Extended Study of Pitch Shifted Speech by Preserving Tempo: An Experimental Study

S Choudhury, C Singh, M Thakar

Citation

S Choudhury, C Singh, M Thakar. *Extended Study of Pitch Shifted Speech by Preserving Tempo: An Experimental Study*. The Internet Journal of Forensic Science. 2006 Volume 2 Number 1.

Abstract

The overall pitch of a recorded speech sample could be subjected to pitch shift techniques available with the advancement in digital technology. Effect on speech characteristics due to time domain pitch shift technique have been undertaken using time warping. Study on the effect of frequency domain pitch shift by preserving tempo has been conducted with the speech exemplars of 15 speakers at a stretch ratio of 90, 95, 105 and 110 as compared to the original speech exemplar. Effect due to frequency domain pitch shift on F1, F2, F3, nasal formant frequencies, duration of word segment and mean period are analyzed with respect to the overall shift in the mean F0. The change in pitch due to stretching is found independent of the position of F1, F2 and F3. However, the change in the values of F1, F2, F3 and mean period for a speaker is linear.

Note: The paper was presented at XVI All India Forensic Science Conference 2004, Hyderabad, India and appeared in the Proceedings.

INTRODUCTION

A change in overall pitch results in a change in the speech characteristics, which makes the forensic expert a challenging task in the process of identifying the speaker [1,2,3,4,5]. Automatic systems for speaker identification based on pitch detection technique suffer from similar problem [6,7,8]. The shift in pitch may be circumstantial or intentional. Recording of speech in a low-grade recorder, recording with off-speed due to low battery or power supply, malfunction of the tape recorder etc. lead to pitch change. Secondly, the difference between standards used for film and for video generates problems when converting from one format to another. Since all the images are displayed, the change of frame rate induces a pitch change on the sound. Another suitable example may be considered as to fit a specified duration of a video footage or speech to a fixed length of time. These are all circumstantial. Effect of change in the playback speed of an analog recorder in authenticity examination has been discussed [9]. In certain situations, factor like tape stretch can also contribute to pitch shift and timing errors, which are significant in contrast to the NAB & DIN specifications as described by McKnight [10]. Advances in technology and processing of audio data digitally by applying different signal processing techniques have contributed a wide number of tools to shape audio data. It has become possible to alter data in a desired manner with the advent of computer-based tools. The methods used are either time domain or frequency domain or time-frequency domain. Time domain uses autocorrelation technique while frequency domain uses phase-vocoder technique based on the concept of analysis, transformation and/ or synthesis applied to the original sound. Time-frequency domain is based on constant bandwidth and modification of phase. The study on the effect of time warping on speech characteristics has been carried out [11] and its impact on speaker identification has been discussed. An extended study has been conducted considering the speech characteristics due to frequency domain pitch shift technique by preserving tempo.

METHODOLOGY & EXPERIMENTATION

SELECTION OF SPEECH MATERIAL

Text containing vowels and nasals are prepared in Hindi. A total of 15 speakers, both male and female in the age group of 25-45 are selected and asked to read the text. Two utterances of each speaker are recorded in a semiprofessional type analog tape recorder. These samples are digitized at a sampling rate of 22050 using 16-bit quantization in mono mode. The sentence of interest “Das din tak banirahi” is chosen from the whole text and it was segregated either from the first or second utterance, whichever is clearly spoken from each of the speaker.

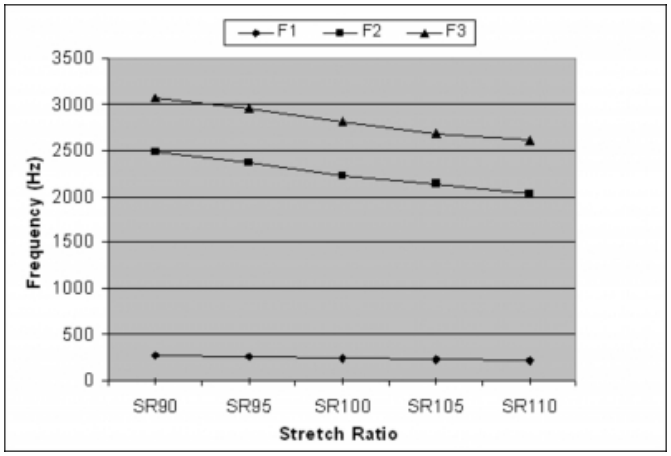
Exemplars are prepared by subjecting these samples to a constant stretch ratio of 90, 95, 105 and 110 by preserving tempo. Splicing frequency of 50 Hz and overlapping of 30% is used for stretch ratio of 90, splicing frequency of 49 Hz and overlapping of 29% is used for stretch ratio of 95, splicing frequency of 47 Hz and overlapping of 28% is used for both 105 and 110 stretch ratio. These exemplars are analyzed in Computerized Speech Laboratory (4003B). Mean fundamental frequency (F0); first (F1), second (F2) and third formant (F3) frequencies at a particular location (/d□s/, /b□ni/), duration of word-segment (/din/) & number of periods and nasal formant frequencies (/din/) are measured. The word /d□s/ and /b□ni/ are chosen to study the vowel characteristics with fricative and nasals.

RESULTS AND DISCUSSION

Fig.-1 shows the first formant frequency (F1), second formant frequency (F2), third formant frequency (F3) at /d□s/ for the speaker (S7) having minimum value of mean F0.

Figure 1

Figure 1: Formant frequencies at for the speaker (S7) having minimum mean F0

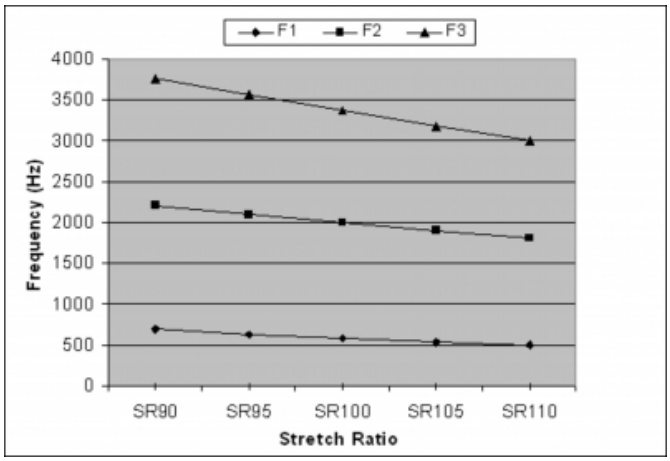


Variation of F2 and F3 is more than twice from the variation of F1 on changing pitch from stretch ratio of 90% through 110%. Stretching an exemplar with a ratio of 90 or 95 either add periods or reduce the duration of each period in the syllable of a word by using a complex algorithm to increase the overall pitch. The extra periods added to the existing periods as appear from the waveform are the mean of the previous and the following period at the center of the syllable. Similarly, stretch ratios of 105 or 110 either remove periods or elongate the existing the periods of the syllable and thereby lowering the overall pitch. The removal of periods cause a loss in formant information and a shift in the formant is observed. Addition or deletion of periods in the syllable results in a decrease or increase in the silence region respectively, even if the total duration of the exemplar is constant. The introduction or removal of periods takes place in such a way that the mean period decreases linearly for stretching below 100 and increases for stretch ratio higher than 100. In case of time warping, pitch changes by elongating or compressing the whole sample in time but the number of periods in the syllable remains unchanged.

The variation of F1, F2 and F3 at /b□ni/ for the speaker (S9) having maximum value of mean F0 is shown in Fig.-2. Like other speakers, the variation in F1 is found to be lesser than the variation in F2 and F3 for the word /b□ni/. The change in the value of F1, F2 and F3 due to stretching is found to be linear for all the speakers.

Figure 2

Figure 2: Formant frequencies at for the speaker (S9) having maximum mean F0



The change in the formant frequency is equally effective in other regions also. No such noticeable difference is observed in the fricative region /s/ in the wideband spectrogram.

Nasal formant frequencies measured at /din/ as shown in Table-1 is found to vary in a similar way as it varied at /d□s/ or /b□ni/ for the corresponding speaker. The variation of N2 is more prominent than N1, which indicates that the higher formant frequencies are more affected when a change of pitch is carried out by preserving tempo.

Figure 3

Table 1

Speaker	Stretch Ratio	F1 (Hz) /bAni/	F2 (Hz) /bAni/	F3 (Hz) /bAni/	Mean F0 (Hz)	Duration /din/ (ms)	No. of Periods	Nasal N1 /din/ (Hz)	Nasal N2 /din/ (Hz)
S1	SR90	452	2600	3171	149.78	87.82	13	346	3060
	SR95	430	2413	2980	141.5	77.91	11	319	2940
	SR100	390	2270	2801	134.99	75.01	10	293	2790
	SR105	340	2123	2600	127.4	62.85	8	265	2630
	SR110	310	1980	2430	122.16	57.59	7	239	2502
		/dAs/	/dAs/	/dAs/					
S2	SR90	664	1476	2662	136.21	83.78	12	385	2570
	SR95	620	1400	2600	128.65	82.08	11	365	2387
	SR100	561	1360	2505	120	78.01	10	348	2244
	SR105	520	1294	2400	113.99	65.64	8	327	2087
	SR110	449	1200	2280	108.97	60.25	7	308	1964
		/dAs/	/dAs/	/dAs/					
S3	SR90	554	1698	2757	138.23	61.38	8	267	2589
	SR95	504	1620	2638	132	64.87	8	243	2487
	SR100	430	1530	2521	124.91	51.12	6	221	2365
	SR105	379	1420	2400	120	37.05	4	200	2220
	SR110	300	1320	2270	114.02	36.75	4	180	2110
		/bAni/	/bAni/	/bAni/					
S4	SR90	520	2474	3047	208.77	81.44	18	253	3105
	SR95	469	2335	2916	196	71.87	15	244	2930
	SR100	410	2237	2798	186	65.92	13	224	2770
	SR105	360	2139	2661	178.89	65.02	12	197	2600
	SR110	305	2050	2505	171.67	62.75	11	181	2454
		/bAni/	/bAni/	/bAni/					
S5	SR90	621	1807	3235	152.1	74.08	10	309	3072
	SR95	591	1701	3110	147.37	62.35	8	297	2914
	SR100	500	1590	2980	143.8	65.49	8	286	2770
	SR105	450	1500	2850	141.32	69.04	8	278	2628
	SR110	385	1417	2736	139.42	54.26	6	270	2500
		/bAni/	/bAni/	/bAni/					
S6	SR90	537	2407	2722	150.23	99.34	15	180	2676
	SR95	500	2260	2627	144	96.59	14	173	2600
	SR100	440	2130	2522	137.25	95.87	13	170	2500
	SR105	350	2000	2390	130.38	86.41	11	154	2387
	SR110	250	1867	2253	124.66	82.2	10	148	2272
		/dAs/	/dAs/	/dAs/					
S7 (F)	SR90	700	2214	3755	241.3	52.11	16	298	3536
	SR95	635	2090	3562	236.72	62.43	18	281	3308
	SR100	580	1992	3379	233.2	61.84	17	261	3116
	SR105	539	1906	3177	230.14	65.65	17	250	xx
	SR110	500	1800	3000	227.5	51.31	13	241	xx
		/dAs/	/dAs/	/dAs/					
S8 (F)	SR90	800	1888	3255	210.1	87.76	25	303	2806
	SR95	718	1780	3107	208.26	84.67	23	279	2774
	SR100	660	1670	2970	206.93	66.0	17	260	2698
	SR105	620	1575	2820	205.89	45.3	11	245	2644
	SR110	587	1487	2684	204.96	47.5	11	231	2542
		/bAni/	/bAni/	/bAni/					
S9	SR90	278	2482	3067	128.4	67.57	10	296	3060
	SR95	258	2364	2958	121.9	64.3	9	279	2965
	SR100	242	2230	2811	114.7	59.8	8	265	2820
	SR105	230	2133	2689	108.9	39.1	5	256	2670
	SR110	210	2024	2611	102.42	41.4	5	245	2520
		/bAni/	/bAni/	/bAni/					
S10	SR90	404	2140	3033	143.12	58.38	8	293	2618
	SR95	382	2058	2950	138.28	52.99	7	275	2500
	SR100	357	1980	2868	132.9	48.05	6	250	2390
	SR105	339	1899	2762	128.2	42.66	5	228	2250
	SR110	300	1810	2640	123.9	34.79	4	213	2145
		/bAni/	/bAni/	/bAni/					
S11	SR90	588	1520	2925	139.97	60.99	9	251	2666
	SR95	523	1415	2800	137.53	64.53	9	244	2598
	SR100	450	1330	2660	135.65	59.96	8	231	2530
	SR105	395	1225	2521	134.29	47.51	6	214	2470
	SR110	359	1145	2405	133.21	41.39	5	197	2387
		/dAs/	/dAs/	/dAs/					
S12 (F)	SR90	570	1858	3779	185.63	79.1	18	182	3140
	SR95	534	1782	3600	182.2	69.08	15	176	3060
	SR100	460	1670	3400	179.48	67.24	14	169	2970
	SR105	427	1580	3220	176.55	60.17	12	161	2840
	SR110	370	1489	3060	173.81	53.41	10	150	2700
		/dAs/	/dAs/	/dAs/					
S13	SR90	590	1600	3085	144.96	62.33	9	386	2811
	SR95	584	1560	2987	138.64	73.29	10	340	2700
	SR100	545	1484	2850	132	69.29	9	304	2540
	SR105	500	1410	2701	125.75	57.28	7	277	2380
	SR110	467	1310	2559	120.65	60.05	7	258	2197
		/bAni/	/bAni/	/bAni/					
S14	SR90	446	2246	3031	139.19	40.69	6	313	2564
	SR95	390	2170	2900	133.5	35.72	5	305	2491
	SR100	325	2075	2790	127.6	30.08	4	287	2350
	SR105	280	2005	2690	122.46	31.26	4	269	2181
	SR110	248	1930	2574	117.25	24.79	3	250	2037
		/dAs/	/dAs/	/dAs/					
S15	SR90	482	1694	2725	146.25	82.69	14	270	2816
	SR95	460	1610	2610	139.03	81.17	13	253	2700
	SR100	430	1523	2478	131.77	65.81	10	243	2555
	SR105	402	1423	2348	125.2	48.09	7	236	2421
	SR110	382	1335	2210	117.8	51.52	7	227	2264

SR100 represents the original sample; F – Female, xx – Measurements could not be taken

Fig.-3 (a) shows the percent variation of F1, F2 and F3 with respect to mean F0 at /bni/ and Fig.-3 (b) shows variation of F1, F2, F3 at /dis/ for stretch ratio of 110. This indicates that the percentage of decrease of F1, F2 and F3 is not same for all the speakers.

Figure 4

Figure 3 (a): Percent variation of F1, F2 & F3 with respect to Mean F0 at for stretch ratio of 110

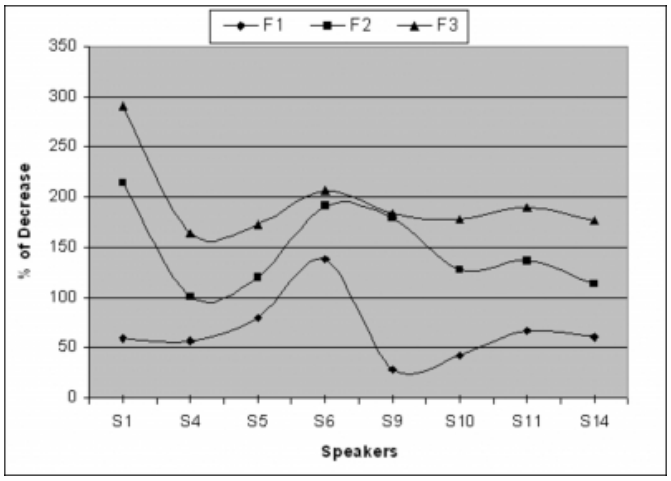


Figure 5

Figure 3 (b): Percent variation of F1, F2 & F3 with respect to Mean F0 at for stretch ratio of 110

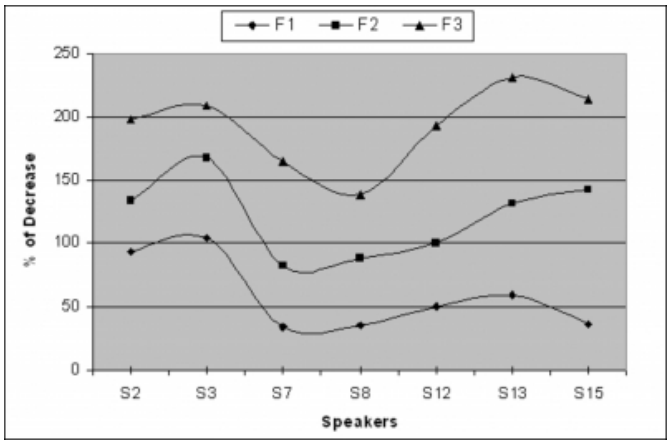


Fig.-4 (a) shows the percent variation of F1, F2 & F3 with Mean F0 for stretch ratio of 105 at /bni/ for the speakers S9, S14, S10, S11, S6, S5, S4 respectively. Percent variation of F1, F2 & F3 with Mean F0 for stretch ratio of 105 at /dis/ for the speakers S2, S3, S15, S13, S12, S8, S7 respectively is shown in Fig.- 4(b). These two plots indicate that the percent variation in the values of F1, F2 & F3 is independent of their initial values in the original exemplar.

Figure 6

Figure 4 (a): Percent variation of F1, F2 & F3 with Mean F0 for stretch ratio of 105 at for the speakers S9, S14, S10, S11, S6, S5, S4 respectively

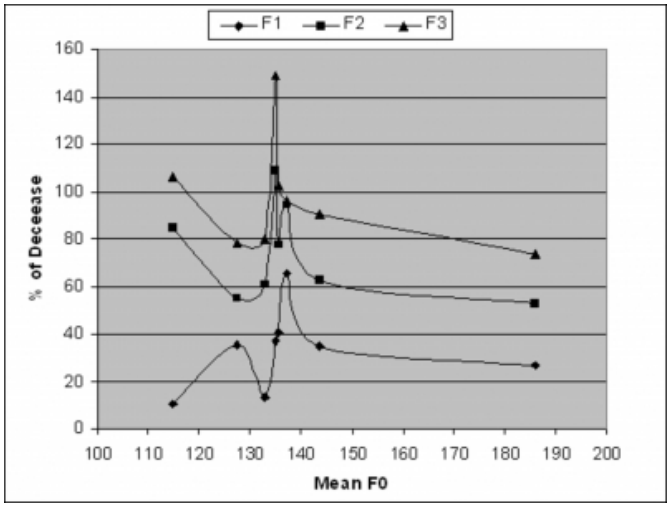
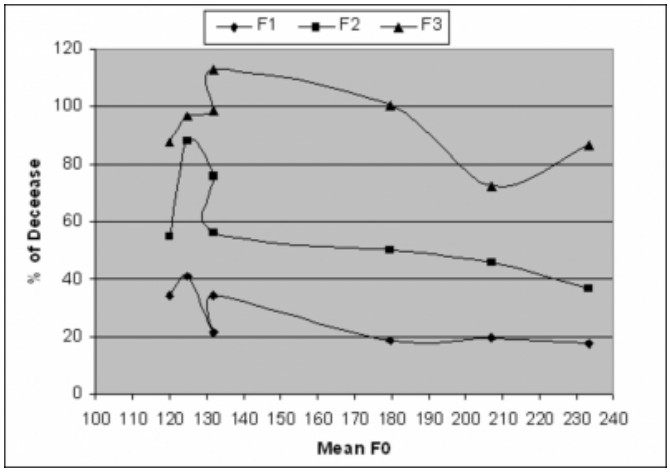


Figure 7

Figure 4 (b): Percent variation of F1, F2 & F3 with Mean F0 for stretch ratio of 105 at for the speakers S2, S3, S15, S13, S12, S8, S7 respectively



CONCLUSION

The change of overall pitch by preserving tempo affects the higher formant frequencies more than the lower formants with linear change in the measurable speech parameters. The amount of change in the values of F1, F2 & F3 is found to be different for each speaker. The attempt to bring back the changed speech samples to the original by reversing the change in the formant frequencies could be brought near to the original. However, information contained in the removed period of the speech sample is lost while moving from higher to lower pitch. Still the sample could suitably be used for speaker identification purposes, as the characteristics

pertaining to speaker dependent feature parameters are found preserved in the process.

References

1. Mead KO. Identification of speakers form fundamental frequency contours in conversational speech. Joint Speech Research Unit1974; Report No. 1002.
2. Stevens SS and Volkman J. The relation of pitch to frequency: A revised scale. American Journal of Psychology 1940; 53: 329-353.
3. Jassem W. Pitch and compass of the speaking voice. Journal of the International Phonetic Association 1971; 1: 59-68.
4. Steffen-Batog MW, Jassem and Gruszka-Koscielak H. Statistical distributions of short term F0 values as a personal voice characteristic. In: W. Jassem (ed.) Speech analysis and synthesis, Warsaw: Police Academy of Science; 1970: Vol.2.
5. Horii Y. Some statistical characteristics of voice fundamental frequency. Journal of Speech Hearing Research 1975; 18 (1): 192-201.
6. Atal BS. Automatic speaker recognition based on pitch contours. Journal of Acoustic Society of America 1972; 52: 1687-1697.
7. Green N. Automatic speaker recognition using pitch measurements in conversational speech. Joint Speech Research Unit 1972; Report No.1000.
8. Robert CL. Speaker verification by computer using speech intensity for temporal registration. IEEE Transaction on audio and Electro Acoustics. 1973; AU-21 (2): 80-89.
9. Koenig BE. Measurement of recorder speed changes in authenticity examinations. Crime Laboratory Digest 1987; 14 (4): 140-152.
10. McKnight JG. Speed, pitch and timing errors in tape recording and reproducing. Journal of Audio Engineering Society 1968; 16: 266-274.
11. Singh CP, Manisha K and Choudhury SK. Study of speech characteristics due to pitch shift by time warping method and its' impact on forensic speaker identification. Proceedings of the XV All India Forensic Science Conference 2004: 56.

Author Information

S. K. Choudhury
Central Forensic Science Laboratory

C. P. Singh
Central Forensic Science Laboratory

M. K. Thakar
Department of Forensic Science, Punjabi University