The Vocal Joystick Data Collection Effort and Vowel Corpus

Kelley Kilanski, Jonathan Malkin, Xiao Li, Richard Wright, Jeff A. Bilmes

Dept. of Linguistics , Dept. of Electrical Engineering
University of Washington, Seattle, Washington, USA
bilmes@ee.washington.edu

Abstract

Vocal Joystick is a mechanism that enables individuals with motor impairments to make use of vocal parameters to control objects on a computer screen (buttons, sliders, etc.) and ultimately will be used to control electro-mechanical instruments (e.g., robotic arms, wireless home automation devices). In an effort to train the VJ-system, speech data from the TIMIT speech corpus was initially used. However, due to problematic issues with co-articulation, we began a large data collection effort in a controlled environment that would not only address the problematic issues, but also yield a new vowel corpus that was representative of the utterances a user of the VJ-system would use. The data collection process evolved over the course of the effort as new parameters were added and as factors relating to the quality of the collected data in terms of the specified parameters were considered. The result of the data collection effort is a vowel corpus of approximately 11 hours of recorded data comprised of approximately 23500 sound files of the monophthongs and vowel combinations (e.g. diphthongs) chosen for the Vocal Joystick project varying along the parameters of duration, intensity and amplitude. This paper discusses how the data collection has evolved since its initiation and provides a brief summary of the resulting corpus.

Index Terms: Speech corpora, data collection procedures, speech recognition, Speech HCI for individuals with impairments, Speech/voice-based human-computer interfaces

1. Introduction

Vocal Joystick is a continuous control device that uses human vocalizations (non-linguistic sounds) as inputs [1] [2]. It is intended to be an assistive device for people with motor impairments to allow users to control objects on a computer screen (buttons, sliders, mouse, etc.) and ultimately electro-mechanical instruments (e.g., robotic arms, wireless home automation devices). Standard spoken language is inefficient for such continuous control tasks as it is discrete in nature and is often recognized poorly by automatic speech recognizers. Therefore, the input vocalizations should be robustly recognized in a variety of environments, and easy to learn regardless of the native language of the user while minimizing repetitive strain and maximizing ease of use. Moreover, the input parameters should be flexible enough to accommodate a wide range of continuous control applications. Fortunately, the anthropophonic repertoire includes an ample range of continuous and discrete sounds to draw from. In this paper we describe the rationale for and development of a training corpus for the continuous inputs to the device. Although there are a relatively large number of continuous signals that were possible candidates, some of which we intend to explore in the future, we settled on three types of vocalizations for the current version: vowel-like sounds (representing two dimensions), pitch, and intensity.

The goal of the Vocal Joystick data collection effort has been to collect vowel samples that would not only be language independent and robust in a variety of environments, but also be representative of how a user would utter vowels when using the VJ-system. Over the course of the Vocal Joystick project the data collection effort has gone through many changes in order to collect the desired vowel samples. Without the availability of a suitable speech corpus, we began a large data collection effort in order to create an appropriate speech corpus that would serve the needs of the Vocal Joystick project. The following sections discuss the data collection effort. Specifically, Section 2 discusses why the continuous set of sounds used in data collection were chosen, Section 3 discusses the procedures for data collection, and Section 4 discusses general details of the resulting corpus.

2. Background

In the world's languages (e.g. as described in [6]) continuous sounds can be drawn from three main classes: 1) *vocalic (vowel like) sounds* that result from the resonances of the vocal tract shape which can change continuously depending on the jaw, lip, and tongue position as long as there is no significant obstruction in the vocal tract; 2) *pitch (rate of vocal fold vibration)* which results from a complex interaction between sub glottal (lung) pressure and vocal fold tension (resulting from a variety of muscular adjustments) again as long as downstream adjustments do not impede airflow across the vocal folds; and 3) *intensity* that generally results from changes in sub-glottal pressure (for voiced sounds).

In general, vocalic sounds can be described as occupying points in or as movement through a two dimensional space that is made up of primarily the first two resonances of the vocal tract (e.g., [4, 7]) Steady state vowels like [i] or [æ] are points and vocalic transitions caused by glides such as [w] or diphthongs such as [oi] are transitional sounds. This feature makes vocalic sounds an ideal candidate for two-dimensional movement. Pitch and intensity add two additional dimensions that can be exploited for directionality, velocity or other purposes.

Moreover, vocalic signals, manipulations of pitch, and manipulations of intensity are found as quasi-independent but coexisting elements in every spoken language. That is, in every language vowels, pitch and intensity can be manipulated mutually independently for linguistic purposes. With these considerations, the resulting continuous set chosen for the Vocal Joystick was based on physiological capabilities of the human vocal tract: a question was how many equidistant vowel sounds and vowel-to-vowel transitions are possible to make. The resulting continuous set includes nine monophthongs: /i,

ɨ, e, ə, a, æ, ɑ, o, u/; and 12 vowel-to-vowel transitions: /i-u, u-i, æ-ɑ, ɑ-æ, æ-i, i-æ, æ-u, u-æ/.

2.1. Data Collection Effort

Initially, the VJ-system was trained using the TIMIT speech corpus, a speech database consisting of a number of read-speech utterances that were phonetically transcribed. We used this training material for our vowel classifiers for VJ directional control. The vowels in TIMIT, however, are uttered in a way quite different than a user of a VJ-system will utter them. Specifically, in TIMIT the vowels contain significant coarticulatory effects, which will not exist when using vowels to control the VJ-system. We researched the availability of speech corpora containing vowels with human-validated format tracks for use in training. But it appears that no such appropriate corpus is available.

In order to significantly improve the VJ-system accuracy, we completed a large data-collection effort. We defined a large set of vowels and diphthongs with various combinations of intonation, volume, and length. We then designed and implemented a Mac OSX application which allows experimental subjects, via a simple dialog box interface, to listen to example recordings of complex vowels and then be recorded pronouncing them. This allows for rapid collection of a corpus of vowel sounds that is pre-segmented and pre-labeled.

3. Methods

3.1. Participants

For the Vocal Joystick data collection effort, a total of 97 participants were recruited ranging in age from 18 to 60 with a median age range of 18-25. In order to collect a representative sampling of the target vowels, both native English speakers from various dialectal regions of North America and non-native speakers of English were recruited for participation. Of the 97 participants, 24 were non-native speakers of English, while 73 were native speakers. While we would have like to have had more dialectal variation in the data collected, 53 of the speakers were from the Pacific Northwest. Additionally, 6 participants participated in more than one task resulting in a greater representation in the data for those users.

3.2. Data Collection Procedures

Recordings took place in a sound attentuated recording booth in the University of Washington Phonetics Lab. Through the use of the Record Dialog program the vowels were recorded onto a Power Mac G5 at a sampling rate of 44kHz and 16 bits. Participants wore a Shure SM10A head-mounted mono-microphone attached to a Shure FP32A Amplifier, which in turn was attached to an M-Audio FireWire 410 audio/midi interface. Although the resulting sound files were in 2-channel recordings, only the left channel contained the actual sound information, while the right channel recorded silence.

While there were several versions of the Record Dialog program used over the course of the data collection effort, the same general procedures were used throughout. Upon arrival at the Phonetics Lab the participants were asked several questions: their age, native country or dialectal region, gender, and whether they had had phonetic training. The participants were then led to the sound booth, and asked to sit in front of a 17" monitor. Sound levels were set in order to avoid clipping by asking the participant to say different vowels at different amplitudes. The vowels /æ, a, i, u/ were used for setting the levels as they correspond to the two loudest and two quietest vowels respectively. After the levels were set, the target task was opened on the screen with the Praat recording meter in the background to monitor the sound levels after initial adjustment. The participants were then given basic instructions concerning the interface of the program and what sounds they would be producing. Initially, participants went through a training phase consisting of one vowel iteration. The training phase, which was not recorded, was created so the participant could get a feel for the interface and for the different lengths, amplitudes and intonations being elicited in the program. Upon completion of this part of the program, the participant began the target task.

3.2.1. Procedures for Record Dialog Version 4

The vowel quality of the recorded vowels in versions 1, 2 and 3 of the Record Dialog program was not ideal for all vowels. Specifically, many participants were not able to produce a distinction between /a/ and /ɑ/. This is not surprising since most participants were from the Pacific Northwest where the distinction between these two vowels has been lost ([7]). In order to elicit the necessary vowel qualities, in version 4 a trained phonetician was present in the sound booth to correct vowel quality whenever problems arose. Several strategies were used to elicit the target vowel qualities from the subjects. Initially the subjects listened to sound files and then reproduced the sound with correction and training by the phonetician. However, some of the subjects had difficulty perceiving the target vowel quality of the sound files. This difficulty occurred regardless of whether the participants listened to synthesized sounds or sounds produced by a human subject. To resolve this issue, the phonetician in the sound booth produced all target vowels for the participant. This new procedure resulted in better perception of the target vowel qualities.

3.3. Record Dialog Program

The Record Dialog program was created to facilitate recording vowels of various lengths, amplitudes and intonations. During the data collection effort 4 versions of the program were used. The creation of a new version was motivated by the addition or deletion of a parameter or in consideration of the amount of time the participant spent in the sound booth. Each version had from 1 to 3 tasks each of which differed in the combinations of duration, amplitude and intonation for the nine vowels /i, ɨ, e, ə, a, æ, ɑ, o, u/ and the 12 vowel combinations /i-æ, i-ɑ, i-u, æ-i, æ-ɑ, æ-u, ɑ-i, ɑ-æ, ɑ-u, u-i, u-æ, u-ɑ/

The duration parameter includes: short (1000ms), long (2000ms) and nudge (a very short production of the vowel produced three times within 1000ms), which was included in order to learn the distinct spectral characteristics of extremely short vowel utterances. Accurate modeling of such short vowels is crucial when using the vocal joystick to produce extremely precise or minute computer-screen object adjustments. The amplitude parameter includes: quiet, normal, loud, quiet to loud and loud to quiet amplitude sweeps. The intonation parameter includes: level, rising and falling.

3.3.1. *The Record Dialog Interface*

While the Record Dialog (RD) program allowed for faster and easier data collection of a large amount of data, the interface allowed for functionality for the participants as well as a certain amount of monitoring target vowel quality by the participants before the task ended. The interface that participants used to record the vowels at the specified parameters is shown in Figure 1.

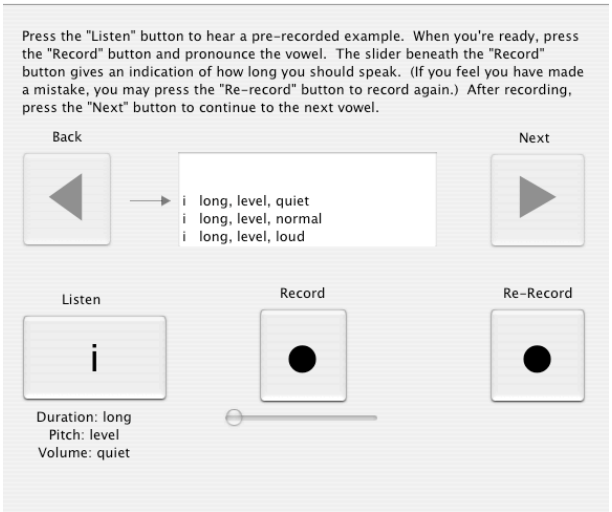

Figure 1: Illustration of Record Dialog Interface.

As shown in the figure above, several features were incorporated into the RD program and interface. Besides the basic functions of recording and advancing to the next vowel, the interface also provides a description of the target vowel and allowed participants to listen to an example of the vowel. The interface also allows for re-recording vowels if a mistake was made, and going back to a vowel that may not have been produced as specified by the description. A crucial component of the interface is the slider underneath the record button. The slider gave the user an indication of the duration each utterance should be.

3.3.2. *Record Dialog Version 1*

The first Record Dialog version had one task for the participant to complete. A total of 18 tokens were produced for each vowel and vowel combination. The tokens varied according to amplitude, intonation and duration. Each vowel was produced at quiet, normal and loud amplitudes for all intonation parameters and all duration parameters. The resulting number of vowel, length, amplitude and pitch combinations were 378. All combinations were recorded in one recording session. The time needed for recording was approximately one hour and 30 minutes. A total of 10 participants were recorded using this task. Amplitude parameters included quiet, normal and loud; intonation included level, rising and falling; and duration included short (500ms) and long (2000ms) lengths.

3.3.3. *Record Dialog Version 2*

The second version of the Record Dialog program included all duration parameters, all amplitude parameters for monophthongs. The number of resulting tokens for all combinations necessitated that the program be divided into three tasks. The first task remained exactly the same as in version 1. The second task included only monophthongs with all duration, amplitude and intonation combinations. The third task included both monophthongs and vowel combinations. The monophthongs were produced with level intonation only for short and long durations at quiet, normal and loud amplitude levels. The vowel combinations were produced with short and long durations; quiet, normal and loud amplitudes; and level, rising and falling intonation. There were 297 tokens in the second task and 270 tokens in the third task. Twelve participants were recorded for task 2, while four were recorded for task 3. The time needed for completing the second task was approximately 45 minutes to one hour and for the third task, one hour to one hour and 15 minutes.

3.3.4. *Record Dialog Version 3*

In consideration of the amount of time necessary to complete each task in Version 2, Version 3 was split into three shorter tasks. The first task included monophthongs for all durations, intonations and amplitudes with the exception of amplitude sweeps, which were only produced with level intonation. The second task was similar to task one with the exception of rising and falling intonations. Rising and falling sounds were only produced with amplitude sweeps. Six participants were recorded for the first task with a recording time of approximately 45 minutes for 225 tokens. Five participants were recorded with task two with a recording time of approximately 45 minutes for 189 tokens. The third task remained the same as task 3 in version 2. Eleven participants recorded using task 3.

3.3.5. *Record Dialog Version 4*

With the length of the program, comments about fatigue in all previous tasks and vowel quality issues for the low central and back vowel as well as rising and falling sounds, further considerations were given as to what lengths, amplitudes and intonations are absolutely needed for the Vocal Joystick project to be successful. In consideration of the program length, it was decided that the short–long length distinctions were not necessary and short duration was eliminated from the program.

Version 4 remains three separate tasks. The tasks are split based on intonation. The first task is level intonation, the second task is rising intonation and the third is falling intonation. All tasks included both monophthongs and vowel combinations with nudges produced only for level intonation and amplitude sweeps only for monophthongs. By eliminating the short length and splitting the task by intonation the resulting number of tokens for each task was greatly decreased. The first task resulted in 108 tokens; the second and third task resulted in 81 tokens each. A total of 51 participants were recorded using version 4 with each task taking approximately 30 minutes.

4. Resulting Data

The data collection process from a total of 108 performed tasks across all 4 Record Dialog versions resulted in a significant corpus of vowels at varying duration, intonation and intensity parameters. The total number of resulting sound files at this point is 23544 and there is approximately 11 hours of recorded vowel data. A summary of the recorded vowels and the amount of time in seconds

is provided in Tables 1 and 2. Table 1 lists the amount of time and the resulting number of sound files for monophthongs. Table 2 lists the amount of time and resulting number of sound files for vowel combinations.

Table 1: Sound File Summary for Monophthongs

Vowel	**Long - 2000ms**		**Short – 1000ms**		**Nudges**
	Time(s)	*Sound Files*	*Time(s)*	*Sound Files*	*Sound Files*
i	1452	726	426	426	144
ɨ	1452	726	426	426	144
e	1452	726	426	426	144
ə	1452	726	426	426	144
æ	1452	726	426	426	144
a	1452	726	426	426	144
ɑ	1452	726	426	426	144
o	1452	726	426	426	144
u	1452	726	426	426	144
TOTAL	13068	6534	3834	3834	1296

Table 1 illustrates the number of resulting sound files and total amount of recorded time in seconds for monophthongs. The total number of sound files for monophthongs across all durations, intonation and intensity parameters resulting from the data collection process is 11664 with approximately 5 hours of total recording time.

Table 2: Sound File Summary for Vowel Combinations

Vowel	**Long – 2000ms**		**Short – 1000ms**	
	Time(s)	*Sound Files*	*Time(s)*	*Sound Files*
i → ɑ	1530	765	225	225
i → æ	1530	765	225	225
i → u	1530	765	225	225
æ→ ɑ	1530	765	225	225
æ→ i	1530	765	225	225
æ→ u	1530	765	225	225
u → i	1530	765	225	225
u → æ	1530	765	225	225
u → ɑ	1530	765	225	225
ɑ → i	1530	765	225	225
ɑ → u	1530	765	225	225
ɑ → æ	1530	765	225	225
TOTAL	18360	9180	2700	2700

Table 2 illustrates the number of resulting sound files and total amount of recorded time in seconds for vowel combinations. The total number of sound files for vowel combinations across all durations, intonation and intensity parameters resulting from the data collection process is 11880 with approximately 6 hours of total recording time. As the data presented in Table 1 and Table 2 illustrate, the Vocal Joystick data collection process has resulted in a significant corpus of vowel recordings for the 9 monophthongs and 12 vowel combinations chosen for the Vocal Joystick project.

5. Conclusion

Due to a lack of an appropriate speech corpus to train the VJ-system on, the Vocal Joystick data collection effort has resulted in a corpus of vowel articulations of considerable size with vowel production varying along the parameters of duration, intensity and pitch. In order to collect data of the quality necessary for training the VJ-system, several versions of the Record Dialog Program and different procedures were used as they became necessary to control quality. Anecdotal evidence from informal user studies has shown that the change in procedures as outlined above has resulted in better functionality of the VJ-system. Future work will include additional analysis to unquestionably establish that this is the case. We also plan to make this corpus available to the community by providing it to the LDC for systemized release.

Details about the larger project and related publications can be found at: http://ssli.ee.washington.edu/vj/.

This material is based on work supported by the National Science Foundation under grant IIS-0326382

6. Acknowledgments

We would like to acknowledge Scott Drellishak for his work on creating the Record Dialog program, Andrea Macleod and Ngaio Halsey for their assistance in the initial data collection effort and Amarnag Subramanya for his assistance in program changes.

7. References

[1] Jeff A. Bilmes, Xiao Li, Jonathan Malkin, Kelley Kilanski, Richard Wright, Katrin Kirchhoff, Amarnag Subramanya, Susumu Harada, James A. Landay, Patricia Dowden and Howard Chizeck, "The Vocal Joystick: A Voice-Based Human-Computer Interface for Individuals with Motor Impairments," *Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing*, Vancouver, Canada, Oct, 2005.

[2] Jeff Bilmes, Jonathan Malkin, Xiao Li, Susumu Harada, Kelley Kilanski, Katrin Kirchhoff, Richard Wright, Amarnag Subramanya, James Landay, Patricia Dowden, Howard Chizeck, "The Vocal Joystick" *IEEE Intl. Conf. on Audio, Speech and Signal Processing*, Toulouse, France, May, 2006.

[3] J. Catford, *Fundamental Problems in Phonetics*. (Edinburgh University Press, Edinburgh, 1977).

[4] G. Fant. The relations between the area functions and the acoustic signal. *PHONETICA*, 37:1865—1875, 1980.

[5] Jennifer Ingle, Richard Wright, Alicia Beckford Wassink. "Pacific Northwest vowels – A Seattle neighborhood dialect study." 149th ASA meeting, Vancouver, BC, Canada, May, 2005.

[6] P. Ladefoged and I. Maddieson. *The Sounds of the World's Languages*. Blackwell Publishers, 1996.

[7] K. Stevens. *ACOUSTIC PHONETICS*. MIT Press, Cambridge, MA, 1998.