SYNTHESIS OF ORAL AND NASAL VOWELS OF URDU

MUHAMMAD KHURRAM RIAZ

ABSTRACT

The following oral and nasal vowels of Urdu were synthesized (i, æ, u, ɑ, ĩ, æ̃, ũ, ɑ̃). HLSyn a High level Synthesizer was used to synthesize the vowels. After synthesizing a perceptual experiment was performed to find out how successfully the synthesized vowels were perceived, moreover the spectrograms of synthesized vowels were compared with that of original vowels recorded. For nasal vowels bandwidths were also compared.

1. INTRODUCTION

Urdu, the National language of Pakistan is widely spoken in different parts of Pakistan and India. Urdu is among those languages that use nasalized vowels to distinguish within pair of words that are otherwise the same.

This paper deals with the synthesis of oral and nasal vowels of Urdu. All the methodology, which was adopted to perform experiment, is described in this paper. An overview of the synthesizer used is also given and all its parameters have been discussed in detail. In the end all the results are discussed and the success of the experiment is concluded.

2. LITERATURE REVIEW AND PROBLEM STATEMENT

Kachru (1990, p. 54) and Khan (1997) say that Urdu has seven long (i, e, ɛ, u, o, ɔ, ɑ) and three short (ɪ, ʊ, ə) oral vowels. Bokhari (1985, p. 6) agrees with Kachru and Alam on long oral vowels but gives seven short oral vowels. All the shorter vowels given by Bokhari are shorter version of longer ones.

Khan (1997) lists ten nasalized vowels containing seven long (ĩ, ẽ, ɛ̃, ũ, õ, ɔ̃, ɑ̃) and three short (ɪ̃, ʊ̃, ə̃) nasalized vowels. Bokhari (1985, p. 6) gives five long (ĩ, ẽ, ũ ,ɔ̃, ɑ̃) and five short nasal vowels. All the shorter nasal vowels given by Bokhari are shorter version of longer ones. Kachru (1990) has not listed any nasalized vowel, but mentions that nasalization is distinctive.

Oral vowels are produced with velum raised thus closing the velo-pharyngeal port. On the other hand nasal vowel are produced with opened velo-pharyngeal port. According to Pickett (1999, p. 70) following important acoustic effects of nasalization have been determined through research on vocal tract models. One effect was that the first formant became lower and broader than before, because of damping of the formant resonance by the loss of energy by opening of nasal tract. Another change due to opening of velo-pharyngeal port is introduction of zeros or anti resonant. The anti resonant is opposite to resonances in their effect on spectrum. They selectively absorb sound that results in reduction of the amplitudes of components near the anti resonant frequency. Further a zero also amplifies components that are sufficiently above the anti resonant frequency. Thus for each zero there is an extra pole introduced.

Stevens (2000, p. 316) gives a general observation about the nasal vowels that they have almost flat spectrum at low frequencies (up to, say, 1200 Hz) the reasons behind this are: 1) widening of the bandwidth of F1 (and, for back vowels F2); 2) introduction of an additional resonance that prevents any one low frequency resonance from being dominant.

Kent and Read (1992, p. 166) give the following effects on spectrograms of vowels after nasalization.

(1) Increase in formant bandwidth, so that formant energy appears broader.
(2) Decrease in the overall energy of the vowel (compared to non-nasal vowels).
(3) Introduction of a low-frequency nasal formant with a center

frequency of about 250-500 Hz for adult males,
(4) A slight increase of the F1 frequency and a slight lowering of the F2 and F3 frequencies.
(5) Presence of one or more anti-formants.

Synthesis is a process of producing speech, which is as near human-like as possible and which is addressed to a human being.

The problem was to synthesize oral and nasal vowels of Urdu using a high level synthesizer so that the usefulness of the synthesizer can be found and it can be used for further synthesis in Urdu language. Another purpose of this paper was to learn some basics of synthesis. The vowels synthesized are [i, æ, u, ɑ, ĩ, æ̃, ũ, ɑ̃].

3. METHODOLOGY

The synthesizer used in this paper is known as HLSyn. HLSyn is a high-level speech synthesizer that provides an integrated graphical environment for specifying, creating, analyzing and comparing synthetic speech files using high-level synthesis. The HLSyn also supports formant synthesis using the underlying SenSyn Klatt-type cascade-parallel formant synthesizer (HLSyn, 1997, p. 3).

The approach of HLSyn is based on the observation that there are constraints on the 40-odd parameters that are available to Control KLSYN88 (Klatt Synthesizer).

These constraints exist because physical process of speech production imposes limits on the combination of synthesis parameters that can exist at a particular time or on the ways in which these parameters can change with time. To account for these constraints, a set of about 10 higher-level (HL) parameters has been proposed. These HL parameters are more closely related to the actual states and movements of the vocal tract than are the lower-level (KL) parameters. A set of mapping relations within HLSyn transforms the HL parameters into the KL parameters that actually control KLSYN88.

The proposed set of HL parameters is listed in the table below (HLSyn, 1997).

TABLE 1 Details of HLSyn Parameters (HLSyn, 1997).

Parameter	Description
f1, f2, f3, f4	First four natural frequencies of the vocal tract. These are the natural frequencies. When the velo-pharyngeal port is 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 by the tongue blade. For non-nasal vowels with a glottal configuration appropriate for modal voicing, these natural frequencies are identical to the formant frequencies.
f0	Fundamental frequency of vocal-fold vibration. This HL parameter is usually identical to KL (Klatt) parameter f0.
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-section area of constriction

	formed by the tongue blade during the production of coronal consonants. A value of 100mm^2 corresponds to the non-constriction configuration.
An	Cross-sectional area of velopharyngeal 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 constriction. The integral of ue over the constricted interval is the total increase or decrease in volume.

Five of these HL Parameters are similar to (and often identical to) the KL parameters. These are the fundamental frequency f0 and the four formant frequencies (f1, f2, f3 and f4) that specify the natural frequencies of the vocal tract assuming no acoustic coupling to the trachea or to the nasal cavity, and assuming that there is no localized constriction formed by the tongue blade or the lips. The time-varying HL formant-frequency parameters in effect specify how the shape of the vocal tract changes with time independent of any nasal or tracheal coupling or local constriction. If there is nasal or tracheal coupling or a local constriction as specified by an, ag, al and ab, then the mapping relations may cause some modification of the HL formant parameters (primarily the first formant) to yield the actual KL parameters (F1, F2 etc) that are used to control the synthesizer components. In effect, the HL parameters f1, f2, f3 and f4 describe the aspects of vocal-tract shape that are determined by tongue-body position, jaw position, pharyngeal shape, and possible lip rounding (HLSyn, 1997).

The very first step of the experiment was to record selected oral and nasal vowels of Urdu. For this purpose a native speaker was selected and each vowel was recorded five times using a high quality recording system. Out of five samples of each sound recorded one best sample was selected for analysis.

After completion of recording data was collected from the sounds recorded. For this purpose software named Praat was used. Praat is freely downloadable software available at http://www.praat.org. All the formants, F0s, bandwidths and spectrograms were found using Praat. Praat gives option to find mean values of formants so mean formants were easily found using this facility. However bandwidths were averaged manually by taking values at different points in time. The purpose of finding bandwidths was not to give them as input to synthesizer but to use them for comparison with that of synthesized vowel bandwidths during discussion. The averaged bandwidth values of nasal recorded and synthesized vowels are given in the results section.

In order to synthesize vowels the synthesizer needed fundamental frequency, four natural frequencies of oral tract and other parameters of HLSyn (ag, al, ab, an, ue). According to HLSyn documentation (HLSyn, 1997) oral vowels have natural frequencies equal to formants so the formants found from recorded oral vowels were given to HLSyn as f1, f2, f3 and f4. f0 was given as found from recorded vowel, ag was set to 4.2 which is among the preferred values given by HLSyn documentation, other parameters (al, ab, an, ue) according to Table 1 play no important role in synthesizing vowels so they were set to default value of 100. The duration of each vowel was set to about 600 milli-seconds. The parameter ag was set to zero at start so that voicing may increase gradually and give a natural effect.

To synthesize nasal vowels the parameter an was used to control nasalization. Since oral and nasal vowels have same oral tract formation, the only difference is the opening of velo-pharyngeal port. Thus the natural frequencies for the nasal vowels were equal to their corresponding non-nasal vowels. f0 was given as found from recorded sound. After giving natural frequencies and f0 the parameter an was increased to such a value that could make vowels nasalized, mostly a value of 40 or 50 was used. The duration was also set to 600 ms and the parameter ag was set to zero in start.

When all the vowels were synthesized they were converted into wave files and were imported into Praat so that their spectrograms can be obtained. Although HLSyn also provides spectrograms of synthesized sounds but their quality is not good that is why Praat was used. After obtaining spectrograms of both synthesized and recorded vowels they were compared with each other to test the quality of synthesis.

Another test was designed to find out how listeners perceive synthesized vowels. All the eight sounds synthesized were converted into wave files and five sets were made using block randomizing. Each set contained every wave file of vowels. Each set was played in front of five listeners and each listener was asked to write what they heard. Based on this test a confusion matrix was made which shows what was played and what was perceived.

4. RESULTS

4.1. Confusion Matrix

Below are the results obtained after conducting the experiment described above.

The rows contain the vowels that were played in front of the speakers and the columns contain what was really perceived by the listeners. To read the matrix just go to a box find the number written in it say X then find the row vowel say A and then column vowel say B for the selected box. After getting X, A, and B you can say that X times A was perceived as B out of total 25. The last column named as miscellaneous contains count of those synthesized sounds that were not perceived among listed nasal or oral vowels.

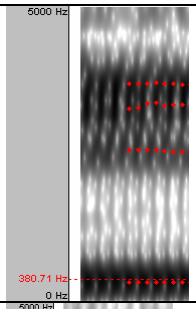
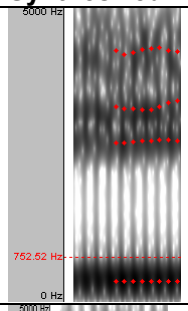
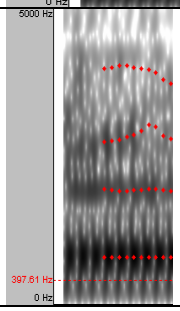
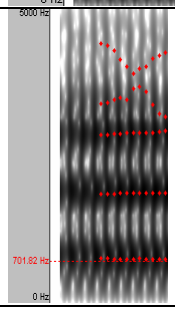
TABLE 2 Confusion Matrix.

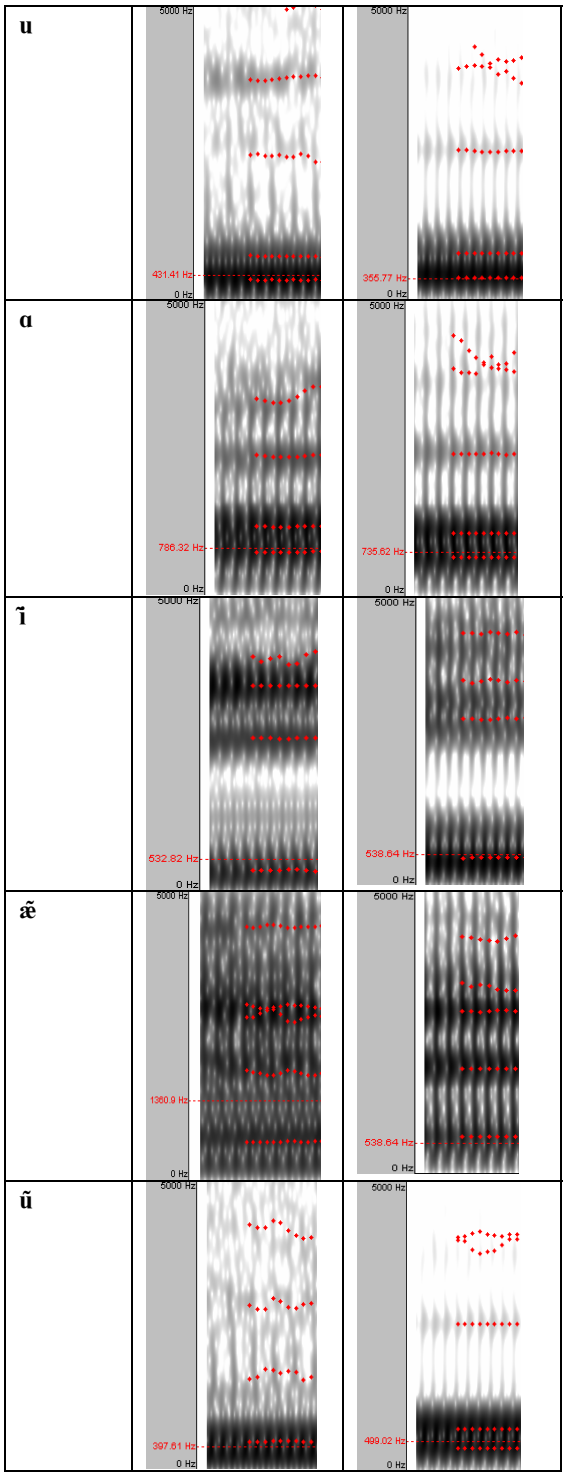
		P	E	R	C	E	I	V	E	D
		i	ĩ	æ	æ̃	u	ũ	ɑ	ɑ̃	Misc
	i	22	3							0
P	ĩ	4	21							0
L	æ			24						1
A	æ̃			1	24					0
Y	u					25				0
E	ũ					15	6			4
D	ɑ							25		0
	ɑ̃						1	3	16	5

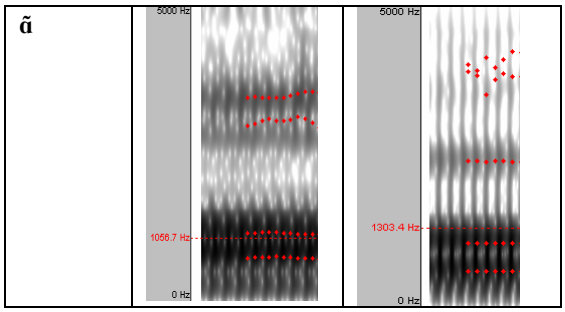
4.2. Spectrograms

The spectrograms of the synthesized and recorded sounds are shown below. The dotted lines in the spectrograms are the formants of the vowels. Frequency range of the spectrograms is also shown against each spectrogram so that the formant values can be realized.

TABLE 3 Spectrograms of Recorded and Synthesized vowels.

Vowel	Recorded	Synthesized
i		
æ		





4.3. Bandwidths of Nasal vowel

Spectrograms are good for viewing the values of formants. But for nasals, bandwidth is also important which cannot be easily measured from the spectrograms. That is why the bandwidths of the synthesized and recorded vowels are provided in this part. These values were found using software Praat.

TABLE 4 Bandwidths of Recorded and Synthesized vowels.

Vowel		Recorded	Synthesized
ĩ	**BWF1**	250 Hz	220 Hz
	BWF2	260 Hz	212 Hz
	BWF3	150 Hz	500 Hz
æ̃	**BWF1**	225 Hz	170 Hz
	BWF2	310 Hz	180 Hz
	BWF3	320 Hz	310 Hz
ũ	**BWF1**	60 Hz	40 Hz
	BWF2	1200 Hz	13 Hz
	BWF3	410 Hz	390 Hz
ɑ̃	**BWF1**	190 Hz	110 Hz
	BWF2	145 Hz	70 Hz
	BWF3	800 Hz	250 Hz

5. 5. DISCUSSION

5.1. Oral Vowels

After conducting this experiment it was realized that oral vowels are easy to synthesize as compared to nasal vowels. More over the results of oral vowels were better than nasal vowels. On viewing the confusion matrix (Table 2) we can see that the speakers perceived almost all of them correctly, even two of them were perceived 100 percent. The spectrograms (Table 3) of the recorded and synthesized vowels are also almost similar. The dotted lines in the spectrograms are showing the formants of

the vowels. The formant values of the recorded and synthesized vowels are almost same that is why all the oral vowels were perceived well.

5.2. Nasal Vowels

Nasal vowels were little hard to synthesize as compared to oral vowels. The reason behind this is, controlling of bandwidth. Among the four vowels the two front vowels [ĩ] and [æ̃] were synthesized good, [ɑ̃] was good and [ũ] was the worst.

On analyzing the confusion matrix (Table 2) it was found that [ĩ] was mostly perceived correctly but sometimes it was confused with [i]. On analyzing the spectrograms (Table 3) the formants found were almost equal to that of original recoded vowel. The bandwidths of first and second formants were reasonably controlled and were near to that of recorded. However the third bandwidth was not properly adjusted that is why there was little confusion in perceiving.

[æ̃] showed excellent results, it was only confused once. The formants of synthesized and recorded [æ̃] were almost equal. The values for BWF1 and BWF3 (from Table 4) were almost equal to that of recorded vowel however there was some difference in BWF2 but still It was well perceived.

[ɑ̃] was satisfactory because of problems in third formant and its bandwidth. The value as well as the bandwidth of F3 was very low from original which effected perception of [ɑ̃].

[ũ] was the worst among all nasal vowels. Among all the formants only F2 caused the problem. It's value was lower than that of original sound recorded. Another big factor was the bandwidth of second formant. The bandwidth found from recorded sound was very large than bandwidth of synthesized sound. Since bandwidth is an important factor (specially bandwidth of second formant in perceiving nasal vowels) that is why vowel was not perceived normally and was confused with [u].

6. CONCLUSION

After performing the experiment we can say that HLSyn was successful in producing selected oral vowels and can be used to produce other oral vowels of Urdu. Talking about nasal vowels there was problem. The major problem was controlling bandwidth of formants f2 and above. Whenever there was a huge increase in the bandwidth after nasalization of vowel, HLSyn was unable to increase the bandwidth to the required value. So we cannot say that all the other nasal vowels of the Urdu can be synthesized successfully.

7. REFERENCES

Bokhari, S. 1985. *Phonology of Urdu Language*. Royal Book Company, Karachi, Pakistan.

HLSyn High-Level Speech Synthesizer. User Interface Manual. , Sensimetrics Corporation , Massachusetts.

Kachru, Y. 1987. "Hindi-Urdu," in *The Major Languages of South Asia, The Middle East and Africa*. Comrie, B (eds.). Routledge, London, UK.

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

Khan, M. 1997. *Urdu Ka Sauti Nizam*. Muqtadara Qaumi Zaban, Islamabad, Pakistan.

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

Stevens , Kenneth N. 2000. *Acoustic Phonetics*. The MIT Press. England

8. APPENDIX A – (HLSYN PARAMETERS FOR ORAL AND NASAL VOWELS)

TABLE A.1 Values for [i]

Time	Ag	al	Ab	an	ue	f0	f1	f2	f3	f4	Ps	dc	ap
0	0.0	100	100	0.0	0.0	1500	240	2650	3150	4015	0	0	0
100	4.2	100	100	0.0	0.0	1500	240	2650	3150	4015	8.0	0	0
200	4.2	100	100	0.0	0.0	1500	240	2650	3150	4015	8.0	0	0
300	4.2	100	100	0.0	0.0	1500	240	2650	3150	4015	8.0	0	0
400	4.2	100	100	0.0	0.0	1500	240	2650	3150	4015	8.0	0	0
500	4.2	100	100	0.0	0.0	1500	240	2650	3150	4015	8.0	0	0
600	0.0	100	100	0.0	0.0	1500	240	2650	3150	4015	0.0	0	0

TABLE A.2 Values for [æ]

Time	Ag	al	Ab	an	ue	f0	f1	f2	f3	f4	Ps	dc	ap
0	0.0	100	100	0.0	0.0	1680	795	1867	2804	3392	0	0	0
100	4.2	100	100	0.0	0.0	1680	795	1867	2804	3392	8.0	0	0
200	4.2	100	100	0.0	0.0	1680	795	1867	2804	3392	8.0	0	0
300	4.2	100	100	0.0	0.0	1680	795	1867	2804	3392	8.0	0	0
400	4.2	100	100	0.0	0.0	1680	795	1867	2804	3392	8.0	0	0
500	4.2	100	100	0.0	0.0	1680	795	1867	2804	3392	8.0	0	0
600	0.0	100	100	0.0	0.0	1500	240	2650	3150	4015	0.0	0	0

TABLE A.3 Values for [u]

Time	Ag	al	Ab	an	ue	f0	f1	f2	f3	f4	Ps	dc	ap
0	0.0	100	100	0.0	0.0	1620	319	784	2520	3709	0	0	0
100	4.2	100	100	0.0	0.0	1620	319	784	2520	3709	8	0	0
200	4.2	100	100	0.0	0.0	1620	319	784	2520	3709	8	0	0
300	4.2	100	100	0.0	0.0	1620	319	784	2520	3709	8	0	0
400	4.2	100	100	0.0	0.0	1620	319	784	2520	3709	8	0	0
500	4.2	100	100	0.0	0.0	1620	319	784	2520	3709	8	0	0
600	0.0	100	100	0.0	0.0	1620	319	784	2520	3709	0	0	0

TABLE A.4 Values for [ɑ]

Time	Ag	al	Ab	an	ue	f0	f1	f2	f3	f4	Ps	dc	Ap
0	0.0	100	100	0.0	0.0	1500	660	1055	2406	3556	0	0	0
100	4.2	100	100	0.0	0.0	1500	660	1055	2406	3556	8	0	0
200	4.2	100	100	0.0	0.0	1500	660	1055	2406	3556	8	0	0
300	4.2	100	100	0.0	0.0	1500	660	1055	2406	3556	8	0	0
400	4.2	100	100	0.0	0.0	1500	660	1055	2406	3556	8	0	0
500	4.2	100	100	0.0	0.0	1500	660	1055	2406	3556	8	0	0
600	0	100	100	0.0	0.0	1500	660	1055	2406	3556	0	0	0

TABLE A.5 Values for [i]

Time	ag	al	Ab	an	ue	F0	F1	F2	F3	F4	Ps	dc	ap
0	0.0	100	100	0.0	0.0	1500	240	2650	3150	4015	0	0	0
100	4.2	100	100	50.0	0.0	1500	240	2650	3150	4015	8.0	0	0
200	4.2	100	100	50.0	0.0	1500	240	2650	3150	4015	8.0	0	0
300	4.2	100	100	50.0	0.0	1500	240	2650	3150	4015	8.0	0	0
400	4.2	100	100	50.0	0.0	1500	240	2650	3150	4015	8.0	0	0
500	4.2	100	100	50.0	0.0	1500	240	2650	3150	4015	8.0	0	0
600	0.0	100	100	0.0	0.0	1500	240	2650	3150	4015	0.0	0	0

TABLE A.6 Values for [æ̃]

Time	Ag	al	ab	an	ue	f0	f1	f2	f3	f4	Ps	dc	ap
0	0.0	100	100	0.0	0.0	1500	795	1867	2804	3392	0	0	0
100	4.2	100	100	40.0	0.0	1500	795	1867	2804	3392	8.0	0	0
200	4.2	100	100	40.0	0.0	1500	795	1867	2804	3392	8.0	0	0
300	4.2	100	100	40.0	0.0	1500	795	1867	2804	3392	8.0	0	0
400	4.2	100	100	40.0	0.0	1500	795	1867	2804	3392	8.0	0	0
500	4.2	100	100	40.0	0.0	1500	795	1867	2804	3392	8.0	0	0
600	0.0	100	100	0.0	0.0	1500	240	2650	3150	4015	0.0	0	0

TABLE A.7 Values for [ũ]

Time	Ag	al	ab	an	ue	f0	f1	f2	f3	f4	Ps	dc	ap
0	0.0	100	100	0.0	0.0	1620	319	784	2520	3709	0	0	0
100	4.2	100	100	40.0	0.0	1620	319	784	2520	3709	8	0	0
200	4.2	100	100	40.0	0.0	1620	319	784	2520	3709	8	0	0
300	4.2	100	100	40.0	0.0	1620	319	784	2520	3709	8	0	0
400	4.2	100	100	40.0	0.0	1620	319	784	2520	3709	8	0	0
500	4.2	100	100	40.0	0.0	1620	319	784	2520	3709	8	0	0
600	0.0	100	100	0.0	0.0	1620	319	784	2520	3709	0	0	0

TABLE A.8 Values for [ɑ̃]

Time	Ag	al	ab	an	ue	f0	f1	f2	f3	f4	Ps	dc	Ap
0	0.0	100	100	0.0	0.0	1500	660	1055	2406	3556	0	0	0
100	4.2	100	100	50.0	0.0	1500	660	1055	2406	3556	8	0	0
200	4.2	100	100	50.0	0.0	1500	660	1055	2406	3556	8	0	0
300	4.2	100	100	50.0	0.0	1500	660	1055	2406	3556	8	0	0
400	4.2	100	100	50.0	0.0	1500	660	1055	2406	3556	8	0	0
500	4.2	100	100	50.0	0.0	1500	660	1055	2406	3556	8	0	0
600	0	100	100	0.0	0.0	1500	660	1055	2406	3556	0	0	0