Ahmed Abdul Rehman - Salman Bashir Awan

Speaker Dependent Features in Transition from a Stop to a Vowel

Abstract: *This paper inspects whether there is some speaker dependent information present in formant transition rates from consonant into a vowel, also if it is computationally possible to extract this information. The first four formants were studied in this regard.*

Keywords: *Consonant, vowel, formant, voice onset time, fundamental frequency, spectrogram*

1. INTRODUCTION

As computers become more powerful and their computational power increased exponentially, new horizons were open to mankind. Now those cumbersome calculations which man never tried to do were within the human reach. One of such fields is Speech processing. The human speech is now analyzed for following reasons: Speaker identification, language identification and speech recognition. Our paper focuses on Speaker identification.

Start of speaker identification is attributed to Stalin by some people, but serious efforts started with its usage in forensic and security systems around 20th century in London and USA (Holien, 2000). Currently it is employed in fields like telephony, Internet security, and other security systems.

Speaker identification is a recognized name in market and media. Use of term voiceprint is as common in use as fingerprints. However it is a misleading concept at the moment because fingerprint speaks of invariant physical characteristic and voice is the product of two mechanisms, which exhibit considerable flexibility i.e. the speech organs and language according to the current knowledge. Despite this crisis of theory, a lot of work is being conducted on Speaker Identification especially in the fields of forensic phonetics and telephony communications. Many research groups are working both independently and together to bring out better results in speaker recognition, of which speaker identification is a part (Holien and Koster, 1996).

2. LITERATURE REVIEW

Speech is continuous process, but it is made of different smaller pieces i.e. words and then phonemes. It is a mixture of vowels and consonants. Every vowel has its specific formants, but due to the presence of consonants at its either ends, these formants tend to change a bit at the ends. This changing of formants occurs when a person speaks one consonant from one articulatory place and then has to position his/her vocal tract to utter the proceeding vowel. This time, when he/she is done saying the consonant and has not yet positioned the vocal tract to say the stable part of the vowel, is very crucial in speech. It is said to be the part that makes speech intelligible, during this time the formants rise or fall, depending upon the speaker and context of that vowel, then they reach the stable part of that vowel (Clark and Yallop, 1992).

This paper is based on the fact that all human beings have a different composition of the vocal tract; so it is assumed that their way of changing the position to say a vowel after they have said a consonant would be different.

3. METHODOLOGY

3.1 Selection of consonants

Initially it was assumed that this study would be text independent since a speaker should make his/her print in every consonant-vowel transition, but a little detail led this analysis to be text dependent.

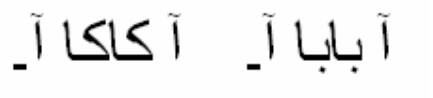
Consonants have different effects on the vowel formants. Some make them rise while going out of the consonant, bilabials, where as some simply distort them for a short time period, retroflex. Some consonants have relatively longer voice onset time, fricatives, while others have short, affricates. Moreover consonants can also be nasal. When these different consonants are spoken with different vowels they result in a large number of distinct consonant-vowel combinations with their own pair specific formant transitions (Pickett, 1999).

The speakers' vocal tracts further filter these transitions and add speaker dependent information in it. So it is computationally not possible to first identify the consonant vowel pair and then distinguish between the source transition and the filtering effect. Hence resulting in context dependent analysis of speech signal.

This limitation also guided the selection of consonants and vowels for this experiment. Out of the wide range of consonants available, stops were found to have very less effect upon vowels (they only have a burst, followed by a short voice onset time). By choosing stops, we expected to get clear and crisp transitions. Further, aspirated stops were discarded, to have even lesser voice onset time hence resulting in clearer transitions. The stops used for this experiment were /b/ and /k/, whereas the vowel used was /a/.

The words containing target consonant-vowel transitions were placed in a carrier sentence to minimize the context effect. Also the sentence was kept small to control the overall speech rate variation.

The sentence made was "a baba a, a kaka a".



3.2 Scope of the experiment

Speech signal not only vary from speaker to speaker, but also for a single speaker no two utterances of same sentence are identical. This Intra-speaker variability was a big issue for this study. In order to cover this variability, a sufficient number of recordings, ten recordings per speaker, were made. Due to this large number of readings per speaker we had to compromise on the number of speakers. Hence limiting the scope of this experiment to fifteen speakers.

3.3 Selection of speakers

To make this experiment independent of sex a good combination of male and female speakers was chosen. Out of fifteen subjects nine were male and six female speakers. All the speakers were adult of age ranging from twenty-two to twenty-four. All the subjects had normal height and they were native speakers of Urdu.

3.4 Experiment

All of the recordings were made in a recording room environment, which was not completely noise free. Further, the equipment also added some noise to the recordings. Hence the fundamental frequency was overlooked in the analysis, which was not distinguishable from noise in data.

All the readings were made in daytime the speakers spoke in a normal daily life manner, i.e. there were no extra-ordinary stress/strain.

3.5 Collection of data

There were four stop-vowel transitions in the sentence, two from /b/ to /a/ and two from /k/ to /a/. The software used for the collection of data was PRAAT. First four formants, F1 to F4, were analyzed for each transition. Fundamental frequency, F0, was ignored since there was noise in the recordings and also F0 was not visible in spectrograms shown by PRAAT.

The values calculated against each stop-vowel transitions, each formant, were:

- ? Difference between the formant value, when it becomes stable after the stop and when it completes its first transition.
- ? The time period in which this difference was taken.
- ? The rate of the transition. It was calculated by dividing the transition values by the time period.

These values were collected by inspecting spectrograms of the recordings. The values were calculated by selecting the region of first transition against each formant, and then calculating the values through PRAAT scripts.

The values against each speaker were stored in separate file. The average, of ten transition rates, against each formant was calculated. Then the standard deviation was also calculated for that average. During the calculation, if a value varied too much from the trend, it was discarded.

3.6 Need for normalization

One major issue, in this experiment, was of variable speech rates. Same speaker might have shown different speech rates during his/her recordings. So there must have been a normalization technique by which the effect of this variable vowel length, on formant transition rate, could be nullified. In order to find this normalization mechanism some additional readings were made. One of these results was expected.

1) *Some linear relationship exists:* The data was analyzed for some linear relationship between the vowel length and the formant transition rate. For this purpose different vowel lengths were measured for the same speaker along with the formant transition time. But data showed that there exists no linear relationship between these two time periods. As sometimes the vowel length was greater, relative to the length of same vowel of another utterance, but the formant transition time was less than the other's. This result could be seen from Table1. Rows labeled 3-4 and 2-4, in Table 1, suggests that as vowel length increases F1 transition time decreases, but row 1-4 denies this trend for F1.

Table 1 Values for vowel and formant lengths

Utterance-Vowel	Vowel length	F1time	F2 time	F3 time	F4 time
1-1	0.103673	0.039754	0.024610	0.010411	0.008992
1-2	0.147479	0.023663	0.025556	0.010885	0.031709
1-3	0.097662	0.033391	0.033868	0.009540	0.013356
1-4	0.126304	0.030529	0.036731	0.028621	0.017650
2-1	0.102294	0.037639	0.024947	0.017506	0.018382
2-2	0.133145	0.051644	0.023196	0.026697	0.009191
2-3	0.089384	0.033426	0.035848	0.025191	0.013079
2-4	0.114923	0.041177	0.041177	0.014048	0.011626
3-1	0.099259	0.034441	0.024874	0.019612	0.013872
3-2	0.146031	0.039224	0.027744	0.018177	0.010045
3-3	0.097999	0.039129	0.035010	0.009267	0.010812
3-4	0.117184	0.037070	0.043248	0.031921	0.009267

2) *No relation exists*: Since there was no linear relation found between the formant transition time and total vowel length so the rest of the experiment was based on conclusion that "If there exists some speaker dependent information in formant transition rates, it is independent of speech rate".

4. RESULTS

The figures obtained from the analysis of spectrograms yielded following results.

4.1 Inter-speaker variability

Closer look to data showed that not necessarily all the four formants gain identical transition rates and standard deviation values for a particular vowel. Table 2 shows this behavior. The data collected for the two occurrences of vowel /b/ for the same speaker gave an almost identical average value for F1 for both occurrence but at the same time F2 showed great variation even for the same vowel.

Table 2 F1 and F2 of two occurrences of /b/ for speaker 1

Readings	Rate of F1	Rate of F2	Rate of F1	Rate of F2
1	7072.776	-2496.01	12215.21	4455.941
2	14477.5	12765.81	11648.02	5313.179
3	15324.03	8854.203	12432.16	5600.245
4	9015.08	1745.976	15112.95	6141.452
5	14353.74	7382.663	8139.145	3566.008
6	13801.46	11837.22	13913.08	1580.609
7	13160.24	7652.866	15566.74	2501.92
8	13289.77	5793.605	14623.69	2028.638
9	12797.16	4691.543	14166.31	4990.859
10	13243.28	9078.147	15891.52	1047.126
Averages	12653.5	6730.602	13370.88	3722.598
STDVs	2583.709	4580.275	2339.575	1831.301

So all the four vowels were treated independently, since this effect was also considered as a result of position of that vowel in speech.

4.2 Intra-speaker variability

Besides inter-speaker variation there was a great deal of intra-speaker variation. Sometimes a formant value helped in distinguishing the two speakers, but it didn't help in distinguishing other two speakers. This result is shown in Table 3. Average F1 of speaker 1 and speaker 2 are quite apart and distinguishable, but this average is not helpful in differentiating speaker 1 and speaker 3, that are distinguishable on the basis of average F2.

Table 3 F1, F2, F3 and F4 for first /b/ vowel

Speakers	Average F1	STDV F1	Average F2	STDV F2
1	8241.313728	1240.89449	-277.593918	2808.25781
2	3243.492323	366.934269	-1107.852701	818.211650
3	9473.385069	2685.93256	3616.652374	2618.10226
4	2984.187971	991.80194	2817.807874	2576.19777
5	7990.463161	1452.12904	2675.072846	1190.16975
6	3714.931512	1551.44119	3731.671083	2470.42575
	Average F3	STDV of F3	Average F4	STDV of F4
1	3175.321366	3113.22465	5905.585305	8067.84972
2	-2371.146126	818.21165	-1521.011591	3516.75885
3	2213.335309	2606.38403	2340.657773	3394.36823
4	1288.685838	1700.75752	803.487011	3671.7745
5	5170.667652	2339.50186	2625.823139	5470.82492
6	1108.762621	2002.9578	-1312.109828	3773.0669

This behavior of data pointed towards the possibility of making a vector that used all the four formants at the same time to distinguish speakers.

4.3 Analyzing all four formants

Sets of all the four formants of a single speaker were compared with other speaker's set to find if they collectively make two speakers identifiable. Table 4 shows the data used to compare two speakers. We can see that if all the formants are collectively seen, while keeping in view their standard deviations we can clearly identify speaker 1 from all other speakers on the basis of F1 alone, for speaker 4, F1 and F2 collectively make it unique in the set.

Table 4 Four formants transitions against /k/

Speakers	Average F1	STDV of F1	Average F2	STDV of F2
1	6638.575801	848.762129	443.234089	3219.04266
2	3487.90591	361.956644	-2285.83948	894.318756
3	4515.63445	1081.69792	-5898.16528	1700.35565
4	3535.3926	1032.6829	-3790.23502	552.117235
5	5010.70953	863.173519	-1487.78133	1491.71362
6	4470.87576	1041.20555	-4228.47837	3614.91331
	Average F3	STDV of F3	Average F4	STDV of F4
1	3875.82063	3724.6492	2494.92592	2362.92933
2	-630.178361	4752.81464	-1825.43595	4438.54912
3	3714.88606	7033.38088	1596.91622	2366.82885
4	369.194375	2430.66096	-2087.04603	3395.14251
5	4401.43663	3345.68136	-2105.00174	7383.07431
6	14540.9715	13550.0807	3399.74927	13689.1256

The formant transitions for first four formants, against one vowel, were not enough to identify all the speakers uniquely.

4.4 Different vowels help to identify different speakers

One last major observation made in this experiment was of vowel dependent identification. i.e. formant transitions of one vowel, collectively, helped to distinguish between one set of speakers, where as formant transitions of another vowel helped to identify exceptionally another set of speakers. We saw that in Table 4 speakers 1 and 4 were very obviously distinguishable from the rest of the data. But in Table 5 speaker 5 is distinguishable from rest of the data on the bases of F1 and F2. This difference is on the basis of average values of F1 and F2, and also by the ranges formed by the standard deviation. If we see the average values of F1 and F2 for speaker 5 and make ranges by considering the standard deviation, we can see that no other speaker falls in these ranges.

Table 5 Four formants transitions against second /b/

Speakers	Average F1	STDV of F1	Average F2	STDV of F2
1	3199.7953	427.03162	-638.464	1187.374
2	7919.9411	1272.632	2229.8598	1304.456
3	4291.0499	2457.9701	2283.015	2392.062

4	6287.0327	1709.9171	782.9611	2922.805
5	3933.6746	882.24849	1317.9762	992.1726
6	7195.0422	1103.6159	-1230.15	4446.129
	Average F3	STDV of F3	Average F4	STDV of F4
1	-580.1168	2791.8666	-2861.885	4320.234
2	4451.055	3770.9661	3785.3972	8345.996
3	-1755.02	7219.0653	-1191.622	5955.495
4	3746.1834	1953.8853	1724.4222	1713.09
5	935.02193	1505.3269	-1890.623	6573.591
6	3684.7234	2243.307	4052.6448	2912.481

The result of such analysis on all the four vowels, provided 80% result, i.e. 12 out of 15 speakers were identified uniquely, according to the average formant values and standard deviation analysis explained above.

5. CONCLUSION

The results gathered from the experiment somewhat agreed with our supposition that everyone has a unique way of going into a vowel from a consonant. But not all the formants of one person behave that way. Sometimes, one formant gives healthy information while other formants do not. Sometimes one particular stop-vowel set yields unique-ness, and other stop-vowel sets do not. This gives us the impression that there is some speaker dependent information in these transition rates. It is quite possible that a person may not have uniqueness for one consonant-vowel set, but he/she maybe has this uniqueness for some other set. This is very natural as some people do have unique ways of saying some words or letters. If we build a matrix that contains all (or most of the) consonant-vowel combinations, and calculate values against them, we might get a unique matrix for one person. But that would require careful experimentation and formulation of statistical methods to analyze data.

REFERENCES

- Clark, John. and Yallop, Colin. *An Introduction to Phonetics and Phonology.* Blackwell Publishers, Oxford, UK, 1992.
- Holien and Koster, Jens-Peter, Speaker Identification by Aural-Perceptual Approaches, BEIPHOL 46(1996): Miszellen IX, pp 135-152.
- Holien, H. Phonetics by Encyclopedia of Forensic Sciences (J. Sigle, Ed) London Academic Press, 2000.
- Pickett, J.M. *The Acoustics of Speech Communication.* A ViaCom Company, USA, 1999.
- Kent, Ray D. and Read, Charles. The Acoustic Analysis Of speech, 1992.