SYNTHESIS OF /ikis/ (21) And /bais/ (22)

Muhammad Mustafa Rafique

ABSTRACT

This paper discusses the entire process which was used for the synthesis of words /ikis/ and /bais/. A high level synthesizer HLSyn was used for this synthesis. A perceptual experiment was done to evaluate the perceptual quality of the synthesized words. Comparison of spectrograms of synthesized and recorded /ikis/ and /bais/ was also done.

1. INTRODUCTION

Urdu is the official language of Pakistan. It is also spoken in South Asian countries like India, Bangladesh and United Arab Emirates. It is phonetically similar to Hindi, but it has different script and historical characteristics. Urdu is regarded as an offspring of Persian (Saleem et al, 2002).

The words /ikis/ and /bais/ are used for the twenty-first and twenty-second digit in Urdu language. The pronunciation of words /ikis/ and /bais/ varies from region to region due to differences in accents.

This paper describes all the efforts which were made to synthesize /ikis/ and /bais/ using a high level synthesizer HLSyn.

2. LITERATURE REVIEW

Manner of airflow differs between different sounds. Phonemes in which airflow completely stops in the oral or nasal cavity are referred to as *stops* (or *plosives*), e.g. /t/ is a stop because airflow is completely stopped in the oral cavity while articulating /t/. Similarly, /s/ has the feature [+continuant] because airflow does not break in the production of /s/ (Napoli, 1996, p. 18). All vowels are [+continuant]. The phonemes, which have the manner [+continuant] but involve air turbulence, are called *fricatives or spirants*. /s/, therefore has the manner [+fricative] (Napoli, 1996, p. 19).

There are at least three different audible characteristics of language sounds: pitch, loudness (or intensity), and quality. Both pitch and loudness are referred to as supra-segmental features. In some languages (e.g. English), only quality is required for the identification of a phoneme, i.e. only by keeping the quality of a sound steady and changing its pitch and intensity does not change the identity of a particular phoneme. But both the pitch and the intensity of a sound are important and when it is desired to put sounds together to make words; special attention is paid on these supra-segmental features (Napoli, 1996, p. 41).

The obvious way to provide speech output from computers is to select the basic acoustic units to be used, record them and generate utterances by concatenating together appropriate segments from this pre-stored inventory of speech units. But it is really very hard to decide the basic units. These basic units could be whole sentences, words, syllables or phonemes.

There are several tradeoffs to be considered while selecting the appropriate unit. The larger the units, the more utterances have to be stored. It is not the length of individual utterances that is of concern, but it is their variety—it tends to increase exponentially instead of linearly with the size of the basic unit. Number provides an easy example of this complexity. There are 10^7 7-digit telephone numbers and it is certainly unfeasible to record each one individually. But it is very easy to record each digit separately and then concatenate these digits to produce any 7-digit telephone number. Since data storage technology is improving day by day, the limitation is shifting more towards the recording factor. At a PCM (Pulse Code Modulation) data rate of 50Kbit/s, a 100 Mbytes disk can hold over 4 hours of continuous speech. With linear predictive coding at 1Kbit/s, it holds 800,000 (222.22 hrs) seconds of continuous speech.

A word seems to be a reasonable size unit. Many applications use a limited vocabulary. An airline reservation system developed in Bell Telephone Laboratories® had a vocabulary of 200 words. Even at PCM data rates of 1 Bit/s, this will consume less than 0.5 Byte of storage. Unfortunately, prosodic and co articulation factors do come into play if word is selected as basic unit (Witten, 1982, p. 153).

Real speech is connected although there are few gaps between words. Co articulation, where sounds are affected by these gaps on either side naturally operates across word boundaries. The time constants of co articulation are associated with the mechanics of the vocal tract (Witten, 1982, p. 153).

Prosodic features, especially pitch and rhythm (stress), span much longer stretches of speech than single words. As far as most speech output applications are concerned, they operate at the utterance level of a single, sentence sized, information unit, which cannot be catered if speech is produced using word as storage unit, since it is almost impossible to alter the fundamental frequency or duration of a time waveform without changing all the formant resonances (Witten, 1982, p. 153).

Historically speech synthesizers fall into two broad categories. One are *articulatory synthesizers* that attempt to model faithfully the mechanical motions of the articulators and the resulting distributions of volume, velocity and sound pressure in the lungs, larynx, vocal tract and nasal tract. The second category of synthesizers is *formant synthesizers* which attempts to approximate directly the speech waveform and spectrum by a simpler model formulated in the acoustic domain (Pickett, 1999, p. 325, Kent, 1992, p. 179)

The HLSyn (High Level Speech Synthesizer) provides a graphical environment for creating and comparing speech files using high level speech synthesis. The HLSyn program also supports formant synthesis using underlying SenSyn® Klatt-type cascade parallel formant synthesizer KLSYN88 (HLSyn High-Level Speech Synthesizer, p. 3, Chapter 1, p.1).

3. METHODOLOGY

Subject for the experiment was a native male speaker of Urdu language. The subject was judged to have a very clear accent of Urdu language. The subject was asked to say /ikis/ in a natural way, which was then recorded using XWaves® utilities. From this recording the average time duration of /ikis/ was noted. The average of the fundamental frequency (f0) of the subject during the production of /ikis/ was also calculated. The values of first four formant frequencies (f1, f2, f3 & f4) of the vocal tract of the subject were also noted after every 0.01 second interval.

HLSyn was selected to synthesize /ikis/. HLSyn requires ten parameters (referred as HL parameters) to synthesize the sound for the time duration of ten milliseconds. The ten HL parameters, with their brief descriptions are:

f0: Fundamental frequency of vocal folds vibration.

f1, f2, f3, and f4: First four natural frequencies of the vocal tract. These are the natural frequencies when the velo-pharyngeal port closed, there is no acoustic coupling to the trachea, and no local constriction is formed near the front of the vocal tract by the lips or the tongue blade.

ag: Area of glottal opening. Range is usually 0 - 40 mm^2. Average opening for modal voicing is usually about 3 – 5 mm^2.

al: Cross sectional area of constriction formed by the lips during the production of labial consonants. A value of 100 mm^2 corresponds to the non-constriction configuration.

ab: Cross sectional area of constriction formed by the blade of the tongue during the production of coronal consonants. A value of 100 mm^2 corresponds to the non-constriction configuration.

an: Cross sectional area of velo-pharyngeal port. Range is 0 - 100 mm^2.

ue: Rate of increase of vocal tract volume that is actively controlled during the constricted interval for an obstruent consonant. Positive values of **ue** correspond to an active expansion of the cavity behind the consonant constriction and negative values correspond to a contraction. The integral of **ue** over the constricted interval is the total increase or decrease in volume.

The values of f0, f1, f2, f3 and f4 from the above experiments were used as inputs for five of the ten parameters of HLSyn. The values of the remaining parameters (**ag**, **al**, **ab**, **an** and **ue**) were adjusted for each consonant under consideration. There are three additional HL parameters in HLSyn, **ps**, **dc** and **ap**, which were not disturbed and were left at their default values.

At the end, the result of the synthesized /ikis/ was compared with the recorded /ikis/. This time, Praat® speech analyzer was used to record the actual utterance of /ikis/. The result of synthesized and actual /ikis/ was compared by comparing the spectrogram of each utterance.

The same methodology was repeated for the synthesis of /bais/. The same subject was used for both experiments.

At the end, a perceptual experiment was conducted to check the perceived quality of the synthesized words. A group of ten native speakers of Urdu language were selected for the experiment. The group was asked to listen the synthesized as well as recorded words and then grade them on the scale from one to ten. The synthesized words were jumbled up with the recorded ones and no distinction was made apparent for the synthesized and recorded words.

4. RESULTS

By adopting the above mentioned methodology, /ikis/ and /baɪs/ were synthesized and the results were noted. The spectrograms of the synthesized /ikis/ is shown in Figure 1 and the spectrogram of the recorded one is shown in Figure 2. Similarly, spectrogram of synthesized /bais/ is shown in Figure 3 and the spectrogram of recorded /bais/ is shown in Figure 4. In all the spectrogram, a small silence is shown before actual words.

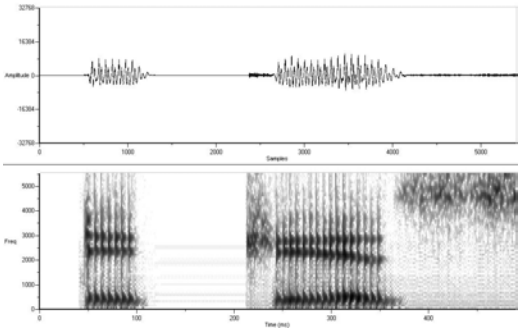

FIGURE 1 Spectrogram of synthesized /ikis/

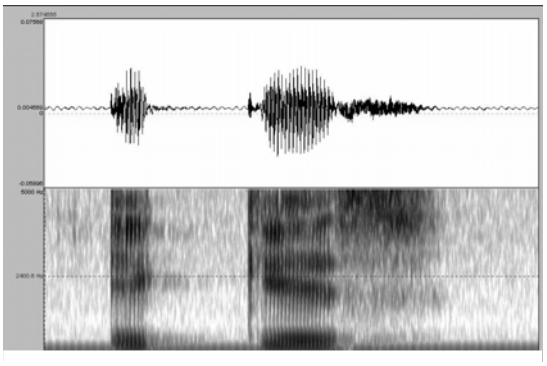

FIGURE 2 Spectrogram of recorded /ikis/

The HL parameter list for the synthesis of /ikis/ is displayed in Table 1 in the Appendix section and for the synthesis of /bais/ is displayed in Table 2 in the Appendix section.

The result of the perceptual experiment shows the quality of the synthesized words relative to the actual words. As described earlier a group of ten native speakers of Urdu language was used to judge the quality of the synthesized words. The average points awarded by the listeners to the

synthesized /ikis/ was 8/10 relative to the recorded /ikis/ which on average got 8.7/10. Similarly, the average score for the synthesized /bais/ was 7.9/10 relative to the recorded /bais/ which got 9.3/10 points.

The data on which the above results are

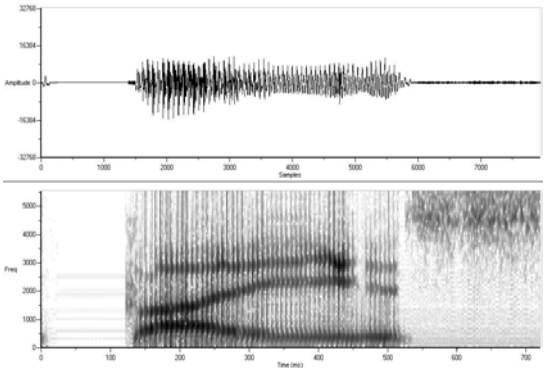

FIGURE 3: Spectrogram of synthesized /bais/

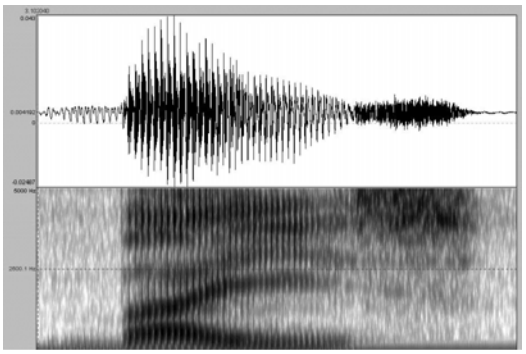

FIGURE 4: Spectrogram of recorded /bais/

based is not important in this experiment and so is not produced here.

5. DISCUSSION

The word /ikis/ consist of four phonemes, out of which two are vowels /i/, one is a velar stop /k/ and the last is a fricative /s/. Similarly, word /bais/ is composed of four phonemes, two of them are vowels /a/ and /i/, one is bilabial stop /b/ and the last is again a fricative /s/.

The time duration for the utterance of any word may change, even if same speaker speaks it at different time. Similarly, duration of any phoneme does not remain constant if they are produced at different times. One phoneme may be stressed for a given utterance in a word, while another phoneme may get the stress in the same word if they are not produced at the same time.

The spectrogram shown in Figure 1 and Figure 2 are almost similar in pattern. As mentioned in the result section, a small silence is shown before each word in all the spectrogram, so the words in the spectrograms of Figure 1 and 2 start with a vowel /i/, then they have a consonant /k/, which is followed by the vowel /i/ and at the end they have the consonant /s/.

In Figure 1 and Figure 2, first vowel in the word /ikis/ does not have the same time duration. This is because the word /ikis/ was not spoken with the same stress pattern by the speaker. Despite of the time duration, the vowel follows the same pattern in both the spectrograms. Since it is occurring before a velar consonant /k/, so it does not show any deviation in its formants and they remain constant through out the vowel. The consonant /k/ in the word /ikis/ shows the same pattern in both the spectrogram but it varies in time duration. It has a complete closure after the vowel, which is followed by a burst. The consonant /k/ is followed by the same vowel /i/ which follows the same formant pattern in both the spectrograms of Figure 1 and 2. At the end, the noise at higher frequencies are shown in both the spectrograms which shows the presence of a fricative /s/.

As mentioned in the methodology section, the three additional HL parameters, **ps**, **dc** and **ap** were kept on their default values. The rest of the HL parameters namely **ag**, **al**, **ab**, **an** and **ue**, which are used to produce constriction in the vocal tract were adjusted according to the need in the synthesis of /ikis/ as shown in Table 1 in the Appendix section.

The value for the parameter **ag** was kept constant for the synthesis of /i/, since area of glottal opening remains constant during the production of any vowel. But value of parameter **ag** was adjusted for synthesizing

/k/ because glottal area increases as the phoneme /k/ is uttered. Its value is brought back to normal for the production of second /i/. Since glottal area remains constant for the production of /s/, so again the value of **ag** was kept constant during the synthesis of /s/.

As there is no constriction at lips in the production of /ikis/, so there is no change in the value of **al**, and it was kept at its default value through out the synthesis of /ikis/. The values for **an** and **ue** were also kept constant since there is no nasalization and no increase in the volume of vocal tract.

The value of the parameter **ab** is adjusted twice for the synthesis of /ikis/, one for the synthesis of phoneme /k/ and the other for the synthesis of phoneme /s/. As there is no HL parameter which deals with the constriction made from the back of the tongue, so the parameter **ab** is used to produce a constriction for the synthesis of phoneme /k/ along with the parameter **ag**. Since /s/ is produced with the slight constriction made by the blade of the tongue, so the value of **ab** is adjusted for the synthesis of /s/.

Figure 3 and 4 shows the spectrograms of the word /bais/. Very fortunately, the same stress pattern is observed in both the spectrograms. In both the spectrograms, there is a complete closure before the burst of phoneme /b/, which is followed by a diphthong /ai/. The formants of the diphthong /ai/ are almost identical to each other in both the spectrograms. After the diphthong, there is the same noise as in the case of last phoneme in the word /ikis/, which shows the presence of a fricative /s/.

The HL parameter list used for the synthesis of the word /bais/ is shown in Table 2 in the Appendix section. Since area of glottal opening remains constant for the production of /bais/, so the parameter **ag** was kept at its default value. Since there is a complete constriction at lips for the production of /b/, the value of **al** for the synthesis of /b/ was set to the maximum constriction i.e. 0. Its value was again changed to 'no constriction' value after the phoneme /b/ and was kept constant for the rest of the word. During the synthesis of /bais/, constriction from the blade of the tongue is made only for the production of phoneme /s/, so value of the parameter **ab** was only adjusted for the synthesis of phoneme /s/. Since there is no nasalization in the production of word /bais/, so the HL parameter **an** was set to 0 value. Similarly, there is no need to change the value of **ue** and it was also kept on its default value.

The results of perceptual experiment are quite satisfactory. HLSyn is a good synthesizer that provides user-friendly interface to synthesized words. No major problem was encountered during the synthesis of /ikis/ and /bais/.

6. REFERENCES

HLSyn High-Level Speech Synthesizer. User Interface Manual.

Kent, Ray D., Charles Read.1992. *The Acoustic Analysis Of Speech*. Singular Publishing Group, Inc. California.

Napoli, Donna Jo. 1996. *Linguistics An Introduction*. Oxford University Press

Pickett, J. M. 1999. *The Acoustics Of* Speech Communication Fundamentals, Speech Perception Theory, And Technology. USA

Saleem, Abdul Mannan; Kabir, Hasan; Riaz, Muhammad Khurram; Rafique, Muhammad Mustafa; Khalid, Nauman; Shahid, Syed Raza. 2002. *Urdu Consonantal and Vocalic Sounds*. Center for research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, Pakistan.

Witten, I.H, 1982. *Principles of Computer Speech*. Academic Press Inc. (London) Ltd.

7. APPENDIX

TABLE 1 HL Parameters for /ikis/

Time	Ag	al	ab	an	ue	f0	F1	F2	f3	f4	ps	dc	ap
0	4	100	100	0	0	0	0	0	0	0	8	0	0
10	4	100	100	0	0	0	0	0	0	0	8	0	0
20	4	100	100	0	0	0	0	0	0	0	8	0	0
30	4	100	100	0	0	0	0	0	0	0	8	0	0
40	4	100	100	0	0	0	0	0	0	0	8	0	0
50	4	100	100	0	0	1360	378.0	2256	2897	3657	8	0	0
60	4	100	100	0	0	1360	355.0	2299	2884	3743	8	0	0
100	4	100	0	0	0	1360	384.3	2255	2802	3403	8	0	0
110	6	100	0	0	0	1360	391.7	2244	2782	3318	8	0	0
180	14	100	0	0	0	1360	254.0	2200	2700	3652	8	0	0
190	15	100	0	0	0	1360	260.0	2200	2700	3725	8	0	0
200	16	100	0	0	0	1360	233.5	2200	2700	3667	8	0	0
210	18	100	0	0	0	1360	248.0	2200	2700	3604	8	0	0
220	20	100	100	0	0	1360	294.0	2200	2700	3658	8	0	0
230	22	100	100	0	0	1360	389.0	2200	2700	3421	8	0	0
240	4	100	100	0	0	1360	372.0	2200	2707	3435	8	0	0
250	4	100	100	0	0	1360	250.0	2255	2714	3435	8	0	0
260	4	100	100	0	0	1360	322.0	2255	2720	3634	8	0	0
270	4	100	100	0	0	1360	327.0	2229	2727	3629	8	0	0
280	4	100	100	0	0	1360	338.0	2234	2734	3633	8	0	0
290	4	100	100	0	0	1360	352.0	2172	2740	3544	8	0	0
300	4	100	100	0	0	1360	378.0	2122	2747	3517	8	0	0
310	4	100	100	0	0	1360	408.9	2122	2754	3517	8	0	0
320	4	100	100	0	0	1360	428.0	2063	2761	3606	8	0	0
330	4	100	100	0	0	1360	350.0	1979	2768	3600	8	0	0
340	4	100	100	0	0	1360	350.0	1926	2774	3551	8	0	0
350	4	100	0	0	0	1360	350.0	1949	2781	3698	8	0	0
360	4	100	1	0	0	1360	350.0	1915	2789	3788	8	0	0
370	4	100	2	0	0	1360	350.0	1906	2930	3759	8	0	0
380	4	100	3	0	0	1360	350.0	1947	2740	3470	8	0	0
390	4	100	4	0	0	1360	350.0	2012	2777	3498	8	0	0
400	4	100	5	0	0	1360	350.0	2046	2757	3508	8	0	0
410	4	100	6	0	0	1360	350.0	2050	2876	3444	8	0	0
420	4	100	7	0	0	1360	350.0	2010	2897	3537	8	0	0
430	4	100	0	0	0	1360	350.0	1949	2781	3698	8	0	0
440	4	100	1	0	0	1360	350.0	1915	2789	3788	8	0	0
450	4	100	2	0	0	1360	350.0	1906	2930	3759	8	0	0
460	4	100	3	0	0	1360	350.0	1947	2740	3470	8	0	0
470	4	100	4	0	0	1360	350.0	2012	2777	3498	8	0	0
480	4	100	5	0	0	1360	350.0	2046	2757	3508	8	0	0
490	4	100	6	0	0	1360	350.0	2050	2876	3444	8	0	0

TABLE 2 HL Parameters for /bais/

Time	ag	al	ab	an	ue	f0	f1	f2	f3	f4	ps	dc	ap
0	4	0	100	0	0	1360	153.0	1399	2890	3475	8	0	0
10	4	0	100	0	0	1360	225.0	1345	2547	3441	8	0	0
20	4	0	100	0	0	1360	241.0	1285	2521	3345	8	0	0
30	4	0	100	0	0	1360	245.0	1326	2552	3292	8	0	0
40	4	0	100	0	0	1360	231.0	1248	2484	3298	8	0	0
50	4	0	100	0	0	1360	224.0	1235	2503	3310	8	0	0
60	4	0	100	0	0	1360	217.0	1253	2488	3237	8	0	0
70	4	0	100	0	0	1360	207.0	1526	2459	3201	8	0	0
80	4	0	100	0	0	1360	204.0	1688	2595	3241	8	0	0
90	4	0	100	0	0	1360	210.0	1442	2734	3102	8	0	0
100	4	0	100	0	0	1360	205.0	1583	2429	3259	8	0	0
110	4	0	100	0	0	1360	213.0	1475	2439	3046	8	0	0
120	4	0	100	0	0	1360	215.0	1407	2614	3329	8	0	0
130	4	100	100	0	0	1360	201.0	1614	2588	3370	8	0	0
140	4	100	100	0	0	1360	474.0	1157	2621	3321	8	0	0
150	4	100	100	0	0	1360	550.0	1203	2445	3435	8	0	0
160	4	100	100	0	0	1360	618.0	1223	2445	3383	8	0	0
170	4	100	100	0	0	1360	672.0	1242	2780	3283	8	0	0
180	4	100	100	0	0	1360	685.0	1270	2732	3249	8	0	0
190	4	100	100	0	0	1360	700.0	1309	2728	3529	8	0	0
200	4	100	100	0	0	1360	711.0	1348	2720	3306	8	0	0
210	4	100	100	0	0	1360	712.0	1365	2734	3325	8	0	0
220	4	100	100	0	0	1360	708.0	1408	2728	3368	8	0	0
230	4	100	100	0	0	1360	684.0	1491	2747	3398	8	0	0
240	4	100	100	0	0	1360	644.0	1590	2769	3372	8	0	0
250	4	100	100	0	0	1360	615.0	1701	2756	3308	8	0	0
260	4	100	100	0	0	1360	562.0	1758	2897	3400	8	0	0
270	4	100	100	0	0	1360	529.0	1800	2751	3311	8	0	0
280	4	100	100	0	0	1360	474.0	1872	2773	3356	8	0	0
330	4	100	100	0	0	1360	348.0	2233	2921	3583	8	0	0
370	4	100	100	0	0	1360	343.0	2254	2967	3620	8	0	0
380	4	100	100	0	0	1360	319.0	2282	3030	3590	8	0	0
390	4	100	100	0	0	1360	315.0	2266	3055	3575	8	0	0
400	4	100	100	0	0	1360	317.0	2254	3044	3407	8	0	0
410	4	100	100	0	0	1360	294.0	2296	3060	3316	8	0	0
420	4	100	100	0	0	1360	291.0	2309	2918	3075	8	0	0
430	4	100	100	0	0	1360	319.0	2270	2816	3083	8	0	0
440	4	100	100	0	0	1360	265.0	2234	2908	3578	8	0	0
450	4	100	100	0	0	1360	218.0	2329	2918	3385	8	0	0
460	4	100	100	0	0	1360	236.0	1818	2900	3652	8	0	0
470	4	100	100	0	0	1360	269.0	2057	2807	3580	8	0	0
480	4	100	100	0	0	1360	262.0	2019	2804	3467	8	0	0
490	4	100	100	0	0	1360	270.0	1940	2760	3535	8	0	0
500	4	100	100	0	0	1360	350.0	1926	2745	3551	8	0	0
510	4	100	0	0	0	1360	350.0	1949	2781	3698	8	0	0
520	4	100	1	0	0	1360	350.0	1915	2789	3788	8	0	0
530	4	100	2	0	0	1360	350.0	1906	2930	3759	8	0	0
540	4	100	3	0	0	1360	350.0	1947	2740	3470	8	0	0
550	4	100	4	0	0	1360	350.0	2012	2777	3498	8	0	0
560	4	100	5	0	0	1360	350.0	2046	2757	3508	8	0	0
570	4	100	6	0	0	1360	350.0	2050	2876	3444	8	0	0

580	4	100	7	0	0	1360	350.0	2010	2797	3537	8	0	0
590	4	100	0	0	0	1360	350.0	1949	2781	3698	8	0	0
600	4	100	1	0	0	1360	350.0	1915	2789	3788	8	0	0
610	4	100	2	0	0	1360	350.0	1906	2930	3759	8	0	0
620	4	100	3	0	0	1360	350.0	1947	2740	3470	8	0	0
630	4	100	4	0	0	1360	350.0	2012	2777	3498	8	0	0
640	4	100	5	0	0	1360	350.0	2046	2757	3508	8	0	0
650	4	100	6	0	0	1360	350.0	2050	2876	3444	8	0	0
720	4	100	6	0	0	1360	350.0	2050	2876	3444	8	0	0