THE COMPUTATIONAL ARCHITECTURE OF ELIJA: A MODEL OF A YOUNG CHILD THAT LEARNS TO PRONOUNCE

Ian S. Howard[1] *& Piers Messum*[2]

[1]*Computational & Biological Learning Laboratory, University of Cambridge, UK.*
[2]*Centre for Human Communication, University College London, UK.*
ish22@cam.ac.uk

Abstract: We describe the architecture and operation of Elija, a computational infant that learns to pronounce speech sounds. Elija is modelled as an agent who can interact with his environment but who has no *a priori* articulatory or perceptual knowledge of speech. His sensory system responds to touch and acoustic input. He judges the value of action and response using a reward mechanism, and can associate and remember the correspondences between his actions, their reward, and prior and subsequent sensory inputs. Elija first develops the ability to babble using unsupervised learning, which is formulated as an optimization problem. Then he takes advantage of tutored interactions with his caregivers. Such interactions consist of naturalistic exchanges in which the caregivers reformulate Elija's output. He uses these to learn the importance of his productions and this process selects for good productions and discards poor ones. In addition, using associative memory, the reformulations build up a correspondence between his output and adult speech sounds. This leads Elija to develop the ability to imitate words spoken by the caregiver by parsing this input, with a DTW recognizer, in terms of previously heard reformulations which he uses as its templates. He thereby identifies the sequence of motor actions he can perform that his caregiver will take to be equivalent to each word. In this way, Elija is able to learn the pronunciation of novel words.

1 Introduction

Learning to pronounce is conventionally assumed to be an imitative process. In contrast, we describe a computational model that is based on exploration, reinforcement and association. Our model first makes use of unsupervised self-exploration to discover articulations that could underlie simple speech sounds. It then learns to associate its speech productions with adult L1 productions using reformulations generated by the mirroring interactions that occur between a young child and his caregivers. Finally it uses these associations to learn words by imitation. This paper concentrates on the computational implementation of Elija. A fuller discussion of the motivation behind Elija is available in previous work [1-3] and also in a companion paper to the current one, which includes the results obtained by teaching Elija to speak first words in American English, Canadian French and German [4].

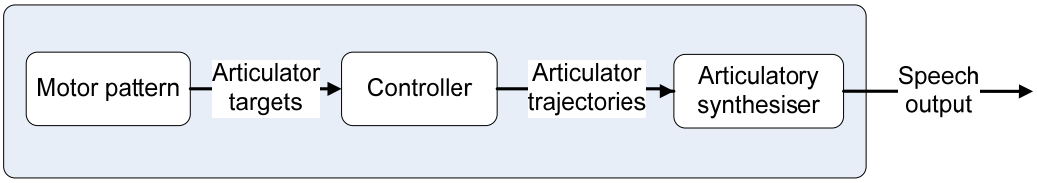

Figure 1 – Elija's motor control system incorporates an articulatory speech synthesiser. A motor pattern is a sequence of vectors of Maeda synthesiser control parameters. These are converted to trajectories by a controller. These resulting sequences of time-varying parameter vectors drive the synthesiser.

2 Computational Architecture

Elija incorporates a speech production capability based on a Maeda articulatory synthesizer [5]. It is driven by a motor system in which representations of motor actions are akin to the gestural score used in the Task Dynamics model [6]. Articulatory targets are used to generate trajectories of the Maeda parameters and this is achieved using a controller to compute their transitions. This flow of information in Elija's motor system is shown in Figure 1. Elija's perceptive system is based on an auditory filter bank front-end [7]. It also implements salience detection as part of a reward mechanism and DTW recognition [8] to enable Elija to discriminate different speech sounds. This is shown in Figure 2.

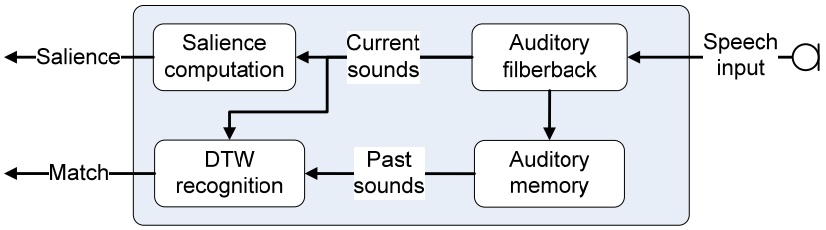

Figure 2– Elija's perceptive system. An auditory filter bank provides input to the salience detector which is used by the reward mechanism, and also to a speech sound recognizer that is based on DTW.

2.1 Maeda articulatory synthesizer

The Maeda articulatory synthesizer [5] uses a 2-dimensional model to represent the cross sectional profile of the vocal tract along the mid-sagittal plane. The parameters in the model were estimated (by Maeda himself) using factor analysis of an x-ray dataset. The dataset consisted of cine-radiographic vocal tract profiles and labiofilm frontal lip shape recordings of 2 female French speakers producing 10 French sentences. The images were recorded at 50 frames per second and in total there were about 1000 frames of data. The vocal tract was divided up into 3 sections – lip opening, principal vocal tract and pharynx. The principal vocal tract was measured in semi polar coordinates, the lips by an elliptical opening and the larynx by its height. A jaw model [9] was then invoked to explain the dataset in terms of the parameters jaw, tongue-body, tongue-tip, lip height, lip width and larynx height. It was assumed that vocal tract shape is determined by a linear combination of the state of these elementary articulators and a directed factor analysis was used to describe vocal tract shape in terms of these parameters. This method allowed the contribution of a particular elementary articulator to be subtracted from the dataset using linear regression, making it possible to explain the input data in terms of the pre-defined elementary articulators. This would not be the case if standard factor analysis had been used, in which case there may be no simple interpretation of the action of the factors. The contributions of control parameters were subtracted in a specific order to find orthogonal parameters. Thus - starting with jaw height - jaw, lip and tongue control parameters were estimated.

In all, 7 articulatory parameters are used to specify vocal tract profile: P1 Jaw position, P2 Tongue dorsum position, P3 Tongue dorsum shape, P4 Tongue apex position, P5 Lip height (aperture), P6 Lip protrusion, P7 Larynx height. In addition, a LF voice source model was added to give control over voiced excitation [10]. This only currently makes use of two parameters: P8 glottal area, and P9 fundamental frequency. In the original VTCALCS implementation a velo-pharyngeal port was added to the basic model and its opening can also be controlled using parameter P10, Nasality. From the vocal tract profile specified by the elementary articulator parameters, an equivalent digital filter is computed and used to filter the excitation from the voice source and noise sources and thereby generate a time waveform output signal. Fricatives are simulated in the model by injecting noise at locations in the vocal tract where turbulent air flow is predicted.

In Elija, the synthesiser used an output sampling rate of 24 kHz. To simulate an infant vocal tract its default physical dimensions, which model an adult female vocal tract, were scaled down by a factor of 0.8. Similarly the mid-range of the fundamental frequency was shifted from 210 Hz to 400 Hz. Although other differences in anatomy exist between infant and adult vocal tracts, it was not necessary to model these for our work because Elija does *not* attempt auditory matching with *his* output. We added proprioceptive feedback of lip and tongue contact, which was generated at times when the vocal tract tube cross-sectional area reached zero. The Maeda synthesizer enabled Elija to produce both oral and nasal sounds. Acoustic output was played to the caregiver from the PC's inboard DAC output via a pair of active loudspeakers.

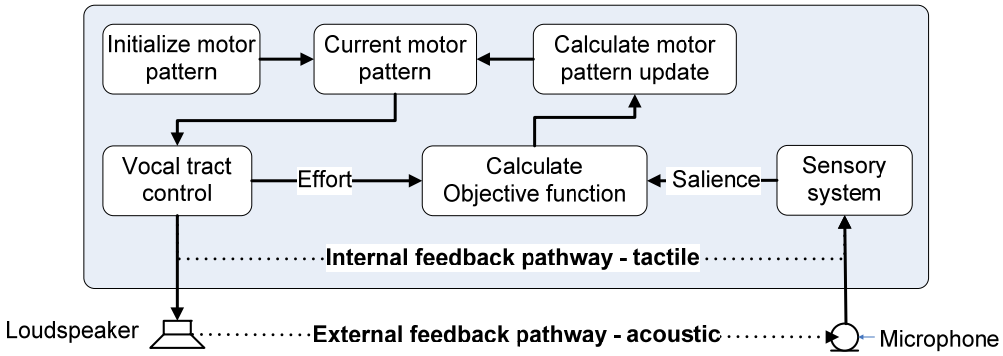

Figure 3 Self organizing discovery of speech sounds using optimization. Firstly the motor pattern is randomly initialized. For the current motor pattern, the objective function computes its utility. Quasi-Newton gradient ascent is performed in motor parameter space to find the "optimal" value of the motor pattern. After several iterations, the resulting motor pattern is recorded in motor memory.

2.2 Motor system

Motor actions in Elija are modelled akin to the gestural score used in the Task Dynamics Model [6] and movement of his articulators between targets is implemented by assuming 2nd order dynamics that follow critically damped trajectories [11].

A motor pattern can be a sequence of up to three vocal tract target positions. Such motor patterns are defined in terms of articulator position vectors. Specifically, the motor patterns consist of the 10 Maeda synthesiser control parameters, the time for which a target is maintained and a transition speed parameter β. A motor pattern generates a vocal action in which the trajectories between targets are determined by articulator dynamics modelled by means of gestural controllers, assuming 2nd order dynamics that are critically damped, leading to movements towards targets without overshoot [11]. Thus, the trajectories of a control parameter is given by:

$$x(t) = x_{endpoint} + \left(\left(x_{startpoint} - x_{endpoint}\right) + \left(x_{startpoint} - x_{endpoint}\right)\beta t + v_0 t\right) e^{-\beta t}$$

where x(t) is the parameter at time t, $x_{startpoint}$t is the starting value, $x_{endpoint}$ is the target value, v_0 is the initial velocity, the constant β is given by the relation $\beta^2 = {}^{k}/_{m}$ where k is the spring constant and m the associated mass of the dynamical system. The value of β associated with the different vocal tract parameters and actions was matched to their dynamic properties. For movements of the articulators during vocalic, sonorant and fricative sound generation, a value of β =40 matches well typical human articulation speeds. However, during plosive sound generation, transitions are much faster due to the build up of air pressure behind the point of vocal tract closure. To account for this phenomenon, transitions following closure have their associated β value increased to 160. This leads to the synthesiser generating more realistic plosive sounds.

2.3 Auditory system

External speech input, which arises from interactions with a caregiver, is digitized using a Rode Podcaster USB microphone. Analysis of Elija's acoustic output by Elija is carried out directly on the digitized signal from the synthesiser (although in principle this could also be achieved by passing acoustic output back from the loudspeaker via the microphone). Elija's auditory system is based on an auditory filter bank.

2.4 DTW speech recognition

Like a real infant, Elija has no *a priori* phonetic knowledge, but is able to discriminate sounds in his environment. We achieve the latter by using a template-based dynamic time warping (DTW) recognizer [8, 12] running with a auditory Gamma tone filter bank front-end [7]. In the final imitation stages of the experiments, the recogniser operates by using previously heard reformulations as its sound templates. In this way, Elija is able to detect speech sounds in incoming speech in terms of those sounds he has heard before. We note that, because we only match caregiver speech with caregiver speech, there is no normalization problem for the classifier to solve, which makes recognition easier. Since words can contain several basic speech sounds concatenated together, a segmentation mechanism is used to present them more individually to the template based recognizer. This requires that the caregiver speaks with pauses between syllables. Segmentation is achieved by detecting regions of silence between sounds on the basis of acoustic power. A calibration phase is used to determine the speech-present threshold.

2.5 Associative memory

When motor patterns are discovered, they are recorded in motor memory. Incoming speech utterances can also be recorded in sensory memory. During the reformulation phase, when Elija generates a speech-like sound, to which his caregiver responds, the caregiver's acoustic response is recorded in sensory memory. In addition, an association is formed between these two memories. During the imitation phase, Elija records input after he is prompted with a keyboard button push.

3 Unsupervised motor action discovery

During their acquisition of speech production, infants progress through several identifiable developmental stages [13]. Elija first discovers potentially useful articulations in an unsupervised manner by finding motor patterns that are solutions to an optimization problem [2]. The objective function of a motor pattern is defined as a sum of its sensory salience, diversity and stability, less the energetic effort involved in action generation. The objective function is given by

$$R = \sum (Salience + Diversity - Sensitivity - Effort)$$

To discover potentially useful articulator configurations, optimization of the objective function is carried out using gradient ascent. The task is to find values of the motor pattern that maximize the reward R. This is achieved by starting from a random initial position and using 3 iterations of a Quasi-Newton gradient ascent algorithm, as implemented by the Matlab function *fmincon*. The discovery of speech sounds using optimization is shown in Figure 3.

In order to increase the variety of available sounds, voicing was explicitly enabled or disabled in each plosive and fricative articulation. Similarly, plosives were generated with or without nasality. In this way, Elija was able to generate distinct articulations useful for the production of a wide range of potential speech sounds. We now describe the individual contributions to the objective (reward) function.

3.1 Somatosensory salience

$$Salience = tactile\ salience + acoustic\ salience$$

Sensory *salience* uses summed low-frequency and summed high-frequency power as features, as well as touch. Elija can selectively focus his attention on these different aspects of sensory feedback by changing their relative contribution in the objective function. Attending to acoustic power at lower frequencies will favour configurations that lead to vocalic sound production, while attending to acoustic output with a dominant high frequency component will favour the discovery of fricatives. Attending to touch will favour configurations where the lips are closed or the tongue makes contact with the teeth or the roof of the mouth.

3.2 Pattern diversity

$$Diversity = tactile\ diversity + acoustic\ diversity$$

The *diversity* term is included in the objective function and assists the formation of a wide range of motor memories. This ensures that Elija performs active learning and explores previously untried articulations. It is computed by comparing the current pattern with all the previously discovered patterns. Using a tactile similarity metric, this leads to the discovery of distinct plosive articulations. Using an acoustic similarity metric, this leads to the discovery of vocalic and fricative sounds that are acoustically distinct. We note that this mechanism has similarities to Liljencrants and Lindblom's Adaptive Dispersion Theory [14].

3.3 Pattern effort

$$Effort = articulatory + voicing\ effort$$

The effort required to make a vocal action makes a negative contribution to reward. This is determined by a combination of the cost of movement, which is calculated as a weighted sum of articulator speed, and the loudness of the voiced excitation. If no penalty is included for voicing, the optimization generally finds a solution with the voicing parameter set to maximum, because this maximizes the sensory salience.

3.4 Pattern stability

$$Sensitivity = \frac{\Delta\ acoustic\ output}{\Delta\ articulatory\ target}$$

Motor pattern stability relates to how a local perturbation affects its corresponding acoustic output. This is an important issue in a real infant because of noise in his motor system and vocal apparatus muscles. More stable articulations will be easier to produce and learn than less stable ones, since they can be repeated without requiring high accuracy movements. Indeed, there is reason to believe that very unstable articulator configurations are not adopted in speech production at all. This issue was previously treated in Steven's Quantal Theory [15] and Gunnilstam's theory of local linearity [16]. Both of these theories hypothesize that preferred regions of articulation exist for use in speech production and that there are, for example, regions of articulator space that provide a natural location for vowel sounds.

To penalize the discovery of vocalic sounds with sensitive underlying articulations, a sensitivity metric was included in the objective function. The sensitivity of the acoustic realization of a given motor pattern was computed by first individually positively perturbing the parameters P1 to P5. In each case the perturbation corresponded to 5% of the full parameter range (i.e. a value of 0.1 was added to each Maeda parameter). All other parameters were set to constant values across all motor patterns analysed to avoid added variability. The output time waveform for the unperturbed motor pattern and for each of the 5 perturbed motor patterns were generated using the Maeda synthesiser and were then analysed using the auditory filter bank. The distance between the auditory representation of each perturbed motor pattern and that of the unperturbed pattern was computed. The overall sensitivity for the given

motor pattern was then taken as the square root of the sum of squares of the 5 components. The insensitive patterns generally had a well defined format structure and sounded like strong vowels. This was not the case for the sensitive patterns which were much less well defined. As well as contributing to the ojjective function, the sensitivity values were used to discard sensitive patterns. In practice, only about 20% of the discovered patterns were retained.

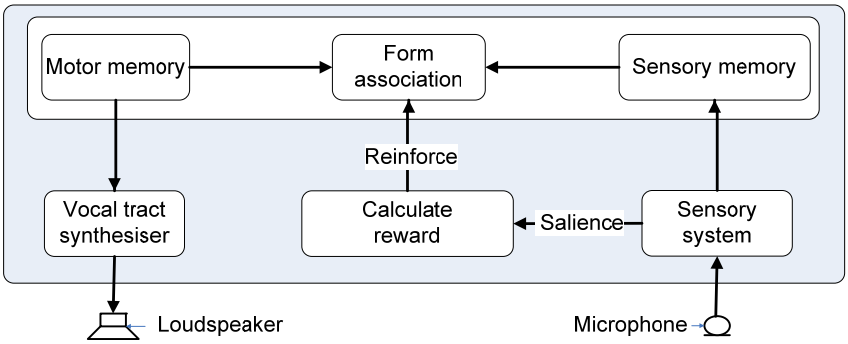

Figure 4 Associations between motor productions and reformulations from caregiver. Elija generates an acoustic output by using previously discovered motor actions. After replay, Elija records any potential response from the caregiver. If the caregiver responds with a reformulation, the presence of her response will generate a reward signal and this will cause Elija to remember both input and output and build an association between them.

3.5 Clustering

For computational reasons, clustering was employed to limit the number of discovered articulations in each type class. Motor pattern clustering was performed using a standard K-means algorithm. Acoustic clustering was performed using a modified version of the standard algorithm that uses DTW as a metric of similarity [1]. Clustering maintained variety but limited redundancy of motor patterns and ensured that at the later stage of the experiment there would be no combinatorial explosion of potential VV, CV and VC configurations. Using acoustic clustering of their sensory consequence, the number of vocalic sounds found was limited to 15. Similarly the number of fricatives was limited to 10. The number of plosives was limited to 15 by clustering on their place of articulation, as determined by somatosensory feedback from the point of closure.

3.6 Combinations of articulations

By concatenating the simple motor patterns discovered by the optimization procedure, it is easy to generate more complex utterances. Thus, after the discovery of Vs and Cs, the recombination process can generate VVs, CVs, VCs and so on.

4 Using caregiver reformulations

Caregiver reformulations

Reformulations of Elija's output by a caregiver reinforce his productions and also provide an adult form of his utterance that the caregiver interprets as equivalent. To ensure Elija can take advantage of reformulations, he is first able to select motor patterns he discovered previously and replay them in turn. Elija is prompted to generate a sound by pressing the keyboard. For a few seconds after each production, Elija records any potential response from the caregiver. When the caregiver reacts vocally with a reformulation, the salience of her response reinforces the motor pattern responsible and also creates an association to the reformulation. This is illustrated in Figure 4.

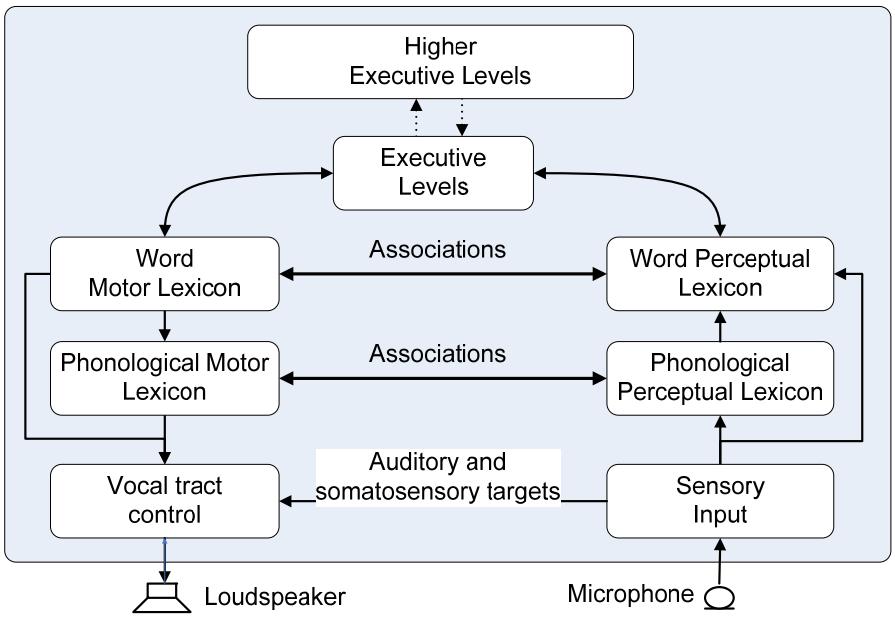

Figure 5 Diagram representing the associations between separate motor and perceptual lexicons. At low levels there can be feedback control to maintain acoustic and somatosensory targets for production. At a phonological level, the reformulation stage builds up associations between motor and acoustic "sound unit" representations. Learning words by imitation builds up associations between motor and acoustic "word unit" representations. The system is administered by executive levels that depend on the task in hand and the rewards associated with the actions and their responses.

5 Learning by imitation

5.1 Parsing input speech

After Elija has learned the associations between his productions and their equivalent adult forms, he can use these adult sounds to analyse incoming speech into component parts. Then he can attempt to generate the word's production by serial imitation of its elemental sounds. To do so, the "reformulations" are used as templates in a DTW speech recognizer. When they are identified, Elija's corresponding motor patterns are also identified. Using this mechanism, Elija can be taught to speak new words by the caregiver. A user interface provides the caregiver with a word from a list. The caregiver then presses a button and speaks the word. Elija then repeats it using imitation and may have up to 4 attempts at recognition, each selectable in the user interface. Acceptance of Elija's response is indicated by the caregiver by pressing a button. Rejection is indicated by pressing a different button. The caregiver may also choose to speak again. An important aspect of this infant caregiver interaction is that they can engage in repetitive loops. A word spoken by the caregiver can be repeated a few times until an appropriate production is performed by Elija or the caregiver chooses to move on. Such a mechanism reduces the effect of any previous incorrect associations. After words have been learned by imitation, Elija has developed both elemental (akin to phonological units) and word motor representations of his productions that are associated to the caregiver's corresponding acoustic elemental (phonological) and word representations of speech. Such associations in Elija's memory and their executive control structures are shown in Figure 5.

Acknowledgments

VTCALCS is a vocal tract model written by Shinji Maeda, modified and distributed by Satrajit Ghosh. Our version has been further extended by Mark Huckvale and ISH. We thank Daniel Wolpert for allowing us to use facilities in the CBL, University of Cambridge.

References

[1] I. S. Howard, and P. Messum, “Modeling the development of pronunciation in infant speech acquisition ” Motor Control, 2011.

[2] I. S. Howard, and P. R. Messum, “A Computational Model of Infant Speech Development,” in XII International Conference "Speech and Computer" (SPECOM'2007). 2007, pp. 756-765

[3] P. R. Messum, “The Role of Imitation in Learning to Pronounce,” PhD., University of London, 2007.

[4] I. S. Howard, and P. Messum, "Modelling caregiver tutored development of pronunciation in a young child," Studientexte zur Sprachkommunikation; ESSV 2011, Aachen, Germany.

[5] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," Speech production and speech modelling, W. J. Hardcastle and A. Marchal, eds., pp. 131-149, Boston: Kluwer Academic Publishers, 1990.

[6] E. Saltzman, and K. Munhall, “A dynamical approach to gestural patterning in speech production,” Ecological Psychology, vol. 1, pp. 333-382, 1989.

[7] M. Slaney, “An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank,” Apple Computer Technical Report #35, 1993.

[8] D. Ellis, "Dynamic Time Warp (DTW) in Matlab: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/," 2003.

[9] B. E. Lindblom, and J. E. Sundberg, “Acoustical consequences of lip, tongue, jaw, and larynx movement,” J Acoust Soc Am, vol. 50, no. 4, pp. 1166-79, Oct, 1971.

[10] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow,” STL-QPSR, vol. 4, pp. 1-13, 1985.

[11] K. L. Markey, “The sensorimotor foundation of phonology; A computational model of early childhood articulatory development,” University of Colorado at Boulder, 1994.

[12] R. Turetsky, and D. Ellis, “Ground-truth transcriptions of real music from force-aligned MIDI syntheses,” 4th International Symposium on Music Information Retrieval ISMIR-03, pp. 135-141, 2003.

[13] D. Oller, The emergence of the speech capacity, Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[14] J. Liljencrants, and B. Lindblom, “Numerical simulation of vowel quality systems: the role of perceptual contrast,” Language vol. 48, pp. 839-862, 1972.

[15] K. N. Stevens, “On the quantal nature of speech,” Journal of Phonetics, vol. 17, pp. 3-46, 1989.

[16] O. Gunnilstam, “The theory of local linearity,” Journal of Phonetics, vol. 2, pp. 91-108, 1974.