Methods for Experiment 3 of “Articulation and Coarticulation in the Lower Vocal Tract.”

Greg Kochanski, John Coleman, Christina Orphanidou,
A. McIntyre, Christopher Alvey, and Steven Golding

June 29, 2008

1 Methods

The experimental methods include the design of the stimuli, the data collection techniques, and the data reduction process. The data reduction is itself a multi-stage process, beginning with raw audio recordings and MRI image sequences The audio and images were first processed separately: the audio yields timing information for phonological segments, and the images yield measured widths (with error bars) for the airway. TThe phonological and articulatory data sets are then combined to yield airway widths at the centres of segments. These are the data used for testing the phonological models in the remainder of the paper.

1.1 Stimulus Design

The aim of our stimulus design process was to find a set of short phrases that would allow us to efficiently compare models for the articulatory positions. We chose to use a restricted phoneme set to keep the complexity of the data analysis and model-building to an acceptable level; within that restricted set, we needed to find phrases with a broad variety of phonological combinations.

We did this in a straightforward manner: we chose a candidate set of phrases, simulated a data analysis, and computed a score based on the measurement errors that analysis would produce. Then, we varied the set of phrases by means of a evolutionary algorithm and chose the set of phrases that gave the best score.

The scoring algorithm was designed to represent a typical measurement error, so that minimizing the score would yield a large number of moderately small measurement errors (rather than a few extremely small measurement errors).

to ensure that the stimulus selection procedure would not somehow favour one of our experimental model over another we actually simulated four analyses and summed their scores, The four simulated analyses were based on our experimental models, but they differed in three ways: (1) they involved coarticulation terms, which the models here do not, (2) they predicted only positions, not velocities, and (3) they treated syllables in stressed phonemes as different from unstressed.

The candidate phrases were chosen from a larger set of 388 phrases. These averaged 1.9 words or 12 characters (counting inter-word spaces). The larger set, itself, was generated by a semi-automated, iterative process. We first selected a limited phoneme set, then selected words from the Hornby (1974) dictionary that were pronounced with only those phonemes. These were randomly combined into short phrases, then awkward or absurd combinations were manually rejected. We then looked at the phoneme bi-gram statistics of the large phrase set (after rejections). If some bi-grams were missing we added more phrases that contained the missing bi-grams using a similar procedure. We went through several iterations of this process until it was clear that we had found most of the bi-grams that could be found in phrases that were comfortable to read and that wouldn't induce too much silliness amount the experimental subjects.

The end result of this procedure is a set of sentences that covers a relatively large fraction of the phoneme bi-grams that are constructible from our target phoneme set. The final set of stimuli consisted of 75 phrases with a mean length of 1.9 words, containing 64 distinct words. These were then assigned randomly to the volunteers.

1.1.1 Phoneme Definitions

To define our set of phonemes, both for the stimulus generation and the later segmentation (§1.3.3), we followed the pronunciations given in Hornby (1974), except that a few of the pronunciations were modified to reflect current pronunciation. Specifically, final /I/ was changed to /i/ in the words "sonny", "infancy", "saucy", "needy", "sunny", and "indecency".

When segmenting, the labellers were given these transcriptions and asked to use them unless the speaker obviously used another pronunciation. The labellers marked a few instances of /eI/, /l/, ZZZ /R/, and /aɪ/ phones[1]; an utterance containing any of these[2] was considered a mispronunciation and dropped. Overall, 12 utterances were dropped for this reason.

1.2 Data Collection

Data collection procedures followed Alvey et al. (2008) closely, with some changes in the details of the MRI sequence used and that we collected data only with gated sequences.

Subjects attended a session at the Phonetics Lab where they were briefed on the experimental procedures, MRI imaging, screening was done, and informed consent obtained. Also, at the session, they spent 5–10 minutes reading all the experimental phrases to the beat of a metronome (typically, they read a few repetitions of each phrase. Our intent was to let the subjects pick a comfortable pronunciation and stress pattern; we did not want them experimenting in the MRI machine were pronunciation reproducibility is crucial.[3] Reasonable care was taken not to coach the subjects toward particular pronunciation patterns. They also had the opportunity to pick comfortable metronome rate(s) to be used for the later MRI session.

Several days later, at the MRI session, subjects completed a standard MRI safety screening before entering the MRI machine. They had access to an intercom and a alarm button during the scanning process. They read the phrases from a screen approximately 3 m away. On occasion, the subject would ask to repeat a phrase; we always accepted. On occasion, the experimenter would notice that the image quality low and suggest that the subject repeat a phrase. [4] Overall, no more than ZZZ 3 of the 26 phrases were repeated for any subject.

[1] Respectively, “eI”, “l”, “R”, and “ai” in the Hornby (1974) phonetic representation.

[2] This check was made after §1.3.3, so alternative pronunciations would only appear and cause the sentences to be dropped if they were the most common pronunciation amongst a subject’s repetitions of a phrase.

[3] One of the contributions to pronunciation variability that we identified during the data analysis was that subjects would sometimes change their breathing pattern during a sequence of repetitions. An extended practice session that included 20 repetitions of phrases might help to reduce this problem.

[4] Note that the experimenter could not hear the subject during the scanning process. The experimenter was listening on a separate intercom, rather than the experimental audio system.

Subjects were offered a break at the midpoint of the experiment, but not all accepted it. Subjects spent approximately 40 minutes in the MRI machine during the experiment.

MRI machines are noisy environments and the scanning process essentially prohibits the use of more than tiny amounts of metal near the subject. As a result, we collected our audio data with an optical, non-magnetic, gradient microphone (ZZZ REF Phon-Or, Inc) placed near the subject's lips. (We used the microphone outputs directly, rather than using Phon-Or's noise reduction software.)

The subjects's head is within a receiver coil that essentially forms a cage around the head, typically leaving 1–4 cm of free space around the head. However, the microphone is large enough so that it will not always fit into the available space between the head and the receiver coils, while staying close to the mouth, and properly aligned so that noise from the breath stream is minimal. Because of this, our microphone placement varied from subject to subject, ranging from the optimal 1 cm distance up to as much as 5 cm.

1.3 Audio Processing

The output of the audio processing consists of a set of audio files, together with a time marker for the centre of each phoneme and a time marker for each metronome tick.

1.3.1 Microphone and Input Processing

The microphone (§1.2) does not have a particularly flat frequency response. Below 1 kHz the response rises slowly, then it climbs to a sharp (Q=7.5) resonance at 3.4 kHz. We designed a Fourier-Transform filter to flatten the microphone's frequency response; the main design constraint for the filter was that we limited the maximum gain of the filter to 8, to make sure that electrical noise would not be excessively amplified. The resulting flattened response rises to a broad peak at 3.7 kHz, with an upper cutoff (-6 dB) of 6 kHz. On the low frequency side, it is -3 dB at 1.6 kHz and -12 dB at 500 Hz. For speech, the signal shows the first formant if it's not too low, shows the second and third formants clearly, and gives good information on fricatives.

1.3.2 Noise Subtraction

We used a gradient microphone, and the performance of such microphones depends on the source-to-microphone distance. As a result, the signal-to-noise ratio varied from subject to subject. We estimate that the SNR of the recordings, at syllable centres relative to the MRI noise ranged between approximately ZZZ-15 dB and ZZZ -3 dB. Subjectively, in the noisiest recordings, it was barely possible to hear that someone was speaking; in the best recordings the speech could be understood though useful phonetic transcription was impossible.

To reduce the noise, we used a novel technique to subtract most of the MRI noise. The technique typically yields a 20 dB improvement in signal-to-noise ratio.

After subtraction, all the recordings are easily understandable with a large enough SNR that at least partial phonetic transcription was possible on all but two of the 390 utterances. A further eight were partially transcribable (typically in the louder regions such as vowels and stressed syllables).

1.3.3 Segmentation and Labelling

The noise-subtracted audio files have SNR at syllable centres of ZZZ to ZZZ dB, generally sufficient for manual segmentation by a trained phonetician. However segmentation was difficult, in that the quietest parts of speech, especially near syllable boundaries or in unstressed syllables often could not be heard well.

We dealt with this problem in two ways. First, we provided an visual aid for the task that was designed to be better behaved in low SNR situations than the standard spectal representation. We showed the labeller a perceptual spectrum, much like Kochanski and Orphanidou (2007). We used broader frequency bins (0.7 erb) and a somewhat wider window than normal for averaging in the time domain (45 ms). This combination averaged away some noise at the cost of some fine detail in the representation.

Second, we made use of the fact that the MRI imaging required the speech to be repeated 18 times. We labelled 5–10 repetitions of each utterance (more repetitions when the boundaries were harder to determine) and then we computed the median position of the boundary relative to the metronome ticks. So, even if one or two repetitions of a particular boundary could not be determined, the result should be fairly precise. Further, as we used only the centre points of phones, we pick up a further reduction in variance because we are averaging the positions of the two edges. All future references to "segment boundaries" refer to these computed medians, and "durations" are derived from the computed medians.

In related processing, we match the labels assigned to each repetition and, for each subject, we compute the most frequent label found at each position. Again, this helps to make the analysis robust against labelling errors under conditions where the labeller cannot clearly hear the speech. Thus, in further steps, we work with the most common pronunciation of each phrase.

To provide some estimate of the segmentation accuracy, we compute the median-absolute deviation of corresponding labels (relative to the nearest metronome tick).[5] This measures a combination of labelling accuracy and the accuracy with which the subjects speak. The median absolute deviation of estimates for the utterance centre about the median is 2.6% of the utterance length. While it is unclear how much if this is due to labelling and how much is due to speaker variation, we believe it is enough smaller than the median phoneme length, 9.6% of the utterance length, so that can reliably sample data from a point near the centre of each phoneme.

References

Alvey, C., Orphanidou, C., Coleman, J., McIntyre, A., Golding, S., and Kochanski, G. (2008). "image quality in non-gated versus gated reconstruction of tongue motion using magnetic resonance imaging: A comparison using automated image processing". *International Journal of Compter Assisted Radiography and Surgery.* In publication. URL checked 6/2008.

Hornby, A. S. (1974). *Oxford Advanced Learner's Dictionary of Current English.* Oxford University Press, third edition. Available from the Oxford Text Archive, `http://ota.ahds.ac.uk/`, as digitized by Roger Mitton,

[5] There may be a few wild labels where the speech was simply not loud enough to be reliably segmented. To allow for this, we compute the median label position rather than the mean: medians have the desirable property that a single outlying datum has little effect on the result. Since we compute the median position, the median absolute deviation offers a better estimate of the uncertainty of the result than the standard deviation for the same reason: a single wild point will have relatively little effect on either the median or the median absolute deviation, while it would have a large effect on the standard deviation.

Department of Computer Science, Birkbeck College, University of London, Malet Street, London WC1E 7HX June 1992 (URLs checked 5/2008).

Kochanski, G. and Orphanidou, C. (2007). Testing the ecological validity of speech. In Trouvain, J. and Barry, W. J., editors, *Proceedings of the 16th International Congress of Phonetic Sciences.* Conference website at http://www.icphs.de viewed 9/2007.