THE WAVEFORM MODEL OF VOWEL PERCEPTION AND PRODUCTION

THE WAVEFORM MODEL OF VOWEL PERCEPTION AND PRODUCTION

MICHAEL A. STOKES

Universal-Publishers
Boca Raton

The Waveform Model of Vowel Perception and Production

Universal-Publishers
Boca Raton, Florida • USA
2009

ISBN-10: 1-59942-888-1
ISBN-13: 978-1-59942-888-8

www.universal-publishers.com

CONTENTS

Acknowledgements vii
Abstract ix
Author's Background xi

Chapter 1
Background **13**

Chapter 2
Experiment I **21**
Method 21
Results 27

Chapter 3
Experiment II **33**
Method 33
Results 37

Chapter 4
Two Exceptions and One Female **43**
Experiment I 43
Experiment II 48

Chapter 5
Discussion **53**
Summary of the Experiments and Waveform Model 53
Future Work 55
Summary 58

References **61**

Appendix A
Measurements from all 20 talkers used
in Experiments I and II **63**

Appendix B
Waveform Displays
Display of small segments of vowel waveforms
for all 9 vowels used in Experiments I and II **85**

ACKNOWLEDGEMENTS

Special thanks to Dr. John Mullennix for his technical support and talker database. Our discussions and his critical feedback have been invaluable in this effort. Early support from Dr. Brian Scott also came at a needed point in time. Finally, this study would not have been possible without the love and support of my parents and family.

ABSTRACT

The visual cues contained within raw complex waveforms have led to a unique method of organizing the vowel space. Formant relationships are used to categorize and distinguish each vowel, with the categorization method associating a formant frequency with a specific articulatory gesture. The method of categorization also provides an explanation for vowel perceptual errors identified in human and computer vowel perception experiments. As a test of the model described in this book, formant frequency measurements were obtained from the mid point of the vowels produced by 20 males (524 total vowels), with the vowels subsequently identified using the features of the Waveform Model in perceptual experiments. Results showed 93.9% vowel identification accuracy across the 20 males using the original method of calculating the relationship of F1 to F0 in Experiment I, with the few errors observed accounted for within the model. Changing the method of calculating the relationship of F1 to F0 in Experiment II produced 97.7% accuracy across the same vowels tested in Experiment I. The focus here will be to introduce and describe the unique concepts of the Waveform Model, and to describe the experimental evidence that supports the model.

AUTHOR'S BACKGROUND

The author has extensive experience analyzing over 15,000 speech waveforms, with this work exhibited in a number of past presentations and published work. The first significant project involved analyzing speech produced in noise (Summers et al., 1988), followed by analysis of speech produced under conditions of cognitive workload (Summers et al., 1989). The author's next presentation represented the initial research showing that vowels can be identified from visual inspection of the waveforms of vowels (Stokes, 1996). To extend this work, the research was replicated with two additional male talkers and one female talker (Stokes, 2001). Beyond the identification of vowels from visual analysis of waveforms, the identification of a talker was achieved from waveforms (i.e., voiceprints, similar to the use of fingerprints, Stokes, 2002). Altogether, this work is innovative in its reliance on waveform displays as an analysis tool, and has contributed to the overall understanding of speech perception in a number of environments.

Since 1998, the author has been employed as a computer programmer working on a number of international business applications. The programming and database skills utilized in these positions have led to the ability to perform the experiments and analyze the data in great detail. The internet programming experience also allowed the author to post early versions of the model under the working name MAS Model of Vowel Perception and Production on the internet in December of 1998. The feedback obtained from researchers in the field proved to be valuable as the work progressed to the level it is now with 97.7% vowel identification accuracy across 20 male talkers.

CHAPTER 1

BACKGROUND

The waveform model is based on combinations of cues originally identified from the visual display of raw complex waveforms. An individual formant from a vowel possesses characteristics of a sine wave at a particular frequency and amplitude. A raw complex waveform is the single wave created from the combination of vowels' individual formants. Spectrograms are a useful tool for acoustic-phonetic analysis of formants because they separate individual formants making detailed analysis of vowels' components possible. However, the interactions of the components are lost, and a great deal of effort has been spent theorizing about how the individual formants interact to give a vowel its quality. Unlike spectrogram displays, raw complex waveforms maintain the interactive quality of the formants.

There has been a fair amount of success in visually identifying speech sounds from spectrogram displays (Cole, Rudnicky, & Zue, 1979; Cole, R., and Zue, V., 1980). However, there was no such success in identifying speech sounds from raw complex waveforms until the presentations by Stokes (Stokes, 1996, 2001). In fact, it has been taught for some time that "[i]t is not possible to look at the waveform of an utterance and say what sounds occurred" (Ladefoged, 1982, p. 168). Although this statement is somewhat dated, it is still being taught. There are only limited descriptions of the visual cues present in a waveform display with one of the earliest works being Chiba & Kajiyama (1941). The best work comes from Scott (1980). Scott's work described regular patterns within the pitch periods of vowels, but was unable to assess how those patterns categorize and distinguish the vowel space. Since Scott's work, there is limited work with speech waveforms despite the advances in other fields derived from waveform analysis (Burke-Hubbard, 1998).

There are a number of competing models of speech perception (see Klatt, 1988), none of which can account for all of the following: (1) extracting formant frequencies and reconstructing vocal tract shapes; (2) accounting for variability, especially cross-speaker variability; and (3) accounting for temporal variability. The Waveform

Model can account for all these factors as well as explain perceptual errors for vowels.

The work by Peterson and Barney (1952) reported average formant values, which were used to create the initial parameters of the Waveform Model. As the specific values of pitch and the formants from 20 talkers were recorded and analyzed in the experiments described below, the categorical boundaries and error predictions became more refined and specific. The specific values allowed for the parameters to be much more detailed than averages could achieve. To simplify and limit the explanation of the model, the initial discussion will be limited to male speakers. Discussion of female talkers will follow Experiment II described below.

The lowest frequency in a complex waveform is the fundamental frequency (F0). Formants are frequency regions of relatively great intensity in the sound spectrum of a vowel, with F1 referring to the first formant, F2 referring to the second formant, and so on. From the average F0 (average pitch) and F1 values reported by Peterson and Barney (1952), a vowel can be categorized into one of six main categories by virtue of the interactions between F1 and F0. The relative categorical boundaries are established by the number of F1 cycles per pitch period, with the following rules determining how a vowel is first assigned to a main vowel category.

Category 1: $1 < F1 \text{ cycles per } F0 < 2$
Category 2: $2 < F1 \text{ cycles per } F0 < 3$
Category 3: $3 < F1 \text{ cycles per } F0 < 4$
Category 4: $4 < F1 \text{ cycles per } F0 < 5$
Category 5: $5.0 < F1 \text{ cycles per } F0 < 5.5$
Category 6: $5.5 < F1 \text{ cycles per } F0 < 6.0$

Each main category consists of a vowel pair, with the exception of Categories 3 and 6, which have only one vowel. Once a vowel waveform has been assigned to one of these categories, identification is reduced to making a further distinction between the vowel pairs.

One vowel within each categorical pair (Categories 1, 2, 4, and 5) has a third acoustic wave present (visually), while the other vowel of the pair does not. The presence of F2 in the range of 2000 Hz can be recognized visually as this third wave, but F2 values in the range of 1000 Hz have no visually discernable features. Since each main category has one vowel with F2 in the range of 2000 Hz and one vowel in the range of 1000 Hz (see Table 1), F2 frequencies provide an

easily distinguished feature between the categorical vowel pairs. In fact, this is directly analogous to the distinguishing feature between the stop consonants /b/-/p/, /d/-/t/, and /g/-/k/ (the presence or absence of voicing). F2 values in the range of 2000 Hz are analogous to voicing being added to /b/, /d/, and /g/, while F2 values in the range of 1000 Hz are analogous to the voiceless quality of the consonants /p/, /t/, and /k/. Considering this similarity with an established pattern of phoneme perception, the model of vowel perception described here appears more than plausible.

Table 1 - *Waveform Model Organization of the Vowel Space*
Formants values from Peterson, G.E., & Barney, H.L. (1952).

Vowel - Category	**F0**	**F1**	**F2**	**F3**	**F1-FO/100**	**F1/FO**
/i/ - 1	136	270	2290	3010	1.35	1.99
/u/ - 1	141	300	870	2240	1.59	2.13
/I/ - 2	135	390	1990	2550	2.55	2.89
/U/ - 2	137	440	1020	2240	3.03	3.21
/er/ - 3	133	490	1350	1690	3.57	3.68
/ ɛ / - 4	130	530	1840	2480	4.00	4.08
/ɔ/ - 4	129	570	840	2410	4.41	4.42
/æ/ - 5	130	660	1720	2410	5.30	5.08
/ʌ/ - 5	127	640	1190	2390	5.13	5.04
/a/ - 6	124	730	1090	2440	6.06	5.89

Identification of the vowel /er/ (the lone member of Category 3) is also aided by the presence of a third acoustic wave. However, the appearance of this wave for this vowel does not conform to the categorical pair's appearance. This particular third wave is unique and provides additional information that distinguishes /er/ from neighboring categorical pairs. The vowel /a/ (the lone member of Category 6), follows the format of Categories 1, 2, 4, and 5, but it does not have a high F2 vowel paired with it, possibly due to articulatory limitations.

A successful model of vowel perception should also be able to explain other relationships associated with vowels. As mentioned above, the categorized vowel space described here is analogous to the stop consonants /b/-/p/, /d/-/t/, and /g/-/k/. To extend the analogy and the similarities, each categorical vowel pair should share a common articulatory gesture that establishes the categorical boundaries. In other words, each vowel within a category should share an articulatory gesture that produces a similar F1 value since it is F1 that varies between categories (F0 remains relatively constant). Furthermore, there should be an articulatory difference between categorical pairs that produces the difference in F2 frequencies, similar to the addition of voicing or not by vibrating the vocal folds. The following section organizes the articulatory gestures involved in vowel production by the six main categories.

From Table 2, it can be seen that the common articulatory gesture between categorical pairs is tongue height. Each categorical pair shares the same height of the tongue in the oral cavity, meaning the air flow through the oral cavity is being unobstructed at the same height within a category. This appears to be the common place of articulation for each category as /b/-/p/, /d/-/t/, and /g/-/k/ share a common place of articulation. It should also be noted that the tongue position provides an articulatory difference within each category by alternating the portion of the tongue that is lowered to open the airflow through the oral cavity. One vowel within a category has the airflow altered at the front of the oral cavity, while the other vowel in a category has the airflow altered at the back. The subtle difference in the unobstructed length of the oral cavity determined by where the airflow is altered by the tongue (front or back) is the likely source of the 30 to 50 cps difference between vowels of the same category. Although this is probably a valuable cue for the auditory system when identifying a vowel, this difference provides little information when visually identifying a vowel from the raw complex

waveform. Fortunately, there is a visual cue that can distinguish between categorical pairs.

Table 2 - *Articulatory relationships*
Articulatory positions from Ladefoged, P. (1982).

Vowel - Category	**Relative Tongue Positions**	**F1**	**Relative Lip Position**	**F2**
/i/ - 1	high, front	270	unrounded, spread	2290
/u/ - 1	high, back	300	rounded	870
/I/ - 2	mid-high, front	390	unrounded, spread	1990
/U/ - 2	mid-high, back	440	rounded	1020
/er/ - 3	rhotacization	490	retroflex	1350 (F3=1690)
/ ɛ / - 4	mid, front	530	unrounded	1840
/ɔ/ - 4	mid, back	570	rounded	840
/æ/ - 5	low, front	660	unrounded	1720
/ʌ/ - 5	mid-low, back	640	rounded	1190
/a/ - 6	low, back	730	rounded	1090

As mentioned above, there is a third wave (high frequency, low amplitude) present in one of the categorical vowel pairs which visually distinguishes it from the other vowel in the category. From Table 3, one sees that one vowel from each pair is produced with the lips rounded, and the other vowel is produced with the lips spread or unrounded. An F2 in the range of 2000 Hz clearly appears to be associated with having the lips spread or unrounded. Therefore, having the lips spread or unrounded is directly analogous to vibrating the vocal folds during the production of /b/, /d/, and /g/ to add the voicing that distinguishes them from /p/, /t/, and /k/, respectively. Furthermore, rounding the lips (resulting in an F2 in the range of 1000 Hz) during vowel production is analogous to the voiceless quality of /p/, /t/, and /k/.

By organizing the vowel space in this way, it is possible to predict perceptual errors. The confusion data shown in Table 3 has Categories 1, 2, 4, and 5 organized in that order. Category 3 (/er/) is not in Table 3 because its formant values (placing it in the "middle" of the vowel space) make it a unique case to explain. Specifically, the distinct F2 and F3 values of /er/ necessitate an extension to the general rule described below. Rather than distract from the general rule explaining confusions between the four categorical pairs, the acoustic boundaries and errors involving /er/ will be discussed in the experimental evidence below. Furthermore, even though /a/ follows the general format of error prediction described below, Category 6 is not shown since /a/ does not have a categorical mate and many dialects have difficulty differentiating between /a/ and / ɔ /.

The Waveform Model predicts that errors occur across category boundaries, but only similar F2 vowels are confused for each other. In other words, a vowel with an F2 in the range of 2000 Hz will be confused for another vowel with an F2 in the range of 2000 Hz. Similarly, this is the case for vowels with F2 in the range of 1000 Hz. Vowel confusions are the result of misperceiving the number of F1 cycles per pitch period. In this way, F2 frequencies limit the number of possibilities as error candidates. Also as expected, confusions are more likely with a near neighbor (separated by one F1 cycle per pitch period) than with a distant neighbor (separated by two or more F1 cycles per pitch period). From the four categories that are shown, 2,983 out of 3,025 errors (98.61%) can be explained by searching for neighboring vowels with similar F2 frequencies.

Table 3 - *Error Prediction*
Error data reported by Peterson, G.E., and Barney, H.L. (1952).

Vowels Intended by Speaker	**Vowels as Classified by Listeners**			
	/i/-/u/	/I/ - /U/	/ ɛ / - /ɔ/	/æ/ - /ʌ/
/i/ /u/	10,267 - - 10,196	*4* --- --- *78*	*6* 3 1 ---	--- --- --- ---
/I/ /U/	*6* --- --- *96*	9,549 --- --- 9,924	*694* 1 1 *51*	2 --- 1 *171*
/ ɛ / /ɔ/	--- --- --- *5*	*257* --- --- *71*	9,014 3 1 9,534	*949* 2 2 *62*
/æ/ /ʌ/	--- --- --- ---	*1* --- 1 *103*	*300* 2 1 *127*	9,919 15 8 9,476

To this point, the vowel /er/ in Category 3 has not been discussed in detail. It is convenient for the waveform model that the one vowel that has a unique lip articulatory style when compared to the other vowels of the vowel space results in formant values that lie between the formant values of neighboring categories. This is especially evident when the F2 and F3 values of /er/ are compared to the other categories. Both the F2 and F3 values lie between the ranges of 1000 Hz to 2000 Hz of the other categories. With the lips already being directly associated with F2 values, the unique retroflex position of the lips to produce /er/ further demonstrates the role of the lips in F2 values, as well as F3 in the case of /er/. The quality of a unique lip position during vowel production produces a unique F2 and F3 value.

The first demonstration of the ability to identify vowels from visual displays of waveforms was presented in 1996 (Stokes, 1996). In that experiment, waveform displays of nine vowel productions from two Midwestern males were presented for identification. For each talker, subject MAS was presented with the nine waveforms of the speakers in random order by an experimenter. The subject was

allowed to review each waveform before a final response was given and scored. The results showed that subject MAS correctly identified 5 out of 9 vowels for both speakers. The 55% accuracy across the vowel space of two speakers was well above chance (chance would be 1 correct out of 9 vowels per talker) and demonstrated the potential success of the model and how the cues are unaffected by talker variability. The 1996 study was replicated and extended to a female talker in 2001 (Stokes, 2001). In that study, subject MAS successfully identified 4 out of 9 vowels and 6 out of 9 vowels for two additional male speakers (again, 55% accuracy across these two talkers) and 4 out of 9 vowels for a female speaker (44% accuracy).

Despite the success of the two visual identification experiments, the method was subjective and difficult to replicate. The present study eliminates the subjectivity and provides a method that can easily be replicated. The analysis of the results also served to refine the categorical and distinguishing feature boundaries.

CHAPTER 2

EXPERIMENT I

Method

Subjects

A talker database of h-vowel-d (hVd) productions (Mullennix, 1994) was used as the source of vowels analyzed for this study. The entire database consists of 33 male and 44 female college students in Detroit, MI, who produced three tokens for each of the ten American English vowels. The recordings were made using CSRE software (Jamieson et al., 1992) and converted to .wav files using the TFR software package (Avaaz Innovations Inc., 1998). Of the 33 male talkers in the database, 20 were randomly selected for use in the current experiment. In the original experiments demonstrating identification of vowels by visual inspection of the waveforms (Stokes, 1996 and 2001), five talkers from the talker database (4 males and 1 female) were used. None of the unique talkers used in those experiments were used in the experiments presented here.

Apparatus

The nine vowels used in this study were / i /, / u /, / I /, / U /, / er /, / ɛ /, / ɔ /, / æ /, / ^ /. Although the vowel /a/ is available in the database, it was excluded from this study since many Midwestern dialects have difficulty differentiating between /a/ and / ɔ /. In most cases, there were three productions for each of the nine vowels used (27 productions per talker), but there were instances of only two productions for a given vowel by a talker. Across the 20 talkers, there were 524 vowels analyzed and every vowel was produced at least twice by each talker.

The research was performed on a Compaq Presario 2100 laptop computer, and the speech signal processing was performed with the TFR software. The collected data was entered into a Microsoft Access database where the data was mined and queried using the SQL query language. The programming language Cold Fusion was used to display the data and results in a browser. Within the Cold Fusion

program, the necessary calculations and the conditional if-then logic were programmed. This allowed for quick modifications to the model parameters used to categorize and identify the vowel space. As errors occurred, modifications to the values used in the conditional logic (for example, the defined range of F2 values) for a vowel could be edited to eliminate the error. The programming and query languages made the error reduction process quick, and provided various statistics such as processing time.

Design

The speech signal processing begins by identifying the center of each vowel (within 10 ms) in order to make the pitch and formant frequency measurements at the most neutral and stable portion of the vowel. Figure 1 shows the display of the production of "whod" by Talker 12. From this display, the cursor can be placed in the relative center of the vowel to identify the time within the display associated with the center. A range of between 20 and 30 ms of that point in time can be used as the point in time that the pitch and formant values will be measured.

Once the relative center of the vowel is identified, the fundamental frequency was measured first so that the specific time that the measurements were made was not associated with an unusually high or low pitch frequency compared to the rest of the central portion of the vowel. The Cepstrum method (i.e., taking the Fourier Transform of the decibel spectrum) within TFR was used to perform the pitch extraction. Figure 2 shows the Cepstrum display for the "whod" production by Talker 12. Using the time identified from the display in Figure 1, the pitch measurement can be made at this central point in time. The point in time and the F0 value were recorded before performing the formant analysis.

The F1, F2, and F3 frequency measurements were made at the same point in time as the pitch measurement using the Formant Module in TFR. Figure 3 shows the Formant Module display of the production of "whod" by Talker 12 which is a typical visual display used during the formant measurement process. These measurements were recorded with the time and pitch from the display in Figure 2 before moving to the next vowel to be analyzed. For each production, the intended vowels identity, the point in time for the measurements, and the F0, F1, F2, and F3 values were recorded and entered into the Access database to be used in the computer recognition perceptual experiment.

Figure 1

This is a display of the entire vowel file to identify the relative center of the vowel.

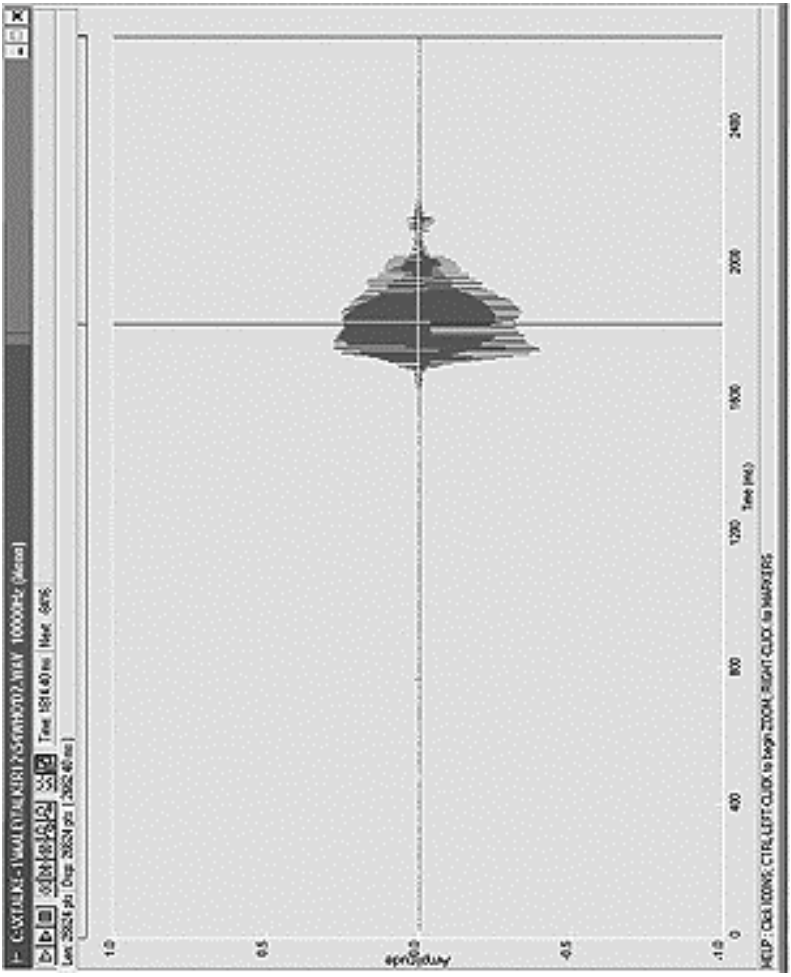

Figure 2

This is a Cepstral pitch display used to measure F0 for the vowel.

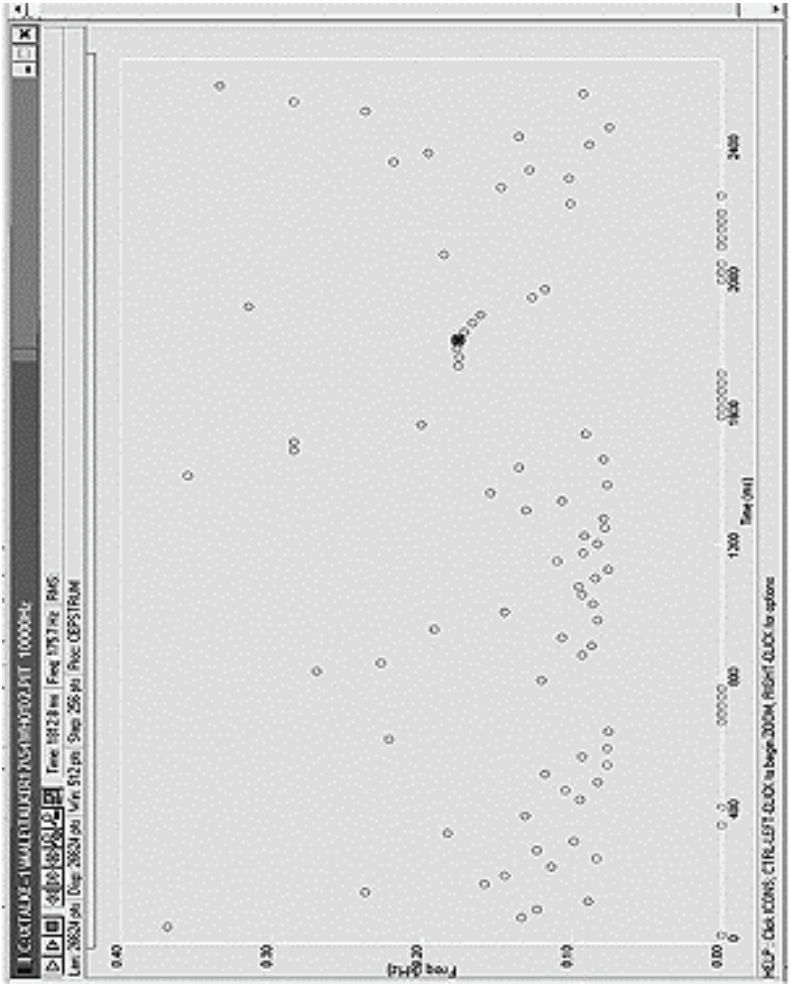

Figure 3

Formant display used to measure F1, F2, and F3 (without color).