ASR Data Cleaning Guidelines

Presenter: Asima Hameed

6th May, 2014 | Center for Language Engineering (CLE)

Data Cleaning Process

The speech files will be cleaned on the basis of:

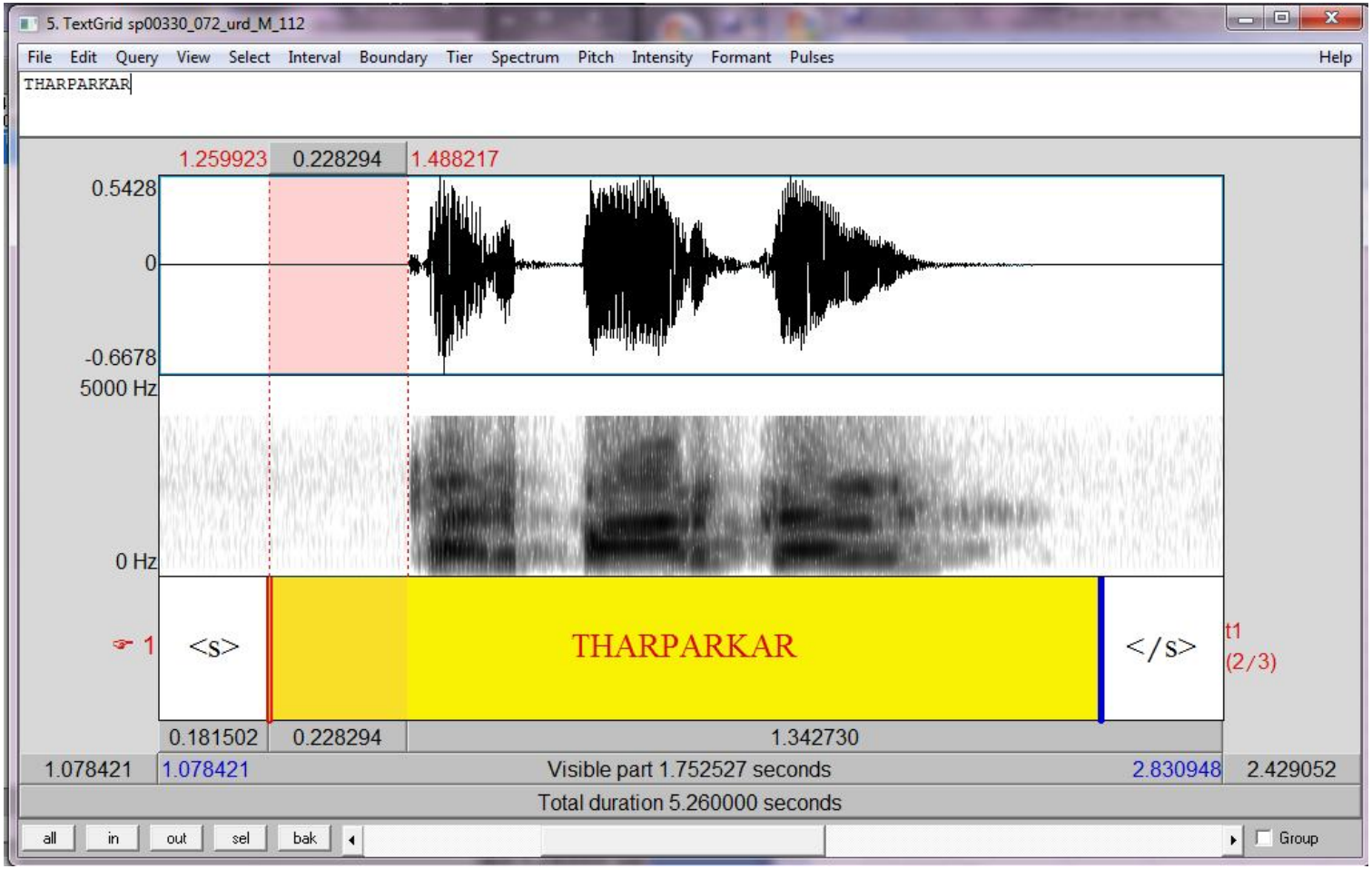
- Silence
- Noise
- Mispronunciation

Silence

1. A silence period of at least 200 ms
2. For voiced and voiceless consonants
3. Noise in silence period
4. No space for silence
   - At onset
   - At offset
   - On both ends of the word

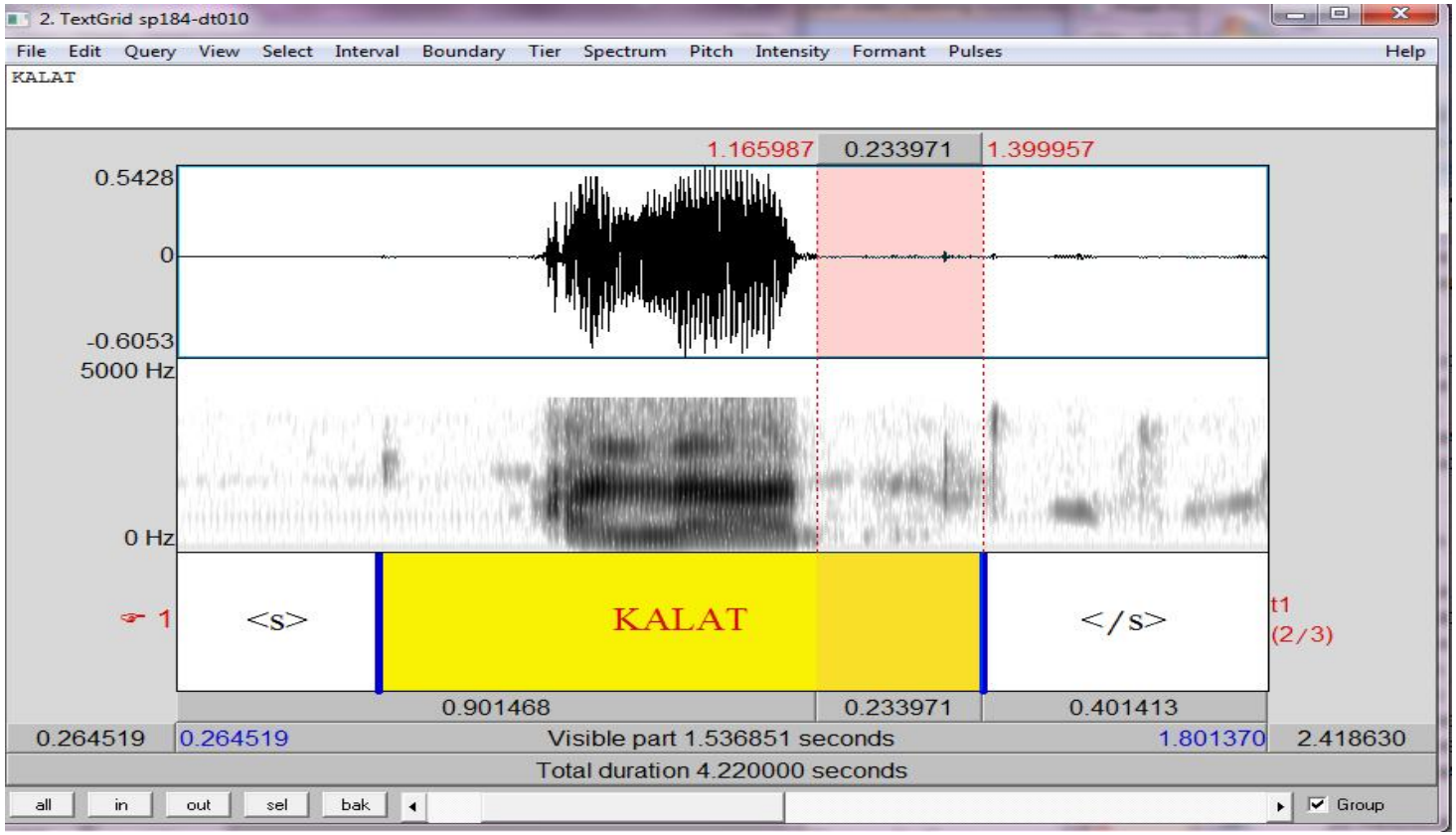
ASR Data Cleaning Guidelines

1. Silence period of at least 200 ms

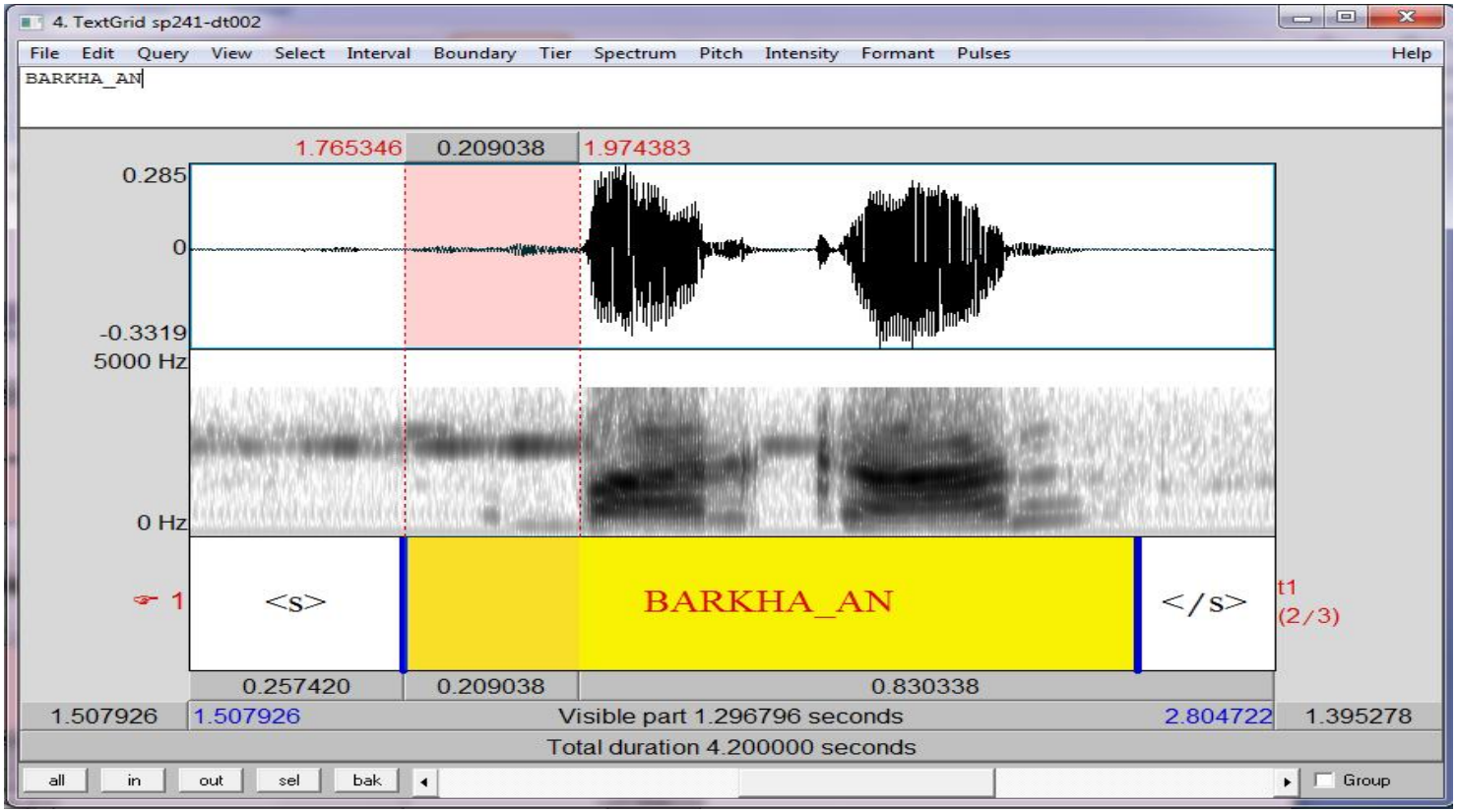


6th May, 2014	Center for Language Engineering (CLE)

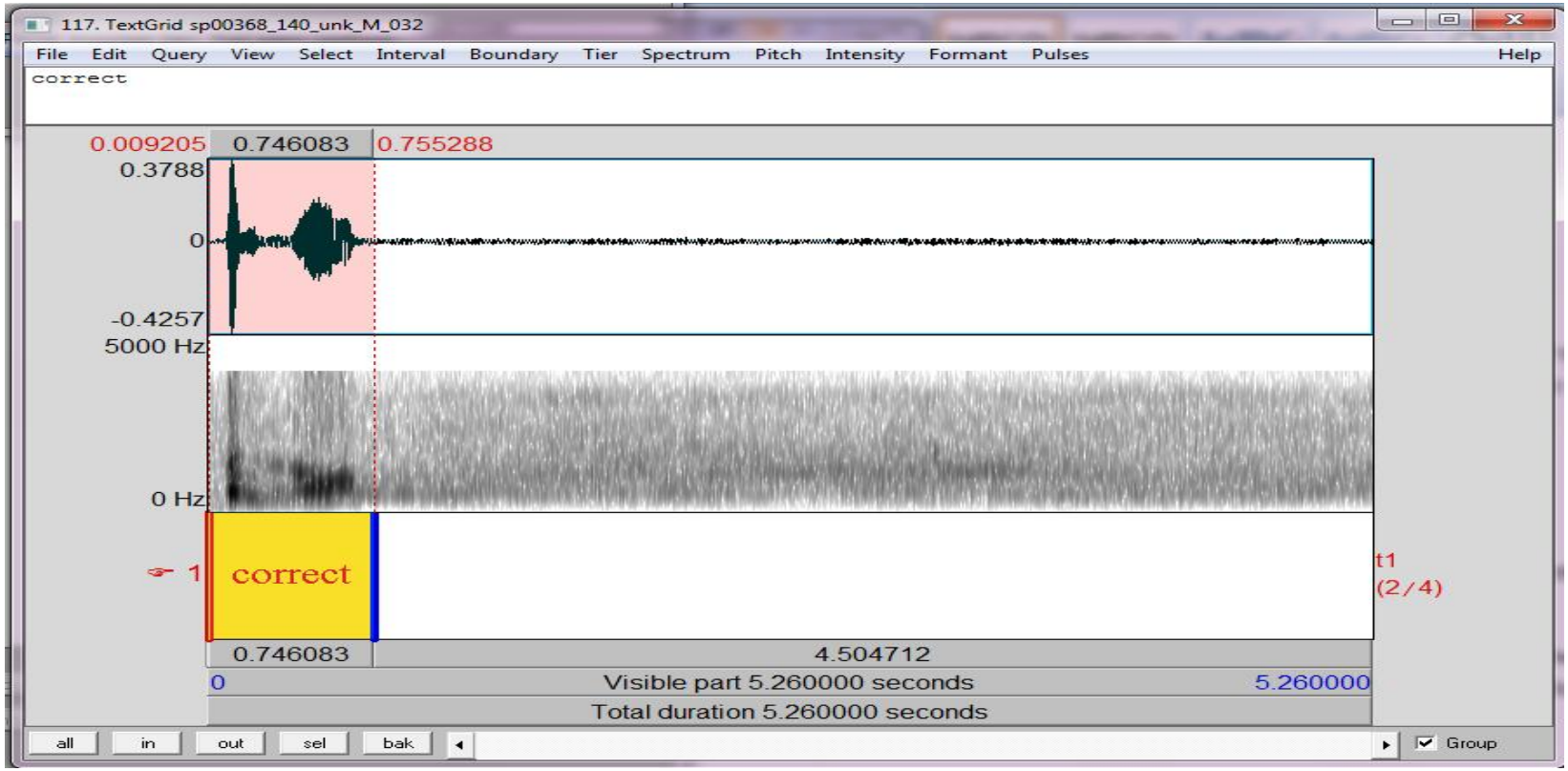
2. Voiced and voiceless regions



3. Noise in silence period



4. No space for silence



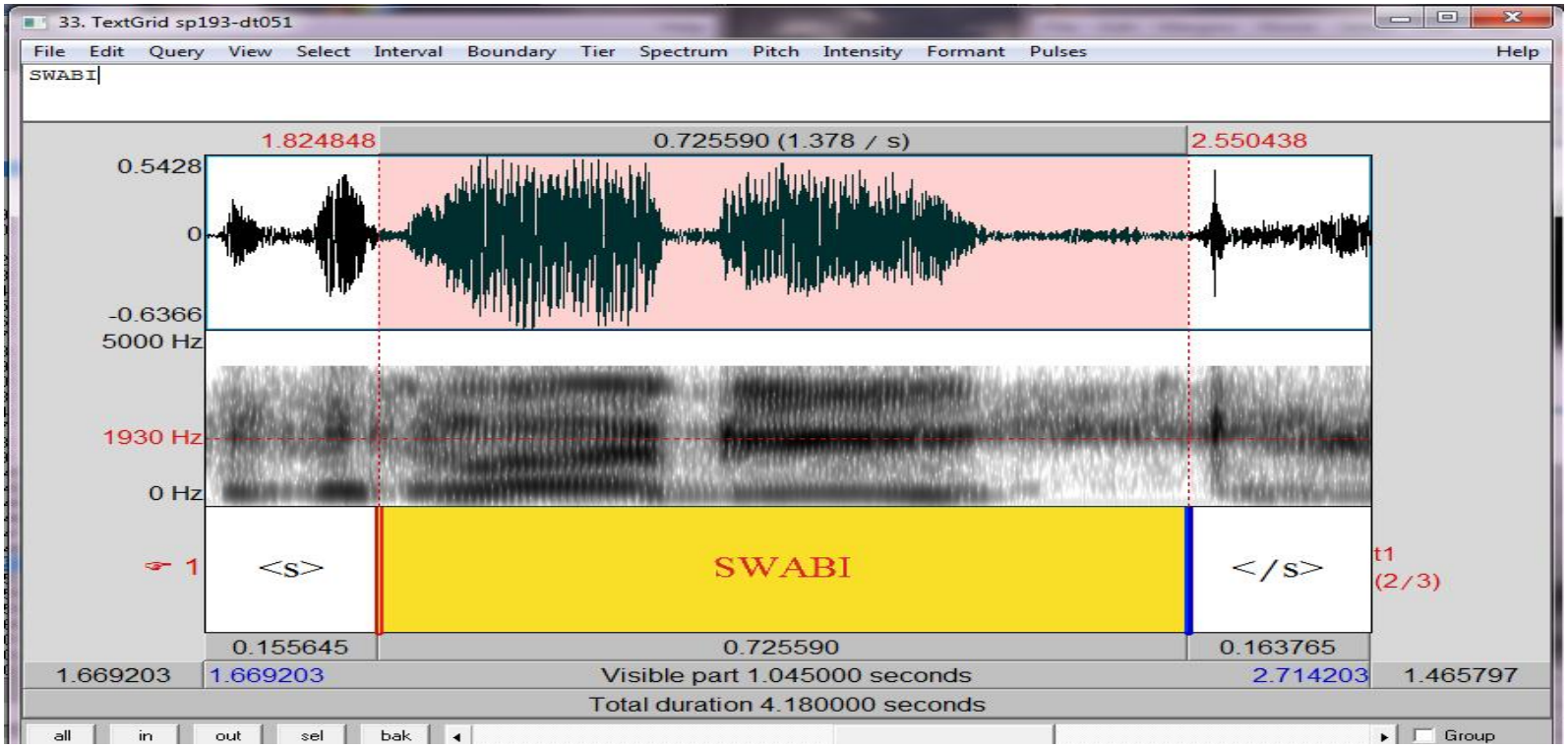
Noise

Generally, three types of noise have been found:

- Traffic Noise
- People Noise
- Babble Noise (any meaningless noise)

Cont'd

- Over-lapping Noise
- Non-over-lapping Noise
  - SNR (signal to noise ratio)
  - SNR >= 10bB



Mispronunciation:

Generally we find three kinds of mispronunciations, related to:

1. Alternate Pronunciation
2. Consonant
3. Vowel

1. Alternate Pronunciation

Acceptable across the accents because of general trend.

E.g. MALAKAND with MLA_AKAND

BATAGRA_AM with BATGRA_AM

FAISALABAD into FAISLABAD

MUZAFFARABAD into MUZAFFRABAAD

BAHAWALNAGAR into BHAWALNAGAR

2. Consonant:

Consonants mispronunciations are taken into account on these parameters.

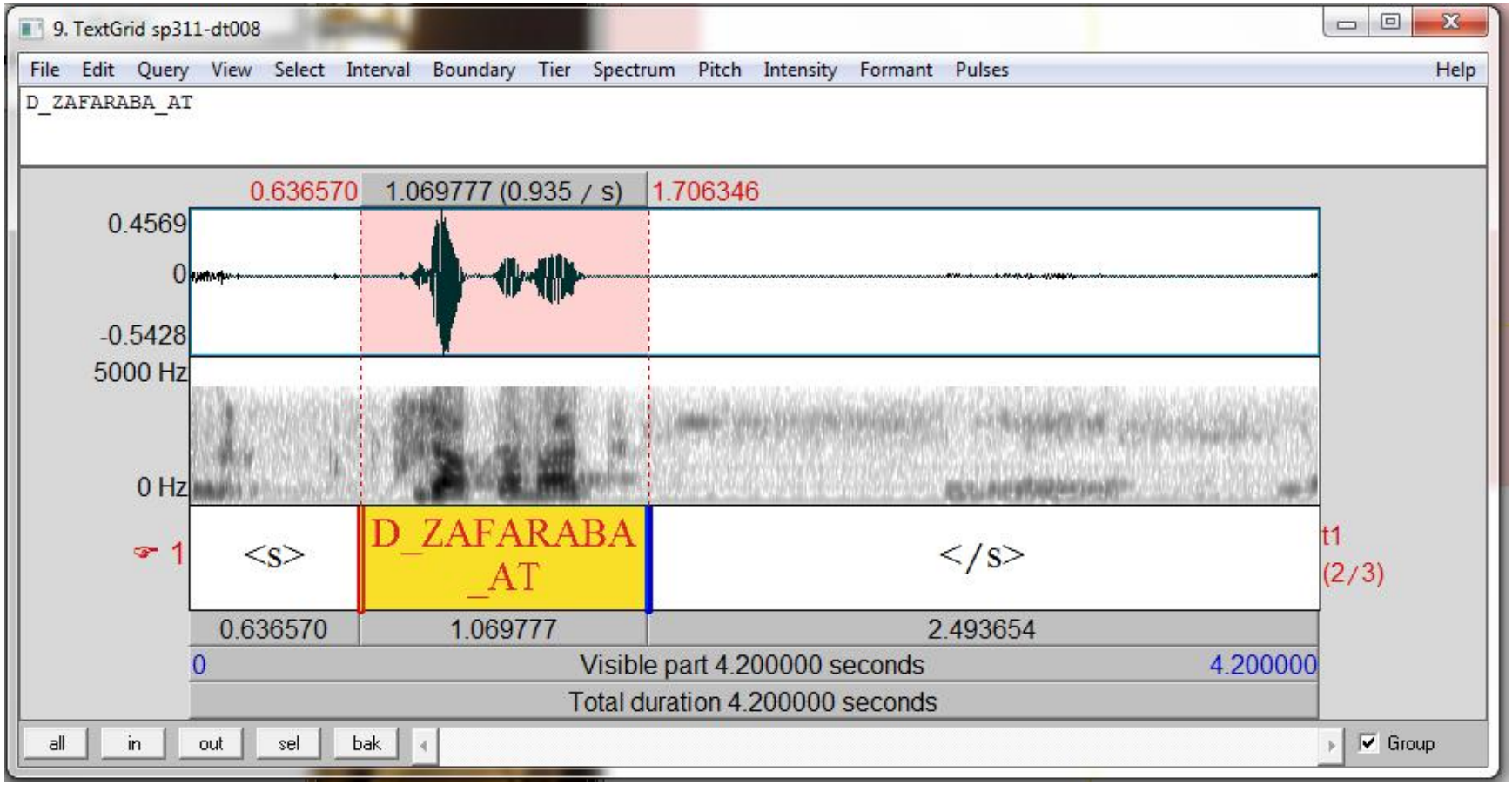
a. Substitution of a Consonant
b. Deletion of a Consonant
c. Insertion of a consonant

a. Substitution of a Consonant

1. Variation in voicing will be acceptable.
   E.g. PARKHAN instead of BARKHAN
2. Variation in place and manner of articulation will not be acceptable. E.g. MATIARI instead of PATIARI.
3. Aspirated consonant into non- aspirated and vice versa will be acceptable.
4. Flap is substituted by trill
5. Trill cannot substitute Flap
6. Exceptions

ASR Data Cleaning Guidelines

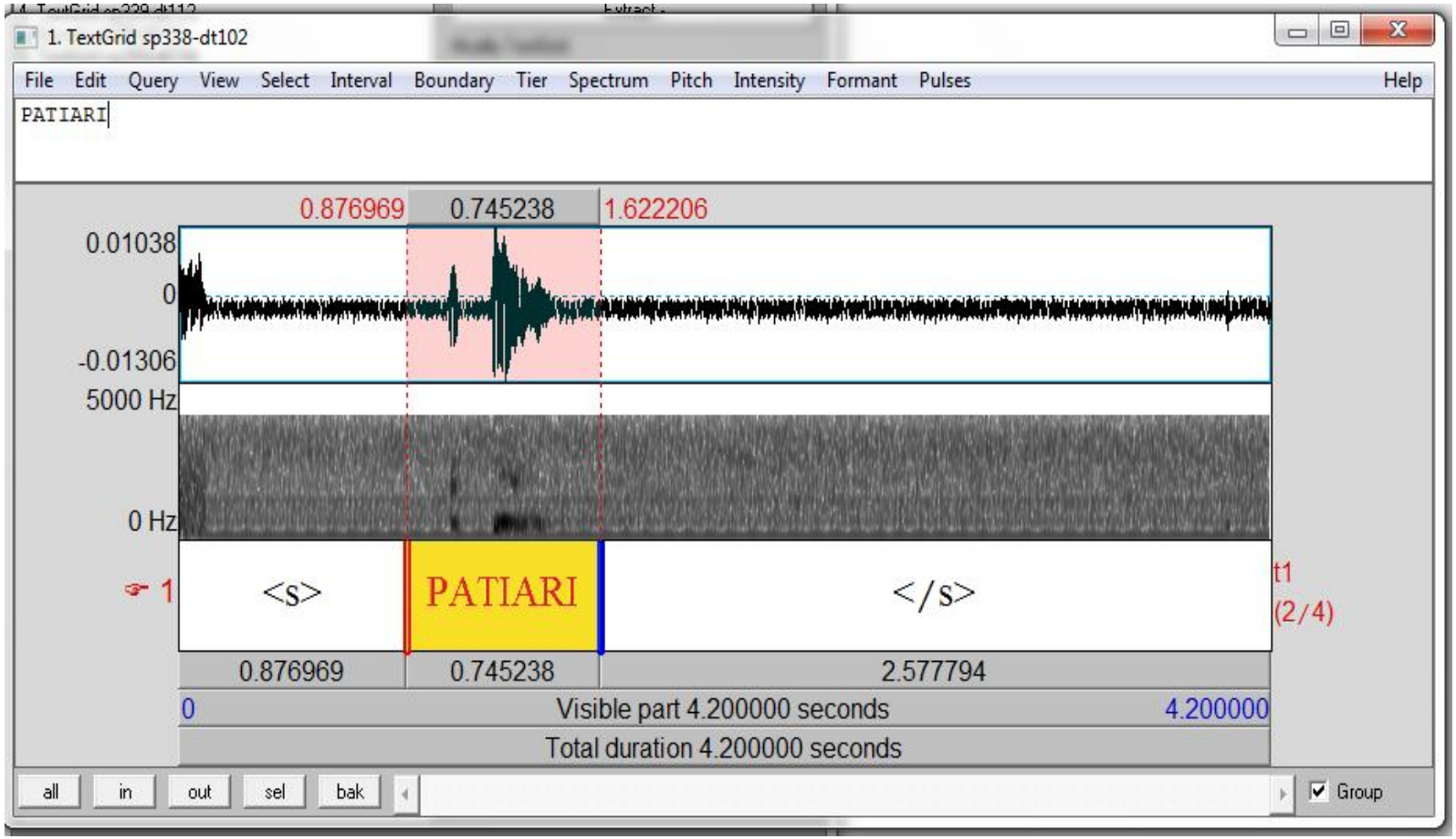
1. Variation in voicing



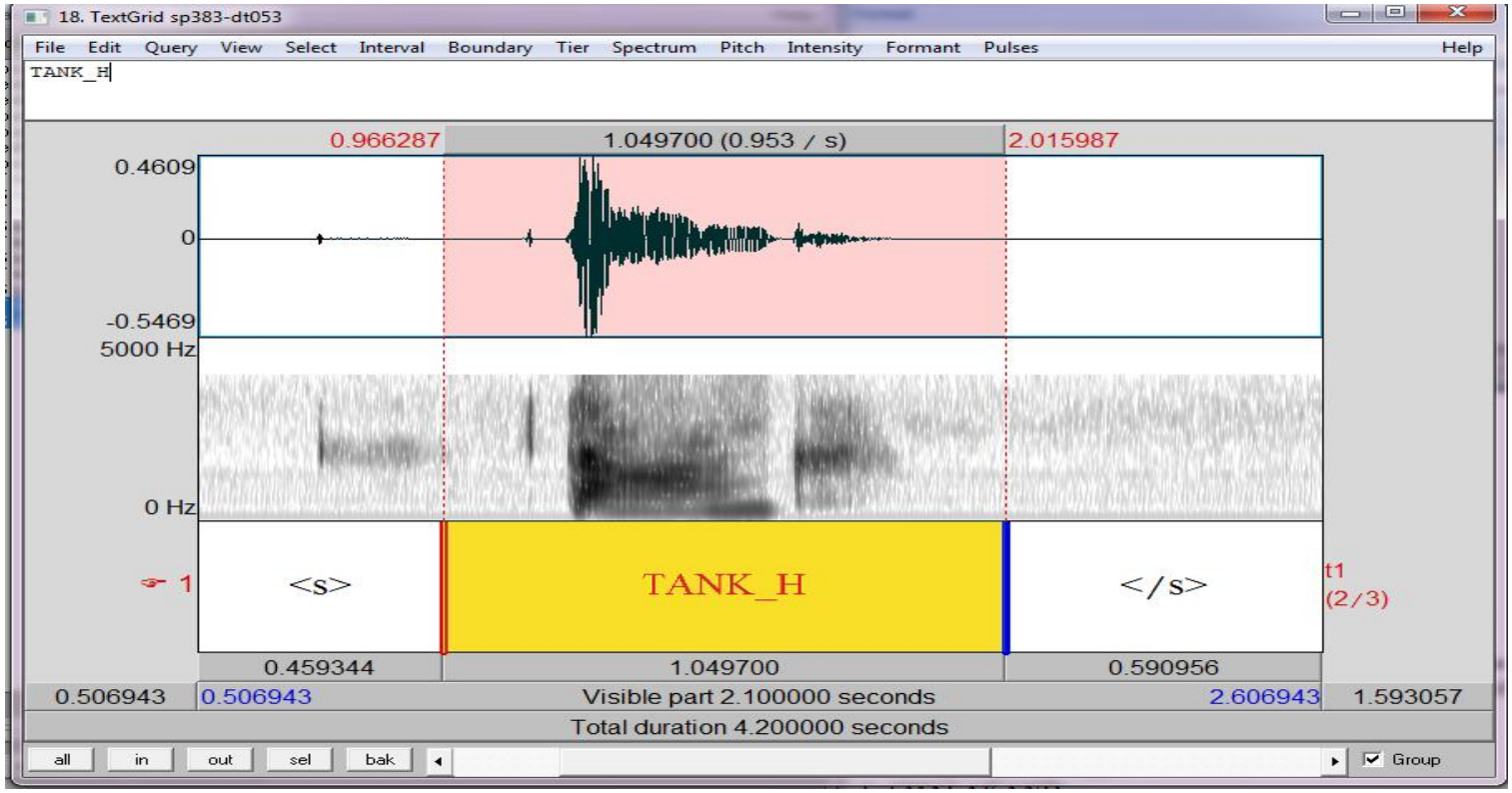
6th May, 2014	Center for Language Engineering (CLE)

ASR Data Cleaning Guidelines

2. Variation in place and manner of articulation

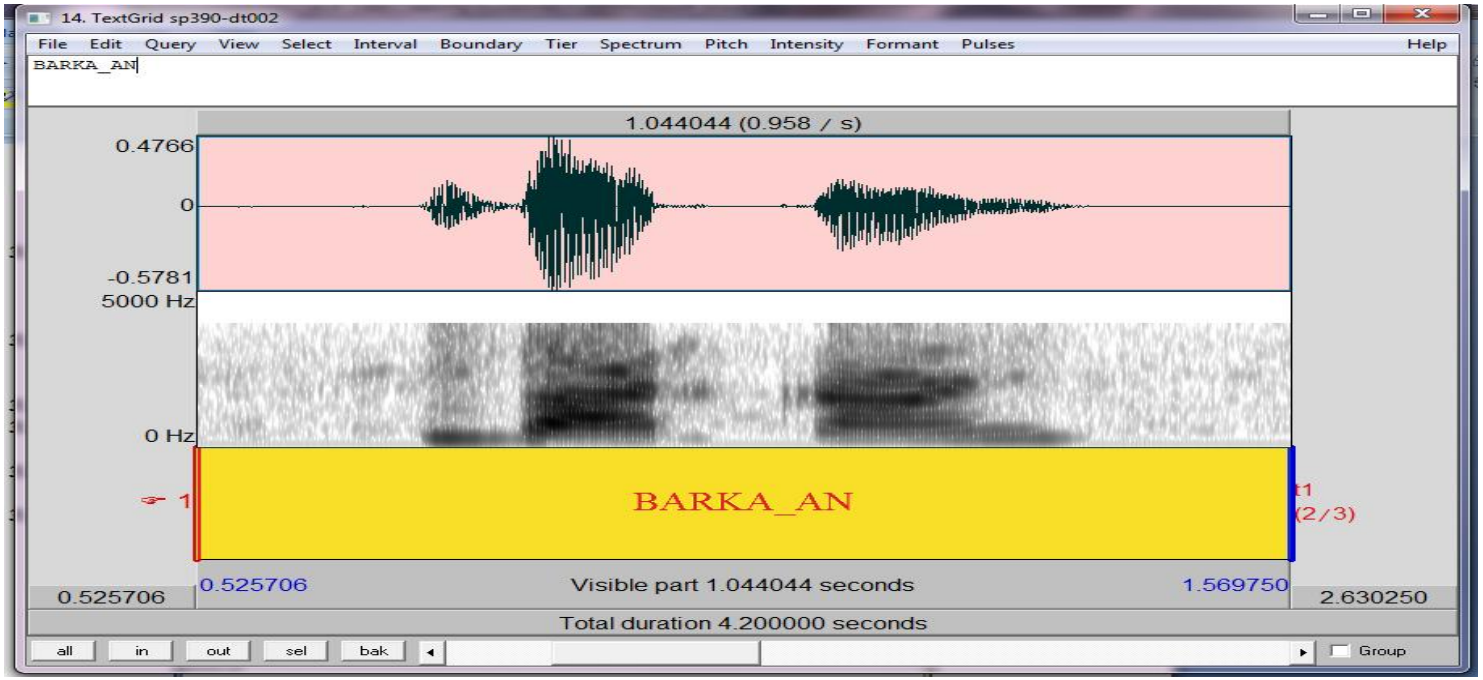


3. Aspirated consonant into non- aspirated and vice versa



ASR Data Cleaning Guidelines

Cont'd



4. Flap; / ɽ / ڑ have been substituted by trill; /r/ ر
   E.g. BAD_ZOR_R into BAD_ZOR,
5. Trill have not been substituted by flap.
   E.g. K_HUZD_DA_AR not into K_HUZD_DA_AR_R
6. Exceptions

b. Insertion of a Consonant

- Initial Position

- Middle Position

  E.g. GUJRA_ANWALA instead of GUJRA_A_NWALA.

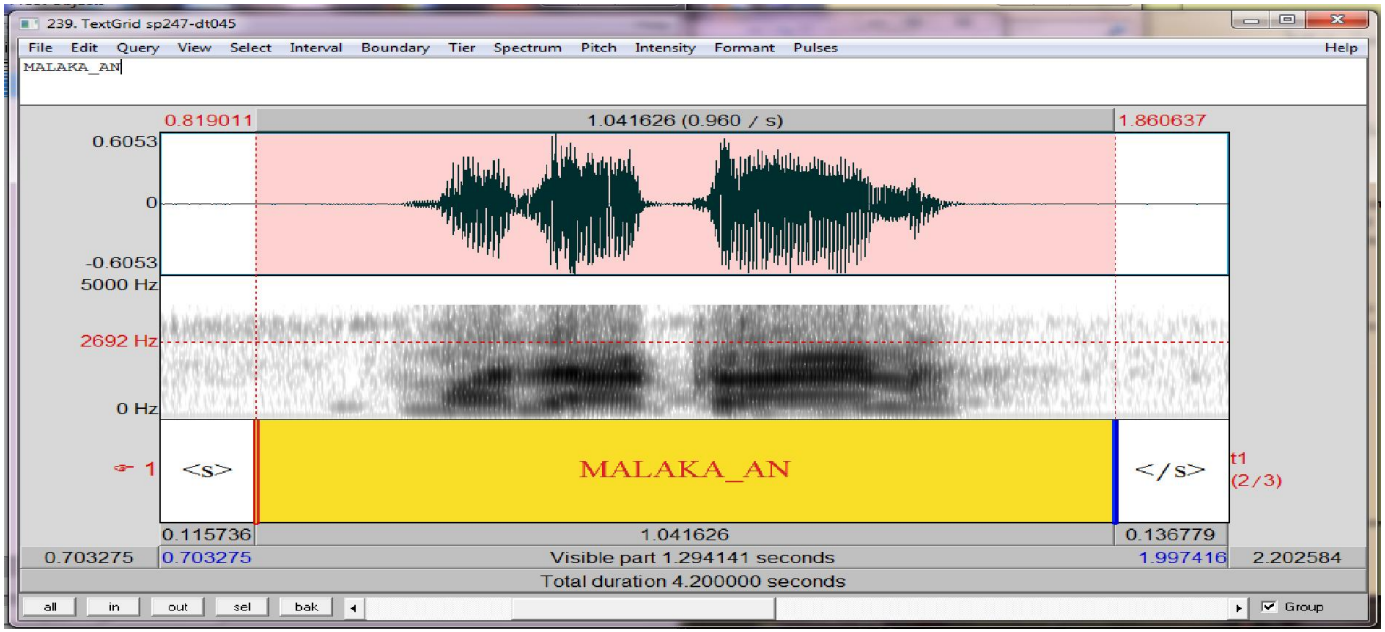
- Final Position

  E.g. LOD_DHRA_AN instead of LOD_DHRA_A_N

c. Deletion of a Consonant:

- Initial Position
- Middle Position
- Final position

E.g BA-AD_ZO_O instead of BA_AD_ZO_OR_R



3. Vowels:

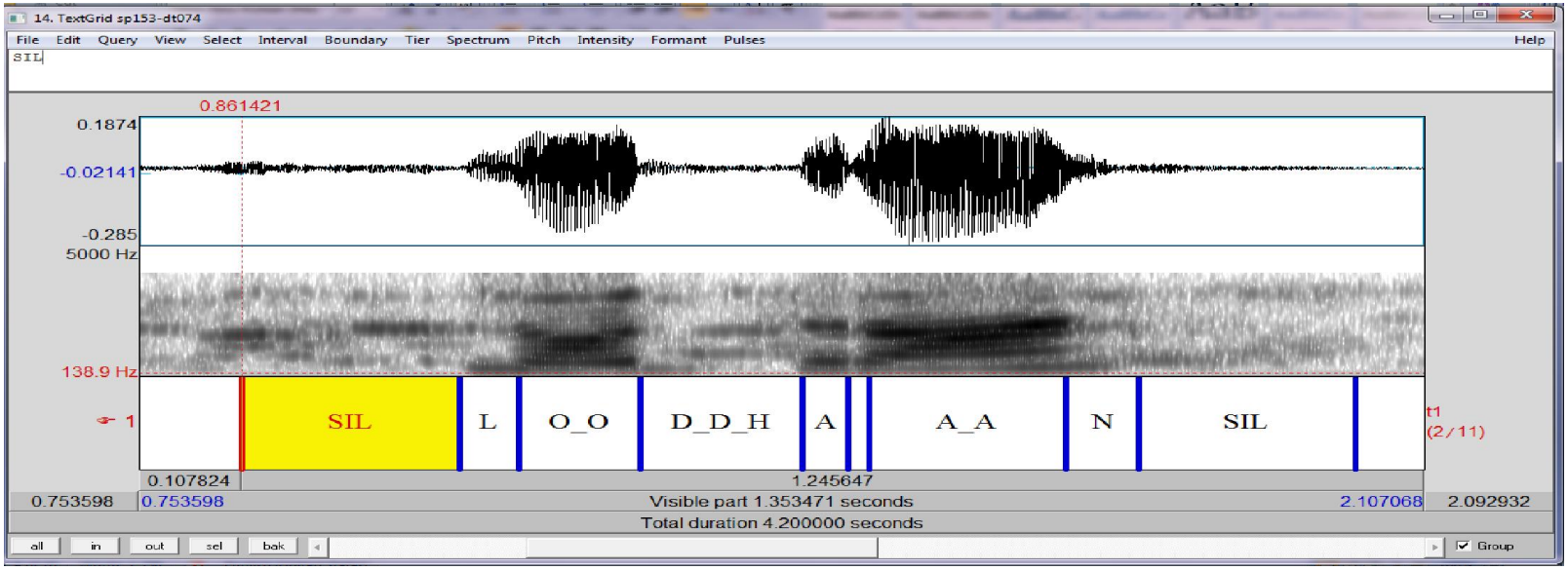
Vowel mispronunciations are taken into account on these parameters.

a. Addition of a Vowel
b. Substitution of a Vowel
c. Deletion of a Vowel

a. Addition of a Vowel:

- Initial Position:
- Middle Position:

  E.g. LOD_D_HRA_A_N into LOD_D_H_ARAN,

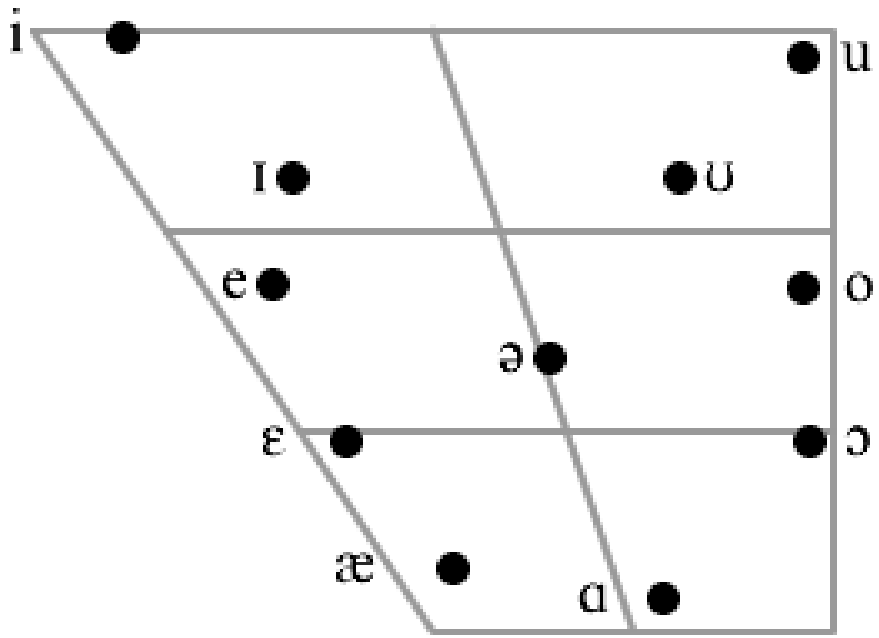


- Final Position:

  E.g. FA_ESALABA_AD_D into FA_ESALABA_AD_DA

b. Substitution of a vowel:

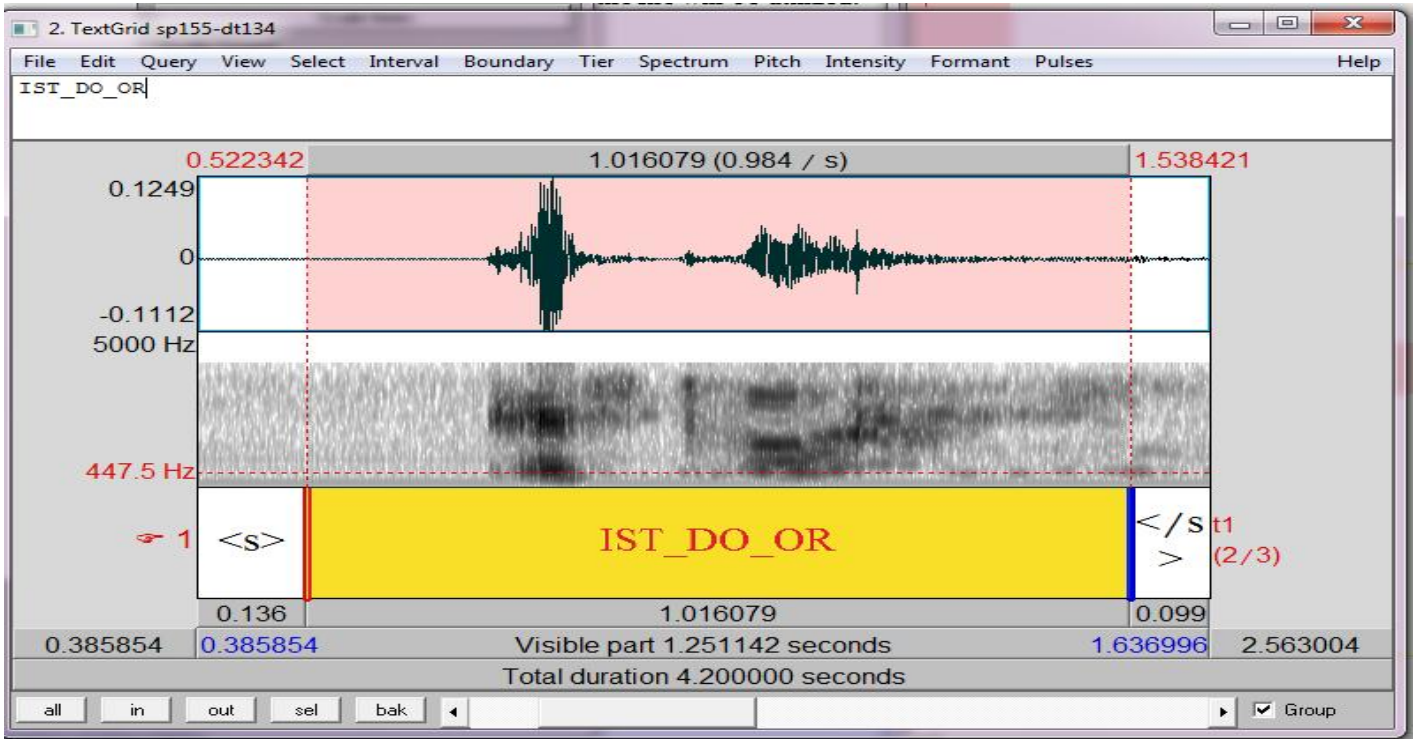
- Neighboring vowels in quadrilateral chart



Cont'd

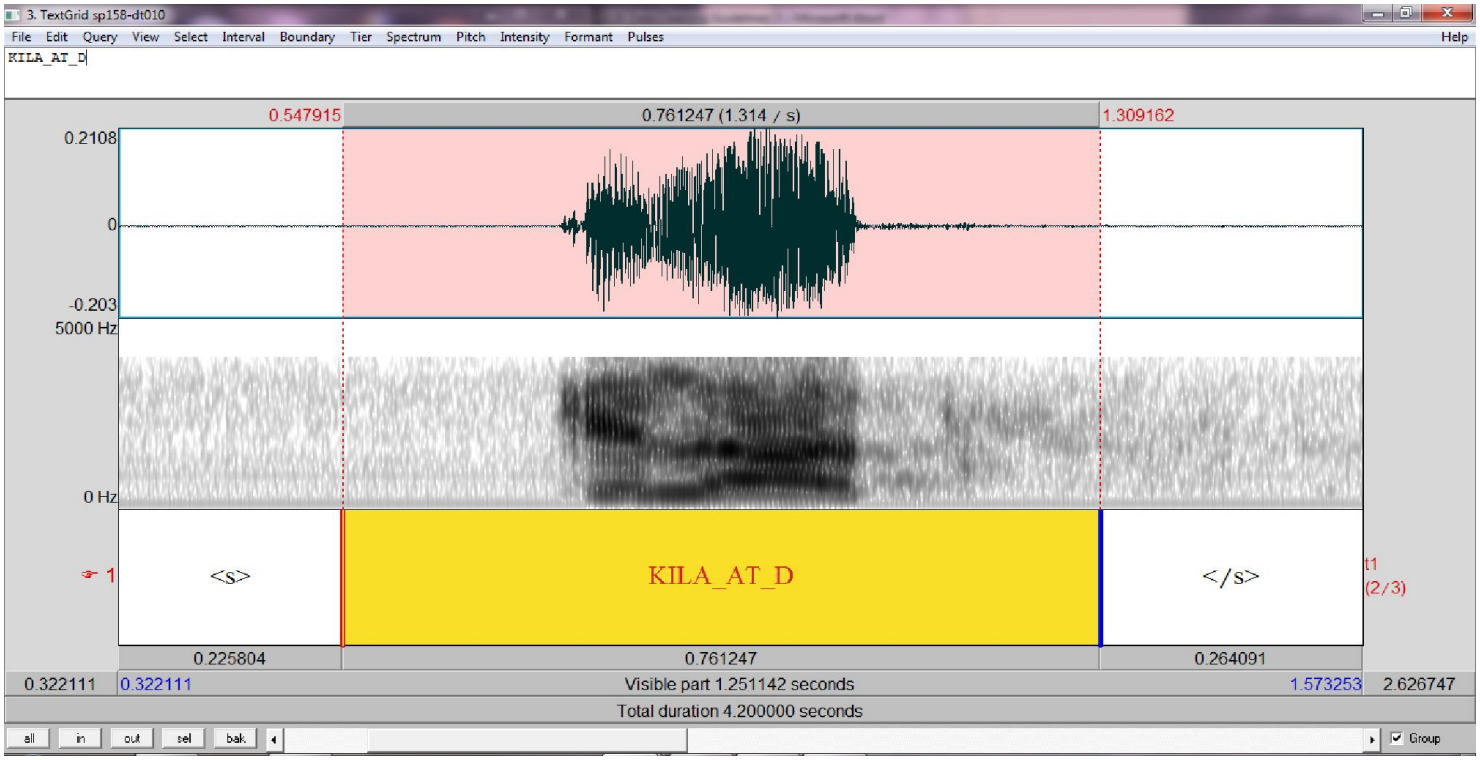
- Initial Position:

  E.g. AST_DO_OR into IST_DO_OR.



- Middle Position:

  E.g. GILGIT_D into GILGAT_D , KALA_AT_D into KILA_AT_D



- Final Position:
  - Long into short or vice versa
  - E.g KOTLI into KOTLI_I

c. Deletion of a vowel:

Initial Position:

Middle Position:

It will be judged on the basis of general trend.

E.g. MALAKAND into MLA_AKAND

BATAGRA_AM into BATGRA_AM

Final Position:

E.g. SHEIKHUPUR instead of SHEIKHUPURA.

Thank You!