# SPEAKER DEPENDENT FEATURES IN APPROXIMANTS OF URDU

***MUSTAFA MUBASHIR RIZVI***

## ABSTRACT

This paper suggests speaker dependent primary and secondary acoustic cues in approximants focusing the Urdu language. Experiment was conducted on seven speakers' samples. The data was collected and analyzed to identify speakers. The paper also suggests various cues, which can possibly identify speakers. It also sheds some light on the problems faced during the experiments and the after effects of increase in speakers.

## 1. INTRODUCTION

Speech processing has been an important area of research for scientists. It can be used for linguistic study of a language as well as in identifying features of speaker.

Speech processing is further classified into various areas like speech recognition, speech synthesis, speaker identification, speaker verification etc. All these fields require knowledge of language of speech.
A lot of research on speech processing has been done for English and some other languages, but this is not the case with Urdu.

This study is an attempt to find out some speaker dependant features of Urdu speech. These features may prove helpful in various research fields. It focuses the speaker dependant features in approximants of Urdu language. The aim of this study was to find out the minimal set of parameters in approximants of Urdu through which a native speaker can be easily identified.
Lack of such efforts in the past, made it a more challenging one. Also the complexity, versatility and uniqueness of the language also made the research pretty interesting.

## 2. LITERATURE REVIEW

The research on speaker identification is so far totally dominated by English Language. In order to proceed with this research, it was required that the previous studies and developments should be explored. For this purpose a good understanding of speaker identifications' domain and the acoustic features of human voice was necessary. So, mainly speaker identification is divided in two domains:

- Speaker verification and
- Speaker identification

In Speaker verification, speaker claims to be a particular person from the existing database of the system. The speaker identification system then matches the voice of the claimer with an existing sample of that particular speaker in the database for verification. As a result the system either accepts or rejects.

In speaker Identification, any user can input voice sample to the system for identification. In this scenario the system has no claimer, so the system searches its whole database to find a match for the input signal or a closest one to validate the user or rejects him or her (Cambridge research laboratories, 2001).

The types of techniques that are used to either for the verification or identification of the speaker are same. So form here onwards all the techniques applies on both and we generally treat it as speaker identification.

Speaker identification is further classified into:

- Text-dependent and
- Text-independent

As the names specifies, the text-dependent recognition requires the analysis of speech done by keeping a particular text into consideration or considering certain language, while in case of text independent there is no language specification or text compulsion under consideration to recognize the speaker, and all the sounds are treated as general IPA phonemes.

The performance of the system depends to a large extent on selecting features that minimize the intra-speaker variability while maximizing the inter-speaker variability. Acoustic analysis of speech plays an important role in finding out more about the speaker dependant features. The various features has been studied are as follows(Hollien and Dr. Jong, 1996):

- Fundamental frequency: is one of the most important features but there are still problems in measuring it especially in a noisy environment and also due to constant dc current in electrical circuits.
- Another robust feature is Long-term spectrum.
- Formant frequencies also exhibited good results in this field but they are hard to measure accurately.
- Other Digital Signal Processing (DSP) techniques are also used in order to identify speaker e.g. Linear Predictive Coding (LPC), Cepstral Coefficients etc. but they are beyond the scope of this paper.

Intensity, bandwidth, nasalization, and lip rounding etc. can also play a part in this regard (Kent and Read, 1992).

## 3. Problem Statement

This paper tries to identify the speaker by finding features with in approximants, which vary from speaker to speaker. While saying approximants the breath stream, passing through the vocal tract, becomes turbulent because of friction offered by the structure of the vocal tract and also the formation of teeth. As a result, generate a sound having constriction more than vowels and less than fricatives. So the speech contains each speaker's vocal tract information. There is a possibility that we can identify a speaker on the basis of his/her vocal tract information in addition to it we can have the source features too (Pickett, 1999). So, if we can some how get the feature out from source and filter we can possibly identify a speaker. In Urdu, we have only /l/ and /j/ as approximants.

## 4. Methodology

### 4.1 Software Selection

The analysis is done using esps/xwaves because it provides us with spectrum, spectrogram, formants and their bandwidths etc.

### 4.2 How the experiment was performed?

#### 4.2.1 Conditions

To perform the experiment it was ensured that none of the outside factors should affect the speech sound. In the earlier research, noise of the environment, distance of microphone from the speaker, speech instances after different intervals of time offered great problems.

So, it has been tried that the distance of the speaker and the microphone remains constant for all the speakers over different instances.

Finally, different sentences of same speakers' were recorded at different instances of time so that the range of a particular subject accent could be defined as clearly as possible.

#### 4.2.2 Speaker Selection

Recordings were gathered from seven speakers, four males (A, B, C, D) and three females (X, Y, Z). These speakers were chosen with extreme care in such a way that the speaker does not know the phonetics of Urdu, so that they do not try to make approximants as they are phonetically but rather speak them in the natural way. In other words non-phonations native speakers of Urdu were selected but they were also aware of other languages like English and Arabic.

Age of the speaker, another important factor, ranges from 18 – 23 years with an average of 20. All the speakers had spent at least 15 years of their lives in Pakistan speaking Urdu and used it regularly in their everyday life. All the speakers are students at FAST-NU, Lahore.

The most important thing in speaker selection was that no speaker had any hint of the experiment.

### 4.2.3 Criteria of words selection

The words were selected such that they were:

1. Actual words of Urdu language and are commonly used in our everyday life.
2. Intervocalic i.e. the target phoneme (the phoneme /l/ or /j/) is preceded and followed by a vowel. So that the transition of vowel to the target phoneme and back to vowel is quite evident making it easy to measure different parameters for approximants.

Phonetic representations of the selected words along with English meanings of the selected words are given below:

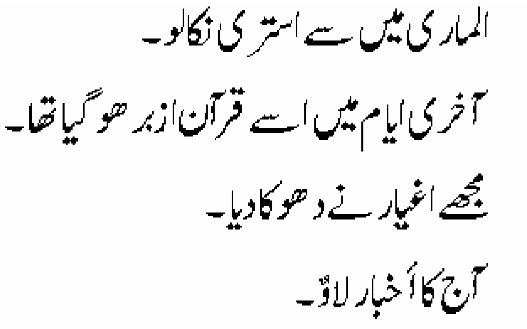
| Phonetic | English |
|---|---|
| əlmɑri | Cup-board |
| əjɑm | Days |
| əɤjɑɾ | Strangers |
| lɑo | Bring |

### 4.2.4 Criteria of sentence selection

An other important thing that was kept in mind during the experiment was to record such sentences, which were spoken in normal everyday life.

Also the sentences were neither too long that the speaker lose his normal voicing routine and nor too short that it effects the words under analysis.

Following were the selected sentences with their transcription and target words underlined:



الماری میں سے استری نکالو۔

آخری ایام میں اسے قرآن ازبر ھو گیا تھا۔

مجھے اغیار نے دھوکا دیا۔

آج کا اخبار لاؤ۔

1. əlmɑri me se ɪstəri nɪkɑlo.
2. ɑɤrɪ əjɑm me ʊsɑ kʊrɑn əzbɑɾ ho gjɑ t$^{h}$ɑ.
3. mʊdʒhe əɤjɑɾ ne d$^{h}$okɑ dɪjɑ.
4. ɑdʒ kɑ ɑxbɑɾ lɑo.

### 4.2.5 Handling of other Speaker Errors

Other speaker errors like saying of sentences in a stretch of ten instances effects speaker’s normal accent because of out of breath. This was handled by recording the same sentence in chunks of three, three and four instances at different time slots.

In addition to that the sentences were recorded one by one in order to keep the normal accent of the speaker i.e. normally in our everyday life we do not say a sentence more than once in a stretch. Further clarifying, if the total number of sentences are four and total number of instances required are ten then the four sentences were recorded in three time slots of three, three and four instances of each sentence and in one slot all the sentences were recorded in sequence i.e. one after the other.

### 4.2.6 Parameters to be measured

As discussed in literature review that various studies come up with different parameters for speaker identification like, genders variation, accent, age, speech rate, place where person lives, and phones realizations, are important issues in speaker identification. This experiment focuses on finding speaker dependent features from the most prominent acoustic cues of former experiments e.g.

- Formant values of F1, F2, F3, F4
- Fundamental frequency

This experiment also measures the following features, which are supported by ensig/xwaves:

- The duration of the approximants spoken by each speaker
- Intensities of all formants
- Bandwidth of all formants
- Intensity of first two peaks (I11 & I12 respectively) of spectrum
- Power analysis[1] (P) of approximants
- Zero cross analysis[2] (Z) of approximants

### 4.2.7 Recording procedure

Before recording, each speaker was instructed, that how he or she is supposed to record his or her voice as discussed above.

## 5. RESULTS

The results of the above recordings were shown in Appendix A. The values were rounded up by two decimal places. Each row corresponds to a speaker showing the average value over ten recordings of all the seven speakers. Furthermore, all the readings are relative to a vowel /a/ following /l/ i.e. the value of power of /l/ divided by the power of /a/. Doing so would overcome the problem of speaking at a different rate than normal because the ratio would remain the same.

### *5.1 Authentication of results*

The authentication of the results were achieved in such a way that the results were averaged over ten samples of each speaker corresponding to each approximant and in case, if any sample showed too much variability with the pervious results, it was dropped.

## 6. DISCUSSION

### 6.1 Analysis of /l/

In Table 1, the first cue of power analysis, could distinguish each speaker from other except X and C. The interesting thing to note here is that X and C have different sex.

Similarly the second cue, again can distinguish speakers on the basis of zero-cross analysis except Y and Z because there values of zero-cross analysis is too close to one another.

The third cue specifies the duration for which /l/ is spoken by a particular speaker, but there was a problem in determining a speaker solely on the basis of duration because some people have very close correspondence.

But a speaker can be identified completely by comparing the first two cues together or first and third cues or even second and third cues.

In Table 2, the first two cues showed the intensity of two maximum peaks in the spectrum of /l/ of each speaker but the above results showed that a person can not be identified completely on the basis of the intensities because it is dependent upon the stress laid by the speaker at different instances. At one instance speakers' recorded the sentences loudly while at other the recordings were converse of the first. So, first two intensities are not a good cue for identifying a speaker. While fundamental frequency is totally opposite to the intensities and a speaker can be identified easily by f0, except C and D. But still it is a good cue for identifying.

Table 3 shows that frequency peak of formant one and its bandwidth can together identify a speaker completely. However, intensity of the formant is not a good cue. On the other hand, the Tables 4 to 6 do not give a reasonable cue for speaker identification.

So, as a result of the analysis of /l/, two different types of cues could be formulated i.e.

[1] For further details of Power analysis see appendix B

[2] For further information of Zero-cross analysis see appendix B

#### 6.1.1 Primary cues

- Power analysis
- Zero-crossing analysis
- Frequency of first formant
- Bandwidth of first formant

#### 6.1.2 Secondary Cues

- Duration
- Fundamental frequency
- Bandwidth of fourth formant

### *6.2 Analysis of /j/*

In Table 7, power analysis and zero-cross analysis both identifies a speaker somewhat but even if both are used together still unable to identify all the speakers completely. In this situation another cue of duration can be used together with power analysis or zero-cross analysis to completely identify all speakers.

Table 8 shows similar results as of Table 2, showed some identification of speaker but there is still some confusion like in determining A and B or C and D.

In Table 9, both frequency and bandwidth of formant one can individually identify each speaker. While intensity showed too much variability.

Tables from 10 to 12, showed extreme detection of speaker by only considering bandwidth of formant two and bandwidth of formant four. However, in case of bandwidth of formant three there was a problem in deciding C and D.

#### 6.2.1 Primary cues

- Frequency of first formant
- Bandwidth of second formant
- Bandwidth of fourth formant

#### 6.2.2 Secondary Cues

- Fundamental frequency
- Frequency of second formant
- Bandwidth of third formant
- Power analysis
- Zero-crossing analysis
- Duration

## 7. SUGGESTED PARAMETERS

In the light of above discussion, this paper finally suggests a number of acoustic parameters some are discussed in this paper while other are expected to show response in identifying a speaker, those are namely:

- Duration
- Formants frequency values
- Formants bandwidth values
- Power analysis
- Zero-cross analysis
- Values of formants entering in to approximants and going out of them
- Fundamental frequency
- Pattern matching of fundamental frequency wave pattern
- Pattern matching of higher formants wave pattern

## 8. PROBLEMS

The major problems faced during the experiments were mainly:

- The analysis software xwaves was unable to calculate the formants efficiently and as a result, all the readings are done manually (i.e. by measuring at different points on the spectrum).
- Other problem was that speakers mix approximants with preceding and following vowels making it quite hard to distinguish between the two.
- Finally, sometimes the variable exclamations of the speakers at different time intervals make the results to vary too much.

## 9. EFFECT OF INCREASE IN SPEAKERS

When the speaker increases the speakers can still be identified up till a certain threshold by combining all the cues, which can identify speakers. But if our database of speakers still keep on increasing then we have to find some other helping cues to do the job.

## 10. REFERENCES

Cambridge research laboratories. 2001. Speaker Recognition". Website Link: http://crl.research.compaq.com/speech/speaker_rec.html

*Entropic, Inc. 1998.* Esps/xwaves manual*, 400 North Capitol Street. NW. Suite G-100 Washington. USA.*

Hollien. H, Dr. Jong. G. D, Martin. C. 1996. "Speaker identification by machine".

*Huang*. C, Chen. T, Li. S, Chang. E. & Zhou. J. 1993. "Analysis of Speaker Variability". Microsoft Research, China, 5F, Beijing Sigma Center, No. 49, Zhichun Road Haidian District, Department of Automation, Tsinghua University.

Kent, R, D. & Read, C. 1992. *The Acoustic Analysis of Speech*. University of Wisconsin-Madison, San Diego, California, USA.

Pickett, J.M. 1999. *The Acoustics of Speech Communication.* A ViaCom Company, USA.

## 11. APPENDIX A

Data of approximant /l/

**TABLE 1 Power, zero-crossing analysis and duration**

| Speaker | P | Z | D |
|---|---|---|---|
| A | 43 | 52 | 52 |
| B | 36 | 41 | 48 |
| X | 60 | 82 | 123 |
| Y | 133 | 29 | 85 |
| C | 62 | 55 | 58 |
| Z | 120 | 30 | 79 |
| D | 84 | 65 | 76 |

**TABLE 2 Intensity of first two peaks in the spectrum and fundamental frequency**

| Speaker | I11 | I12 | F0 |
|---|---|---|---|
| A | 99 | 98 | 145 |
| B | 84 | 82 | 134 |
| X | 95 | 94 | 218 |
| Y | 96 | 95 | 240 |
| C | 98 | 106 | 112 |
| Z | 105 | 90 | 214 |
| D | 98 | 96 | |

**TABLE 3 Frequency, bandwidth and intensity of formant one**

| Speaker | F1 | B1 | I1 |
|---|---|---|---|
| A | 47 | 66 | 99 |
| B | 51 | 82 | 89 |
| X | 72 | 13 | 113 |
| Y | 52 | 18 | 106 |
| C | 82 | 134 | 96 |
| Z | 39 | 21 | 107 |
| D | 62 | 101 | 95 |

**TABLE 4 Frequency, bandwidth and intensity of formant two**

| Speaker | F2 | B2 | I2 |
|---|---|---|---|
| A | 115 | 82 | 92 |
| B | 108 | 123 | 96 |
| X | 67 | 89 | 92 |
| Y | 152 | 92 | 81 |
| C | 150 | 169 | 87 |
| Z | 99 | 233 | 84 |
| D | 95 | 203 | 89 |

**TABLE 5 Frequency, bandwidth and intensity of formant three**

| Speaker | F3 | B3 | I3 |
|---|---|---|---|
| A | 113 | 200 | 81 |
| B | 112 | 104 | 98 |
| X | 90 | 199 | 88 |
| Y | 114 | 194 | 88 |
| C | 95 | 318 | 93 |
| Z | 96 | 92 | 88 |
| D | 69 | 167 | 72 |

**TABLE 6 Frequency, bandwidth and intensity of formant four**

| Speaker | F4 | B4 | I4 |
|---|---|---|---|
| A | 104 | 127 | 102 |
| B | 86 | 130 | 98 |
| X | 106 | 223 | 85 |
| Y | 98 | 136 | 77 |
| C | 86 | 171 | 92 |
| Z | 99 | 158 | 85 |
| D | 114 | 138 | 84 |

Data of approximant /j/

**TABLE 7 Power, zero-crossing analysis and duration**

| Speaker | P | Z | D |
|---|---|---|---|
| A | 63 | 62 | 114 |
| B | 56 | 52 | 125 |
| X | 74 | 72 | 102 |
| C | 96 | 105 | 129 |
| D | 69 | 62 | 136 |
| Z | 131 | 232 | 94 |
| Y | 108 | 48 | 99 |

**TABLE 8 Intensity of first two peaks in the spectrum and fundamental frequency**

| Speaker | I11 | I12 | f0 |
|---|---|---|---|
| A | 98 | 97 | 148 |
| B | 100 | 97 | 142.5 |
| X | 97 | 108 | 285 |
| C | 102 | 98 | 110.4 |

| | | | |
|---|---|---|---|
| D | 99 | 98 | 107 |
| Z | 330 | 101 | 225 |
| Y | 100 | 102 | 236 |

**TABLE 9 Frequency, bandwidth and intensity of formant one**

| Speaker | F1 | B1 | I1 |
|---|---|---|---|
| A | 61 | 41 | 95 |
| B | 68 | 72 | 102 |
| X | 87 | 36 | 130 |
| C | 57 | 91 | 95 |
| D | 73 | 140 | 101 |
| Z | 81 | 53 | 100 |
| Y | 140 | 63 | 108 |

**TABLE 10 Frequency, bandwidth and intensity of formant two**

| Speaker | F2 | B2 | I2 |
|---|---|---|---|
| A | 146 | 77 | 101 |
| B | 154 | 135 | 81 |
| X | 127 | 105 | 79 |
| C | 121 | 403 | 83 |
| D | 133 | 184 | 98 |
| Z | 124 | 364 | 104 |
| Y | 151 | 134 | 95 |

**TABLE 11 Frequency, bandwidth and intensity of formant three**

| Speaker | F3 | B3 | I3 |
|---|---|---|---|
| A | 94 | 579 | 98 |
| B | 101 | 270 | 98 |
| X | 161 | 83 | 78 |
| C | 94 | 112 | 107 |
| D | 92 | 111 | 100 |
| Z | 118 | 91 | 102 |
| Y | 145 | 149 | 103 |

**TABLE 12 Frequency, bandwidth and intensity of formant four**

| Speaker | F4 | B4 | I4 |
|---|---|---|---|
| A | 97 | 113 | 107 |
| B | 104 | 73 | 94 |
| X | 98 | 469 | 94 |
| C | 100 | 94 | 100 |
| D | 99 | 112 | 89 |
| Z | 93 | 103 | 112 |
| Y | 97 | 80 | 101 |

## 12. APPENDIX B

Zero-crossing rate

Zero-crossing rate is computed by multiplying the number of zero crossing by the sampling frequency and dividing by the number of samples in the frame. Thus, the units are zero-crossings per second.

Power Analysis

Power analysis is computed by summing of the squares of the sampled data values and dividing by the number of points in the frame.