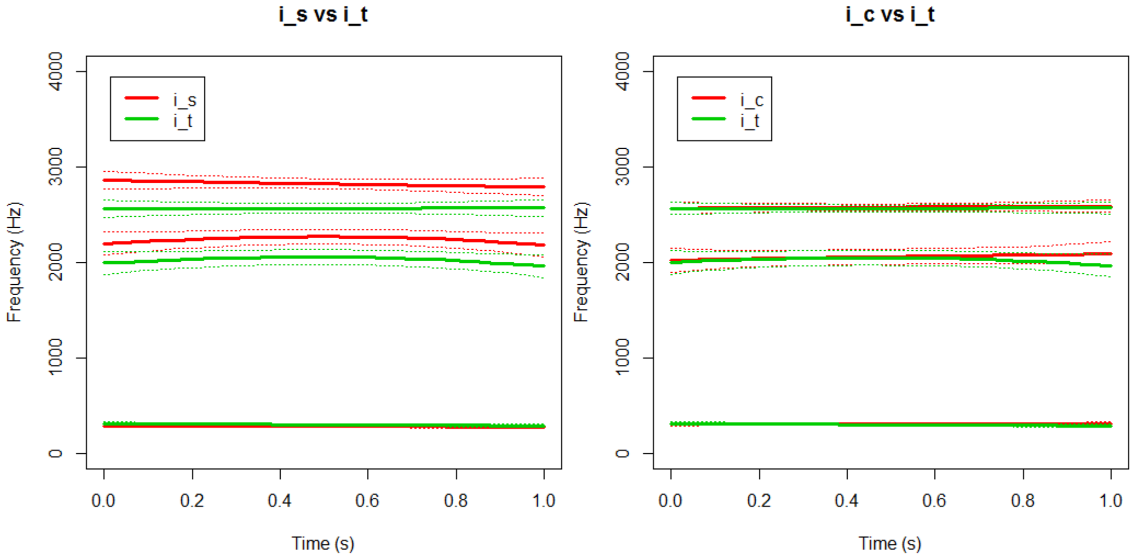


Figure 5.26. Smoothing Splines for vowel /ax/ (a) Source vs Target (b) Converted vs Target



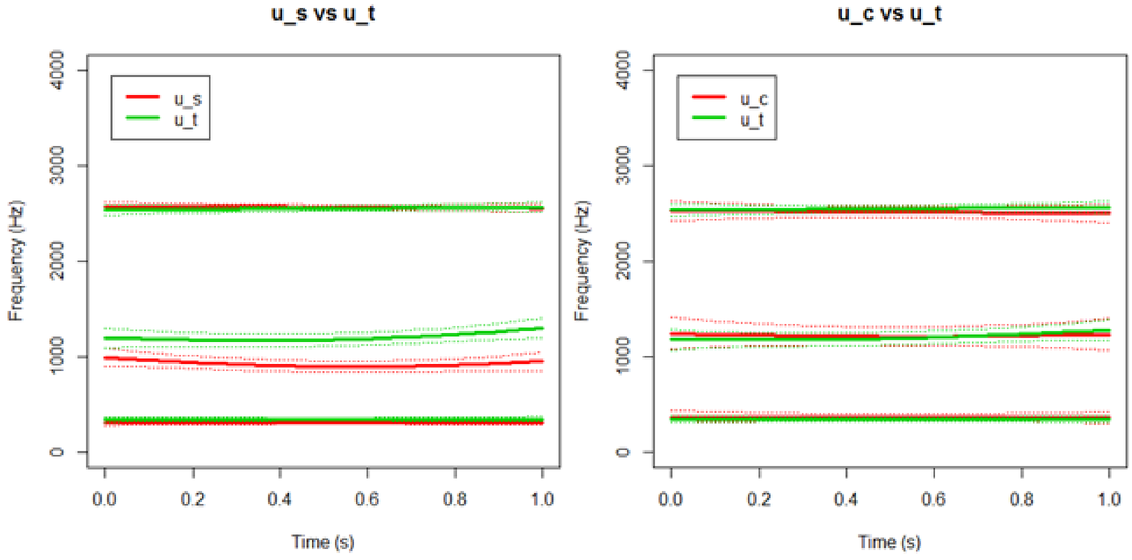
(a) (b)

Figure 5.27. Smoothing Splines for vowel /aa/ (a) Source vs Target & (b) Converted vs Target



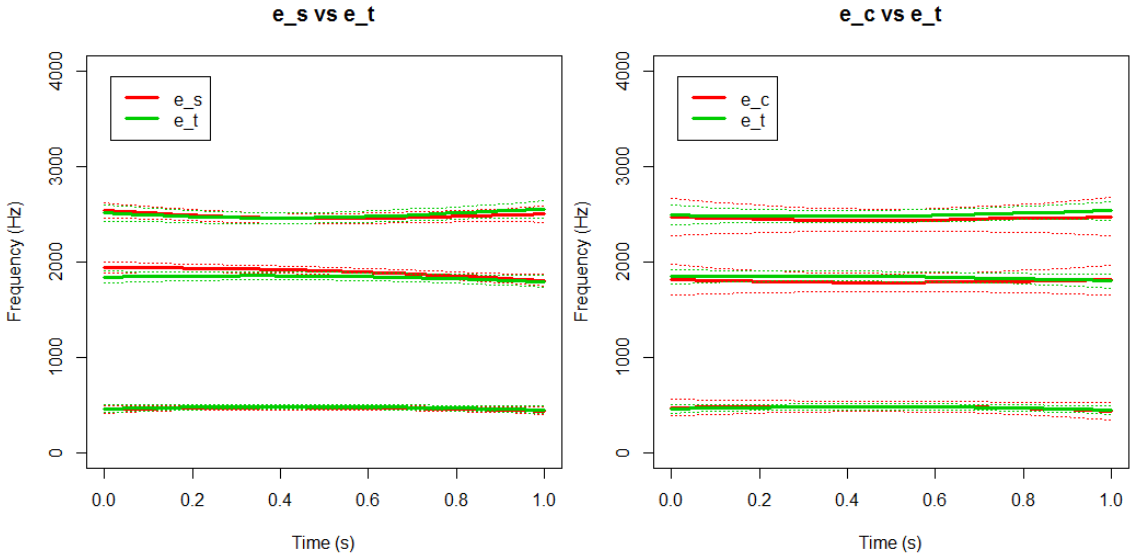
(a) (b)

Figure 5.28. Smoothing Splines for vowel /i/ (a) Source vs Target & (b) Converted vs Target



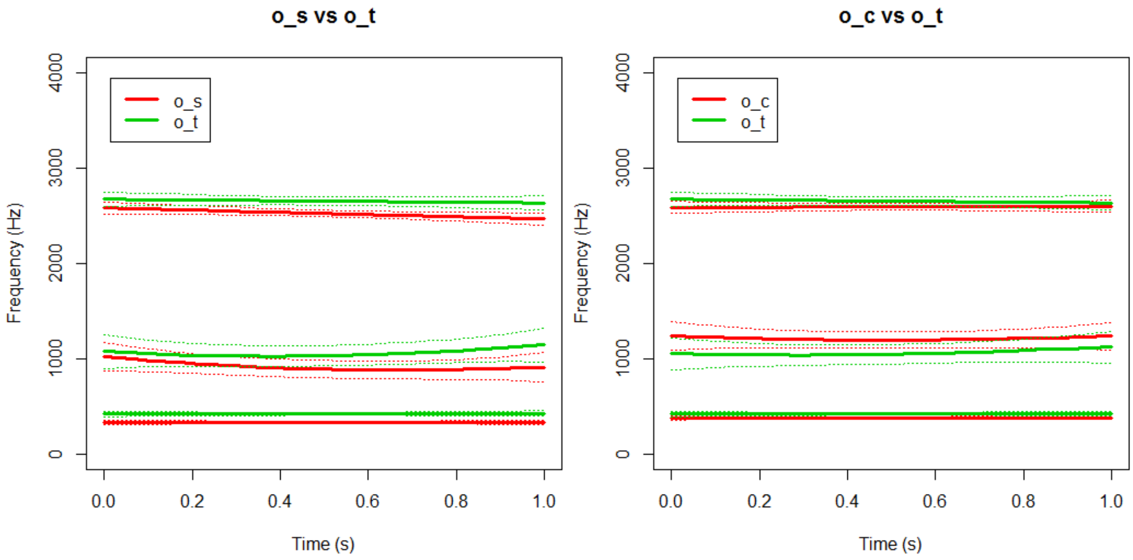
(a) (b)

Figure 5.29. Smoothing Splines for vowel /u/ (a) Source vs Target & (b) Converted vs Target



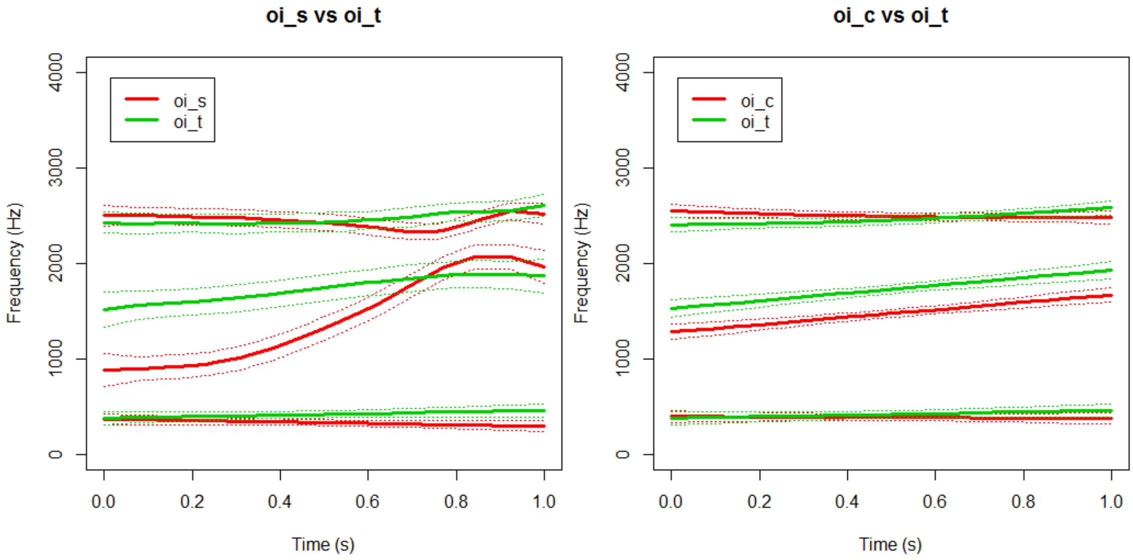
(a) (b)

Figure 5.30. Smoothing Splines for vowel /e/ (a) Source vs Target & (b) Converted vs Target



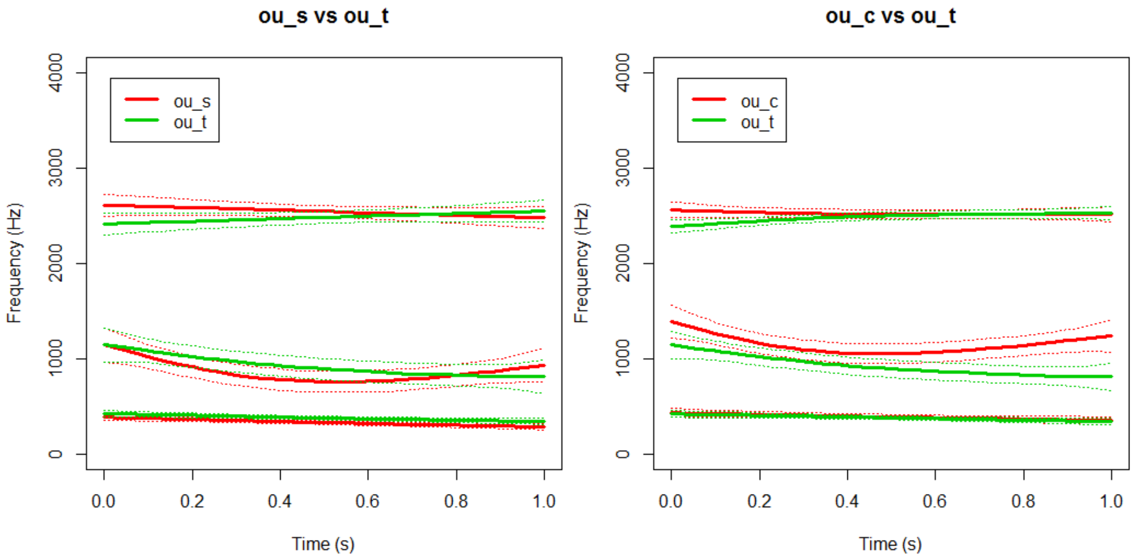
(a) (b)

Figure 5.31. Smoothing Splines for vowel /o/ (a) Source vs Target & (b) Converted vs Target



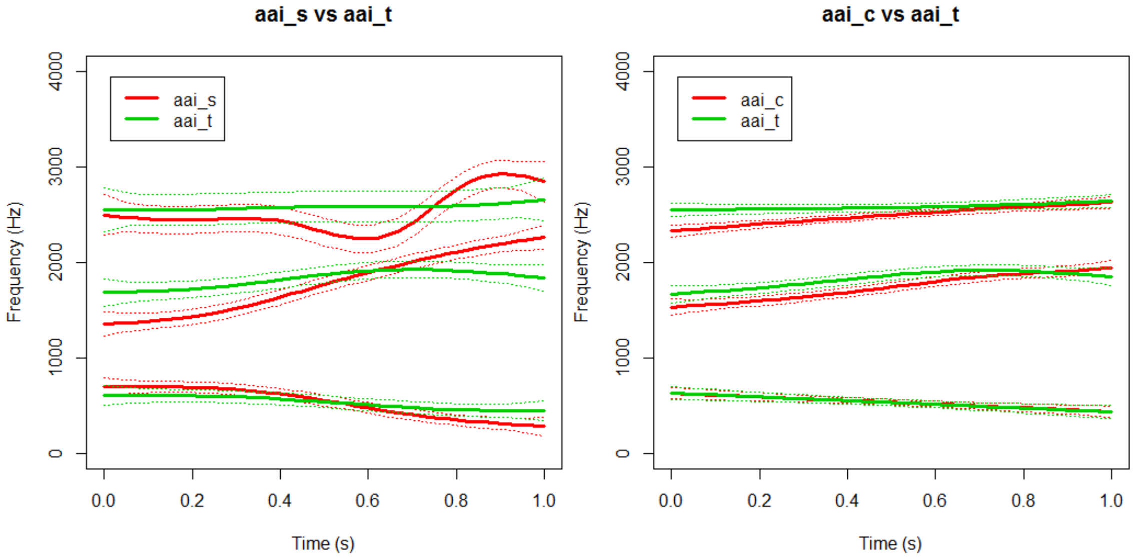
(a) (b)

Figure 5.32. Smoothing Splines for diphthong /oi/ (a) Source vs Target & (b) Converted vs Target



(a) (b)

Figure 5.33. Smoothing Splines for diphthong /ou/ (a) Source vs Target & (b) Converted vs Target



(a) (b)

Figure 5.34. Smoothing Splines for diphthong /aai/ (a) Source vs Target & (b) Converted vs Target

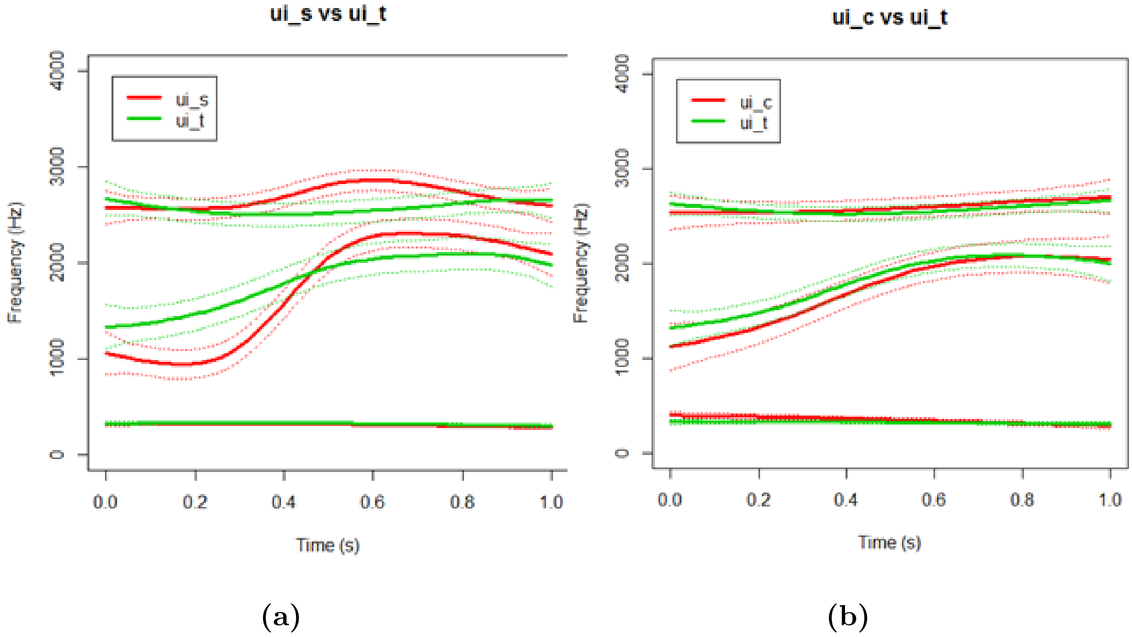


Figure 5.35. Smoothing Splines for diphthong /ui/ (a) Source vs Target & (b) Converted vs Target

5.5.4 Graphical Comparison of formant transformation results obtained using VC and FT

In Chapter4 we have discussed how the technique of VC has been successful to an extent in incorporating naturalness to synthesised dialectal speech. The VC method resulted in the transformation of spectral features (MCEPs) which is further equivalent to the transformation of formant frequencies. In this chapter we have attempted to transform the formant frequencies of vowels/diphthongs using GMM based mapping functions. We now compare graphically the results obtained using VC and FT. We have taken three test utterances together with their respective target utterances, common to both the VC based transformation and the FT based transformation. F1, F2 formant frequencies of the vowels/diphthongs in the target utterances, in the converted utterances using VC and those converted using FT are plotted for comparison in Figures 5.36, 5.37, 5.38, 5.39, 5.40 and 5.41. It is observed that compared to the VC based transformation, the FT based transformation has resulted in better (closer to target formant contours) and smoother formant contours of vowels/diphthongs.

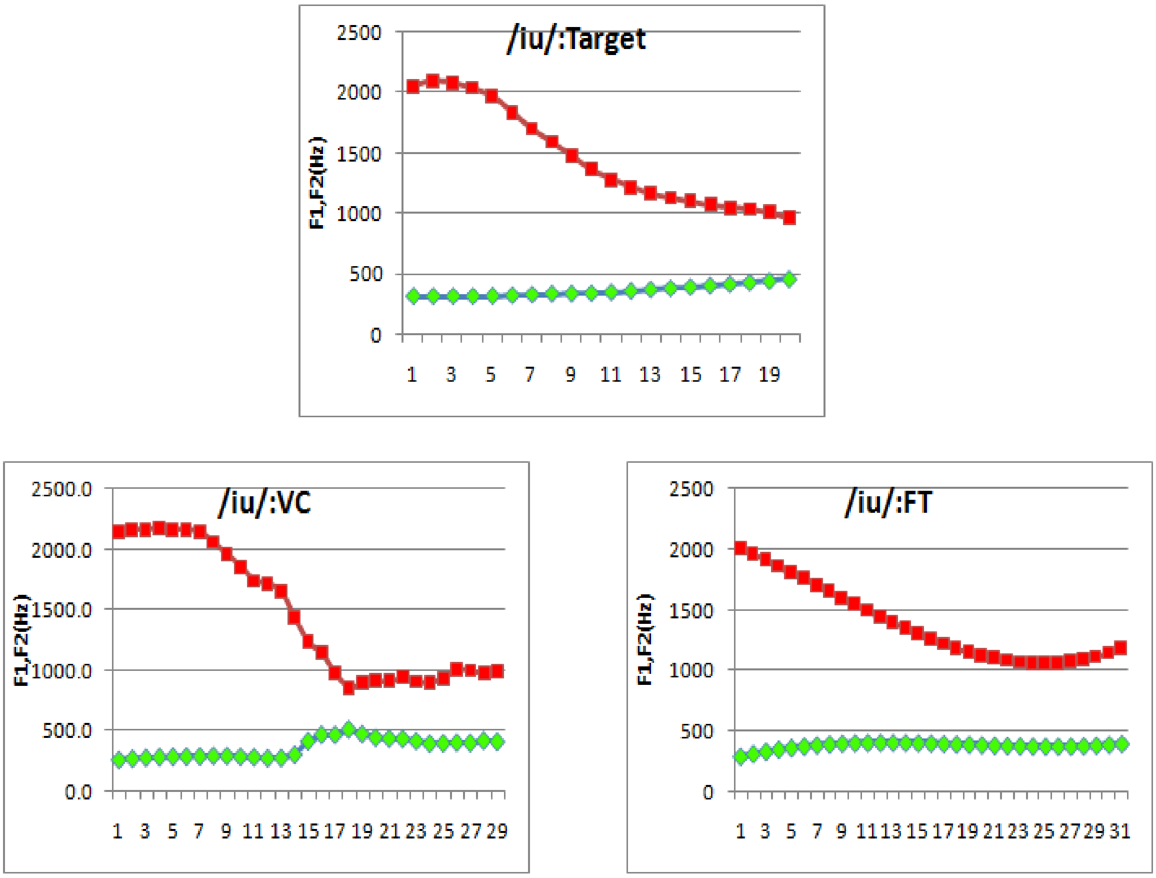


Figure 5.36. Diphthong /iu/ (a) Target, (b) Converted(VC) & (c) Converted(FT)

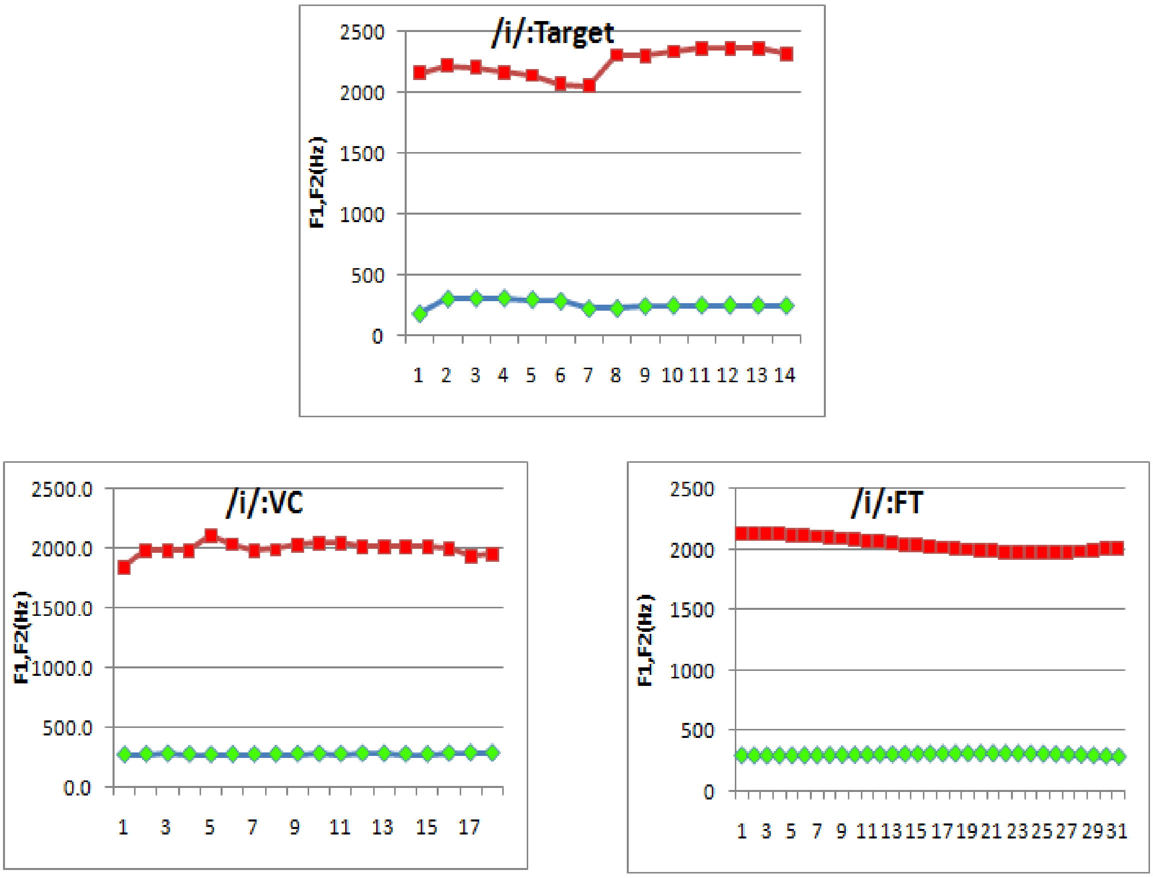


Figure 5.37. Vowel /i/ (a) Target, (b) Converted(VC) & (c) Converted(FT)

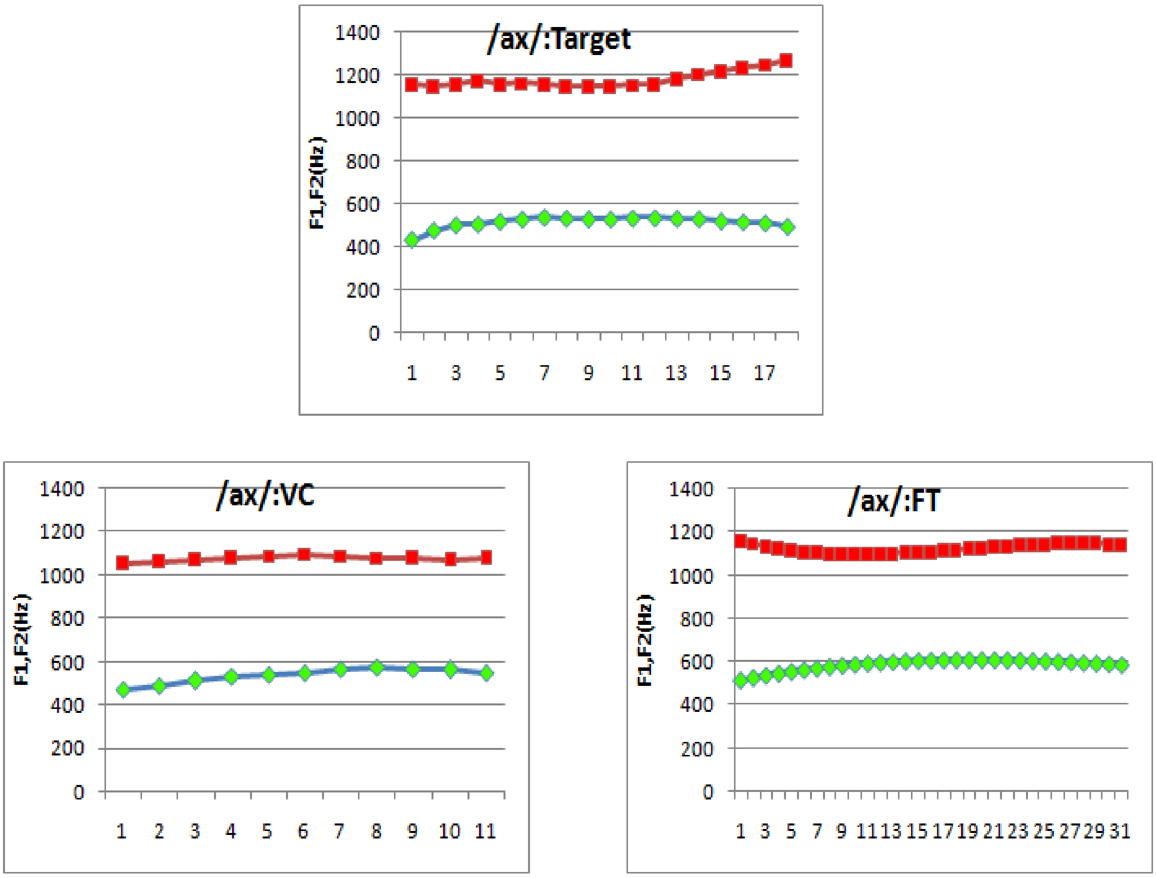


Figure 5.38. Vowel /ax/ (a) Target, (b) Converted(VC) & (c) Converted(FT)

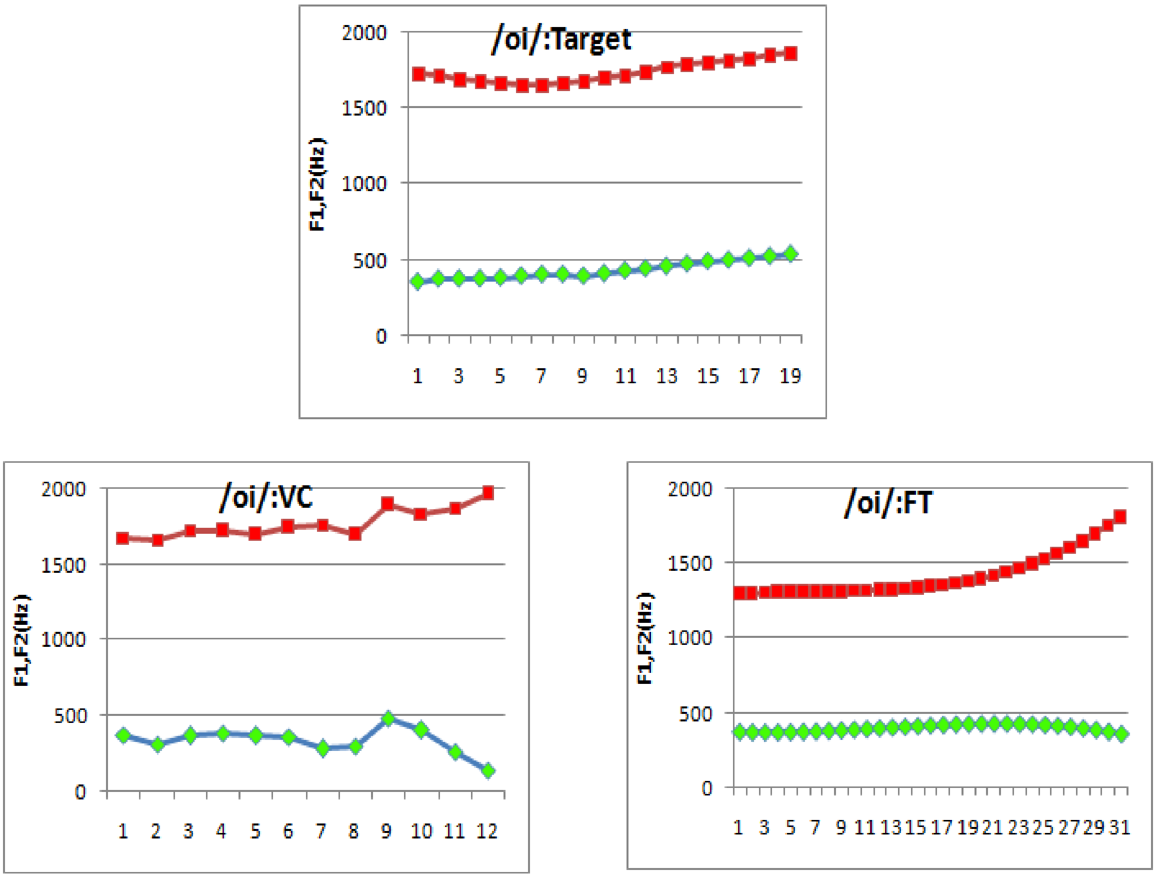


Figure 5.39. Diphthong /oi/ (a) Target, (b) Converted(VC) & (c) Converted(FT)

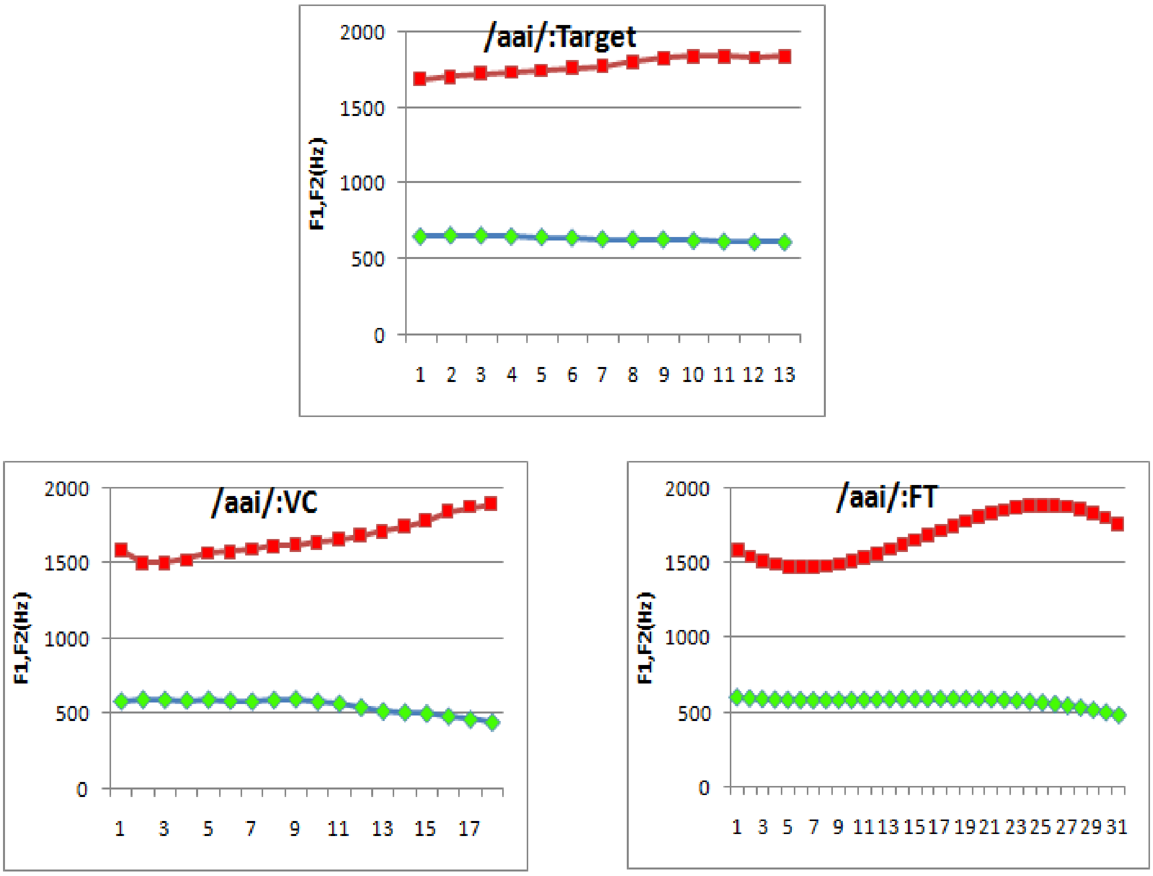


Figure 5.40. Diphthong /aai/ (a) Target, (b) Converted(VC) & (c) Converted(FT)

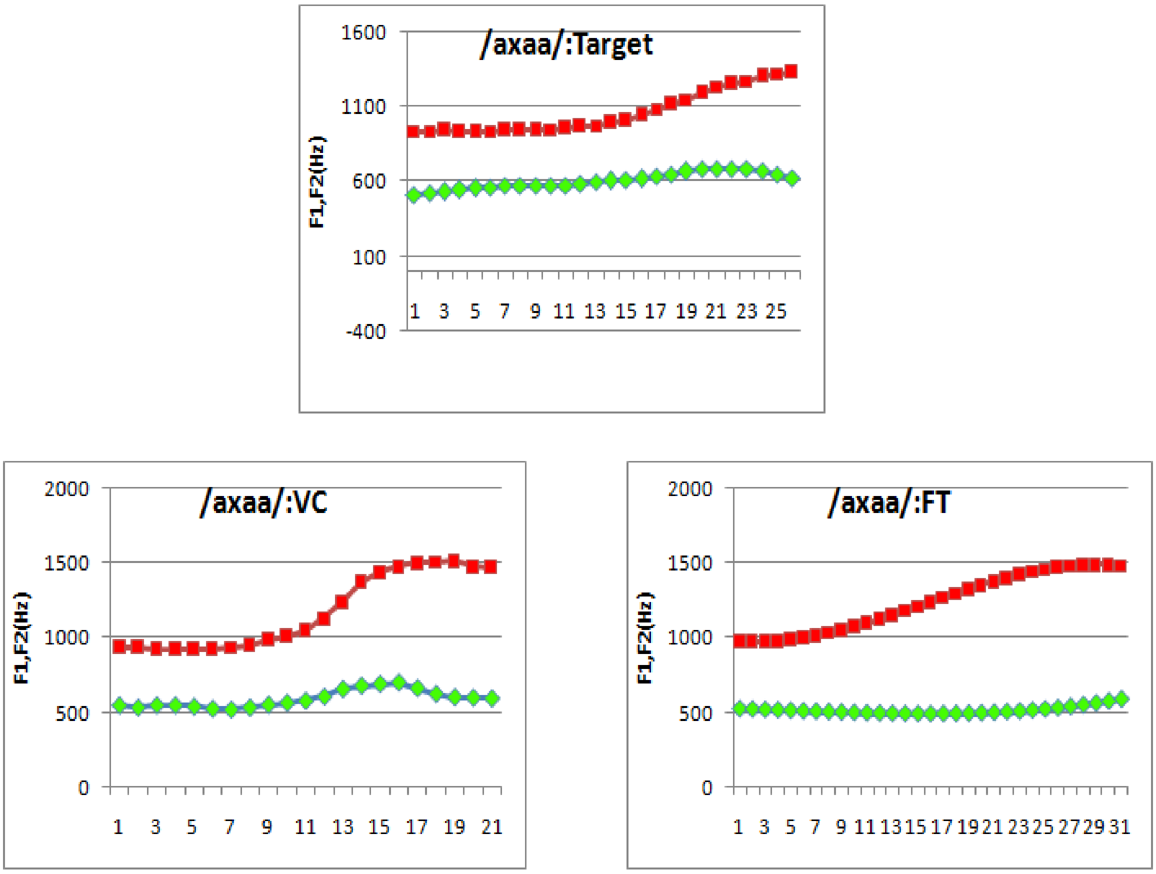


Figure 5.41. Diphthong /axaa/ (a) Target, (b) Converted(VC) & (c) Converted(FT)

5.5.5 Subjective Evaluation

The Klattworks [14] interface to the 1988 Klatt synthesiser is used to generate speech stimuli using the transformed formants. Although Klatt is not state-of-art, it was the best possible way we could resynthesize vowel segments using transformed formants only. A human evaluator is presented with three sets of stimuli, ‘A’, ‘B’ and ‘X’, and asked to find out which one out of ’A’ and ’B’ is closer to ’X’. Set ‘A’ consists of a set of ten test words, i.e., the words generated by the standard TTS, TU-TTS. Vowels/diphthongs in words in set ‘A’ are annotated in PRAAT, formants F1, F2 and F3, are extracted in Klattworks and written to a text file. The formants F1, F2 and F3, in this text file, are converted using the GMM based transformation method which is implemented in Matlab. Klattworks then generates new stimuli with the transformed formants and this forms Set ‘B’. Set ‘X’ consists of target words, i.e., the corresponding words in the NAL variety. However since we are going to compare synthesised stimuli to Set ‘X’, therefore instead of using the target words recorded directly from the speaker, we have used the Klattworks [14] interface to the 1988 Klatt synthesiser, to read the formants from the target words and resynthesise them so that each of ‘A’, ‘B’ and ‘X’ contain synthesised stimuli. Results of this test indicate that the stimuli generated using transformed formants are perceptually closer to the respective words in the target dialect.

5.6 Summary

Objective evaluation results indicate that in most of the transformations, with any of the three approaches, the source formants are brought closer to the target formants. In most cases the third approach, where a separate mapping function is built for each of the vowels/diphthongs, outperforms the other two. The disadvantage of this approach is the additional requirement of label files (containing vowel/diphthong identity and duration) for each of the utterances. Results of the comparison with an existing ANN-based method of FT show that our GMM-based method performs better specially for the diphthongs. Compared to the source vowel space, the transformed vowel space is observed to be closer to the target vowel space. The distance between like vowels in the source and target vowel formant spaces after the transformation is also reduced except for the vowel /o/. It is expected that a careful transcription of /o/ and /u/ vowels in the training data would improve transformation results.

In a bid to study the effects of training data size on the efficiency of the mapping functions developed, the mapping function developed using Approach 1 described in Section5.4.2.3, is trained with varying amounts of training data. The function is trained four times, with first 25%, first 50%, first 75% and 100% of the total speech corpus. Accordingly the functions are tested on the same set of 10 test utterances and the RMSE values of the vowel formants before and after transformations are recorded. In general it is observed that the efficiency of the mapping function increases with the increase in data size. However this is not always the case. A probable reason is the process of selecting the training data. When we increase the data size from first 25% to first 50% of the total corpus, it does not necessarily mean that the frequency of occurrence of a vowel/diphthong also has been increased by the same amount. A better method therefore, would be to increase the training data in a way that would result in the increase of vowel/diphthong samples also, and this can be taken up as a future task.