Speech Development by Imitation

Björn Breidegard
Certec, Lund Institute of Technology
P.O. Box 118
S-221 00 LUND, Sweden
bjorn@certec.lth.se

Christian Balkenius
Lund University Cognitive Science
Kungshuset, Lundagård
S-223 50 LUND, Sweden
christian.balkenius@lucs.lu.se

Abstract

The Double Cone Model (DCM) is a model of how the brain transforms sensory input to motor commands through successive stages of data compression and expansion. We have tested a subset of the DCM on speech recognition, production and imitation. The experiments show that the DCM is a good candidate for an artificial speech processing system that can develop autonomously. We show that the DCM can learn a repertoire of speech sounds by listening to speech input. It is also able to link the individual elements of speech to sequences that can be recognized or reproduced, thus allowing the system to imitate spoken language.

1. Introduction

A robot that gradually develops through interaction with humans can not be built with a limited fixed vocabulary or a fixed mode of articulation. An open ended system should have the ability to learn both the pronunciation and the use of new words in the appropriate linguistic and conceptual context through interaction with a human caregiver (Varshavskaya, 2002). In addition, it should use speech to communicate emotions (Breazeal, 2001, Murray and Arnott, 1993).

Learning to produce and recognize speech sounds can be done through imitation (Doupe and Kuhl, 1999). This imitation can be either externally or internally controlled. When the imitation is internally controlled, the robot attempts to imitate its own speech sounds. An example would be babbling where the machine produces a sound at random that is subsequently recognized by its auditory system. As a consequence, it attempts to produce that sound again through its speech organ. This will result in several repetitions of the same or similar speech sounds. This is a form of circular reaction as described by Piaget (1950) which results in self-imitation.

When the imitation is externally controlled, the speech sounds of the human caregiver are recognized by the robot which attempts to reproduce those sounds. This repetition can either result in the direct imitation of an individual phoneme or in longer sequences of speech. The complexity of this imitation will depend on how advanced methods are used for temporal prediction and sequence processing.

The learning depends on two types of associations. On one hand, the auditory input must be associated with the appropriate commands to the speech system, and on the other, the robot must have the ability to recognize and produce sequences of sounds.

The goal of our research is to design a (biologically inspired) speech imitator based on the Double Cone Model (Breidegard, 2000). The speech imitator is modelled as an executable computer program with interaction and auditive and visual feedback. The program receives real-time sound input and produces real-time sound output.

There are several aims of this research. The first is to imitate child language acquisition. We want to investigate how the ability to understand and produce speech sounds can develop without initially assuming such an ability. This approach can be contrasted with the work of de Boer (2000) and Oudeyer (2002), for example, where the goal is to understand the emergence of the phonetic system itself.

A second goal is to simulate impairments in language acquisition many of which relates to phonological processing (Tallal, et al. 1998). In the future, we hope to compare results from the computer model with the developing child by introducing lesions and other disturbances in the model. As a result, we hope it will be possible to design help and pedagogical methods for children with speech learning impairments. By modeling the different types of aphasia we hope to make conclusions about new methods for pedagogy and training of patients with aphasia.

The third goal is to develop a novel type of system for speech recognition and synthesis based on biological principles. Current speech synthesis methods still require further development before they reach the intelligibility of human speech (Venkatagiri, 2003). The speech recognition systems that are available are even more limited. By letting the system de-

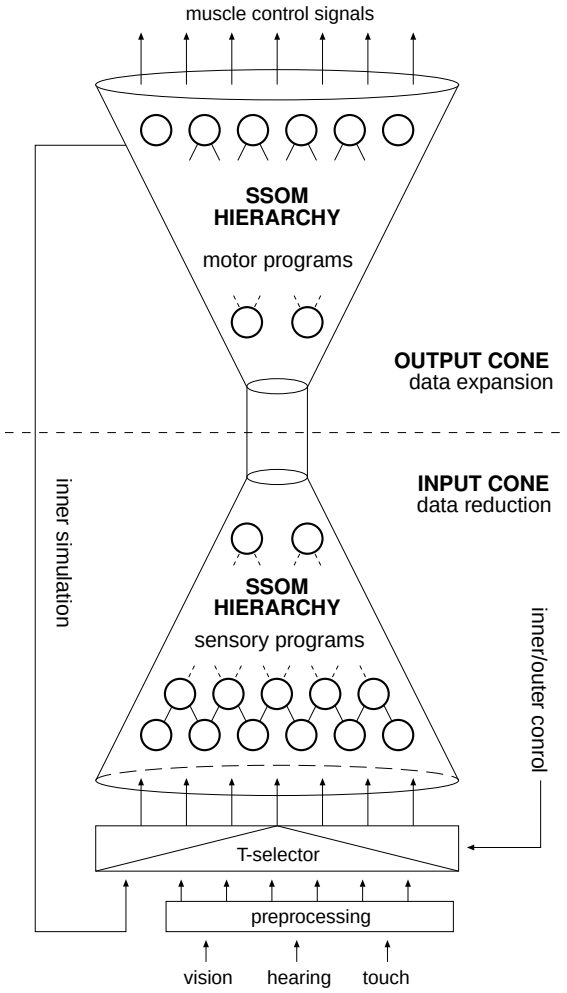

FIGURE 1: *The Double Cone Model*

velop its own speech representation based on its own motor output, we believe it will be possible to design a more successful speech recognition and synthesis system than the current state-of-the-art.

Below we describe the Double Cone Model and shows how it can model the phonological loop in imitation. The goal is to test the basic assumption that the model can learn to imitate speech sequences. We also discuss how the system will be extended in the future to a more complete model of vocal learning.

2. The Double Cone Model

The Double Cone Model (DCM) is a computational model of the human brain (Fig. 1). It is based on two data processing cones that are intended to model the functional role of a number of hierarchically organized cortical regions (Breidegard, 2000).

The input cone performs hierarchical data reduction (lossy compression). The role of the input cone is to reduce the high data rate of the incoming sensory signals. This task is performed in parallel in different parts of the input cone.

The hierarchical processing allows complex sensory programs to be constructed. It also allows for sensory fusion between different modalities such as vision and hearing. The mechanisms can also be used to integrate several sensory codes within the same modality such as several visual submodalities or somatosensory codes for body posture and gaze direction. The highest level of cognitive processing occurs where the tops of the cones meet halfway between input and output where the data rate is the lowest.

The output cone operates in the reverse way. It expands the signals from the input cone to generate parallel muscle control signals with a high data rate. The expansion occurs both in time and space as the output cone will both extend the number of parallel signals and the bandwidth of each signal. This eventually results in the formation of motor programs in the output cone.

Processing in the Double Cone Model is similar to the perception-action cycle described by Fuster (1995) where the input cone corresponds to the sensory hierarchy and the output cone corresponds to the motor hierarchy (See also Grossberg, 1986).

The main feedback loop is via the T-selector. A control signal Inner/Outer determines the proportions between sensory data from the outer world and mental data from the inner world. This makes it possible for the model to be driven either by external or internal perceptions. This also introduces the possibility of performing inner simulations before making movements in the external world. These inner simulations may be the origin of thought (Hesslow, 2002).

Both the input and output cones consist of Sequential Self-Organizing Maps (SSOM), which are intended as models of the cortical feature maps found in the human brain (Gazzaniga, Ivry and Mangun, 2002). The SOM architecture has previously been shown to map the phonetic space on a two-dimensional surface through a self-organizing process (Kohonen, 1988). The SSOM is an extension of the original SOM architecture (Kohonen, 1997) that has been enhanced with a capability to link SOM elements into unique sequences (Fig. 2). In the output cone, the function of the linking mechanism together with the SOM categories are similar to the avalanche network proposed by Grossberg (1986).

First, the SOM is trained until all the SOM elements become specialists on different regions of the input space, e.g. it may develop partitions for the different vowels in a speech signal. In this case, each node in the SOM corresponds to a snapshot of the sound spectrum. Second, the SSOM is trained to recognize (or generate) small sequences of those features, e.g. the syllables of speech. These sequences are constructed by chaining those SOM elements that best match the sequential features in the in-

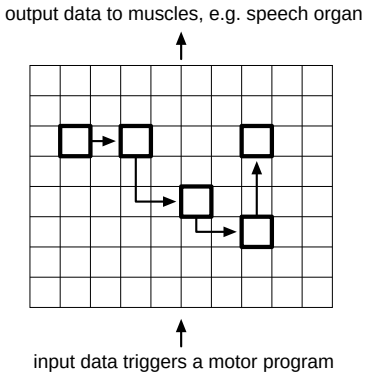

FIGURE 2: *The Sequential Self-Organizing Map (SSOM) consists of a SOM where each SOM element (by training) has become a specialist on some aspect of the input signal. SOM element sequences can also be used to generate motor programs, e.g. to generate speech.*

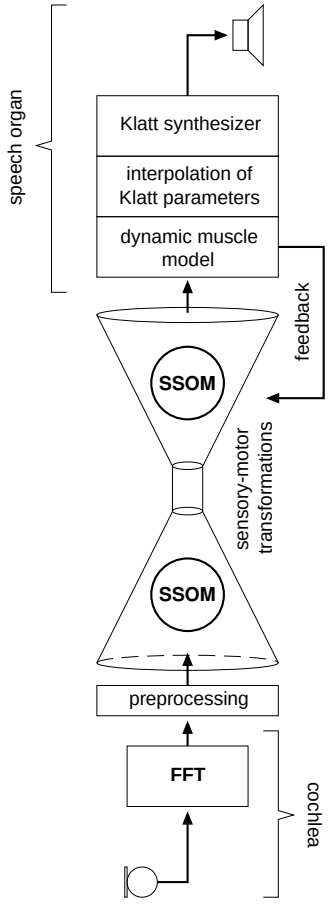

FIGURE 3: *The Phonological Loop is the auditive subset of the Double Cone Model. The ear (cochlea) is modelled by a microphone and Fourier Transform which gives the spectral components of each time sample (snapshots). The speech organ is modelled by a dynamic muscle model that is used to set the parameters for a Klatt synthesizer which generates the speech signal to the loudspeaker. Here the input cone and the output cone consist of one SSOM each.*

put signal. Once a sequence has been learned, it can be triggered by a part of the sequence in the input code and then produced automatically in the output cone.

3. Modeling the Phonological Loop

Our goal is to apply the Double Cone Model to the phonological loop (See Gazzaniga, Ivry and Mangun, 2002). The system that is currently being implemented is shown in (Fig. 3).

To model the auditory input we use the Fast Fourier Transform (FFT) and filtering in space and time as a simple model of the cochlea. This is certainly a simplification compared to the more advanced cochlear models that have been suggested (Roberts, 2002, Nobili, Mammano and Ashmore, 1998), but has the important advantage that it can be run in real time. The output from the FFT preprocessing are time samples, or snapshots, of the incoming speech signal. These snapshots are subsequently learned by the nodes of the input SOM.

In the final implementation, the output SSOM will be connected to a model of a muscle controlled speech organ. This speech organ consists of three parts:

- A dynamic muscle model that adds physical and timing constraints similar to the human speech organ.
- An interpolation model that calculates the large number of parameters needed to control the Klatt synthesizer based on a number of calibration points for speech sounds (e.g. vowels).
- The Klatt synthesizer which produces the sound fed to the loudspeaker.

The output SSOM will learn the possible patterns of muscle activations through feedback from the muscle model about the actual movement made. That is, even when the SSOM is initially producing noise as output, the muscle model will transform that noise into a physically possible articulation which will be learned by the output SSOM. It will thus gradually be tuned to the muscle model.

In the simulations described below, this speech organ was not used since we wanted to investigate how well the model could imitate speech independently of the model of the speech organ. In this implementation, the same spectrum snapshots are used by both the input SSOM and the output SSOM. It was thus possible to use a single SSOM for both input and output (Fig. 4).

4. Materials and Methods

The phonological loop in the Double Cone Model has been implemented as a computer program. A stan-

dard PC with sound card, a microphone and loudspeaker was used to run the model. The program was developed using Microsoft Visual C++.

The real-time simulator is highly interactive with auditive and visual feedback and allows easy exploration of the different codings in the double cone. By clicking SOM elements with the computer mouse (i. e. computerized Penfield probing) it is possible to visualize and hear the SOM elements constituting different SSOM sequences. When a SOM element is clicked, it triggers the imitation and starts the sound output. Depending on which SOM element is clicked, the part of the speech starting with this SOM element is heard. The first element in the sequence is highlighted with a special color and by clicking on it, the whole sequence will be produced.

In the experiments described below we used the simplest possible Double Cone Model where the input cone and the output cone consist of the same SSOM (Fig. 4). This is possible due to the choice of data representation. The input data representation from the 'cochlea' are the spectral components obtained by the Fourier transform (FFT). Consequently, each SOM element becomes a specialist on some aspect of the speech signal and a sequence of SOM elements represent some perceived speech. This sequence also constitutes a motor program. The same sequence of SOM elements are used both to recognize a speech sequence and to produce that sequence by feeding the Fourier spectrum coefficients for each node to the speech organ. In this initial implementation, the speech organ was modeled as the Inverse Fast Fourier Transform (IFFT). This converts the speech data to a time-domain signal that drives the loudspeaker. In the current implementation, each sequence of input features allocates unique SOM elements. These SOM elements can thus not be used in other chains. This implementation has the advantage that it is easy to understand and control but has the limitation that the memory for sequences will eventually be used up unless older sequences are removed or forgotten.

The length of the Fourier transform is 512 and each snapshot is the mean of a time slot of 46.4 ms. The output sound, the imitated speech, is obtained by concatenating snapshots to an auditive sequence. All processing takes place in real time.

5. Experiment 1

In the first experiment we tested the ability of the model to imitate speech without first being trained on any speech. All SSOM elements were randomly initialized and there were no sequences in the SSOM. Since the model had never heard any speech sounds, the SSOM showed no form of partitioning or spatial ordering – all SOM elements are specialists on different random aspects of the input signal. This corresponds to a developmental stage where the child has not yet learned to recognize speech sounds or to produce any sound of its own.

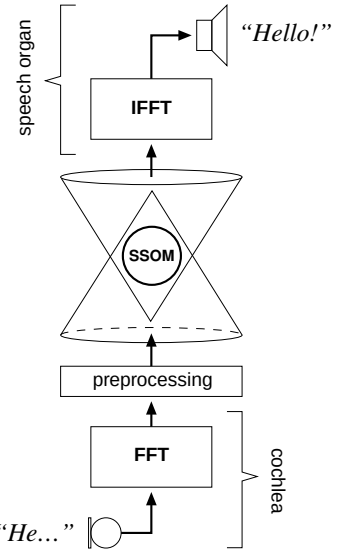

FIGURE 4: *The simplified Double Cone Model. The input cone and the output cone are the same SSOM. This is possible due to the choice of data representation: both input and output data are spectral components. The speech organ here consists of an Inverse Fourier Transform driving the loudspeaker.*

Imitation was tested by giving spoken sentences as input. Interestingly, this immature machine is able to learn to imitate speech in a reproducible way, but a human listener is not able to understand the imitation. When exposed to speech sounds, the model will allocate best-fit SOM elements and create a sequence. When the imitation starts the machine will 'say' the sequence of SOM element features as an auditive sentence. Although we will not hear any intelligible speech, the length of imitation is exactly the same as the original speech.

6. Experiment 2

In the next experiment, the double cone was trained on speech input in two ways. With training 1, three Swedish sentences were repeated for 50 minutes while the SSOM was in learning mode. During this time, each sentence was presented to the system 500 times. With training 2, the system listened to a chapter from an Astrid Lindgren book for 50 minutes. With this training, the material was not repeated. The ability of the system to repeat the three sentences used in training 1 as well as its ability to generalize to a new sentence was subsequently tested with the two types of training.

After the 50 minute training, the SSOM elements developed sensitivity to the different features in the speech input (Fig. 5). The thick grid of lines in-

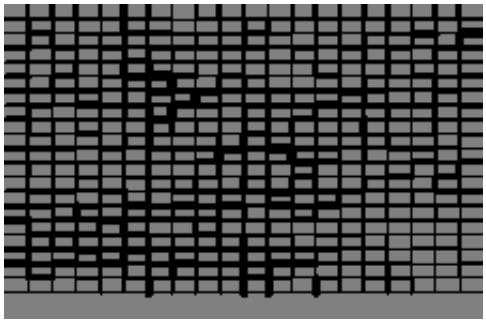

FIGURE 5: *The SSOM has been trained with three Swedish sentences. The SSOM shows some organization.*

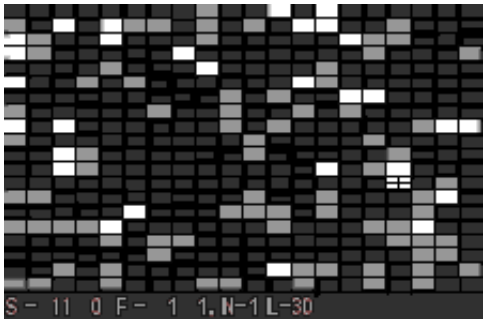

FIGURE 6: *The SSOM has learnt to imitate the three Swedish sentences. The white and light gray elements are the allocated SOM elements for these sentences. The white elements are the sequence "Jag är hungrig" ("I am hungry" in Swedish). By probing other SOM elements, the other sentences are imitated.*

dicates by thickness the distance between neighboring SOM elements. With thinner lines between the nodes, the neighboring SOM elements detect similar features in the speech signal. Probing different SOM elements with the mouse give sounds which bare a resemblance to speech sounds and not only noise as in the first experiment.

The machine is now prepared to learn speech sequences – the mature model can recognize and imitate speech snapshots. Fig. 6 shows the allocated SOM elements for the three Swedish sentences. The highlighted elements are the allocated SOM elements. The white elements are included in the sequence "I am hungry" in Swedish. By probing other SOM elements, the other sentences are imitated. The sonogram (speech spectrogram) in Fig. 7 shows the third sentence: "The square root of nine is three" (in Swedish).

The imitation of the three sentences used in training 1 was tested after training 1 and 2. Regardless of the material used during training, the imitation of the three sentences could easily be recognized (if the sentences were already known), but as expected, the system trained on these sentences (training 1) performed better than the system trained on a book chapter (training 2). This was particularly the case

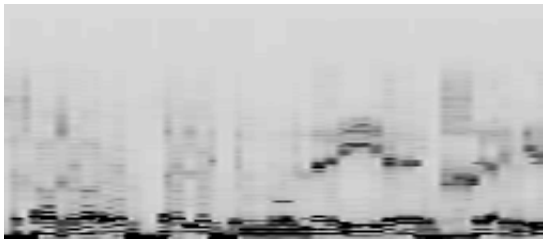

FIGURE 7: *The sonogram (speech spectrogram) for the sentence: "Roten ur nio är tre" ("The square root of nine is three" in Swedish).*

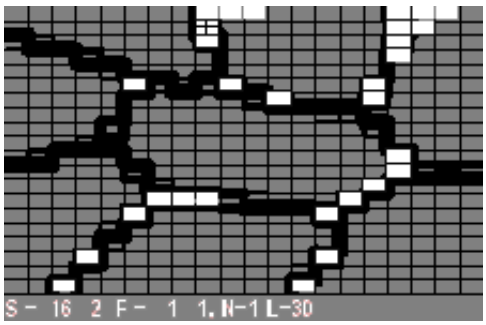

FIGURE 8: *The SSOM was trained with the C major scale of sine tones. The SSOM shows a very specialized SOM organization with eight partitions – one for each tone in the scale. Most SOM elements within a partition are very similar (due to the lack of variation in the rigid sine tones).*

in the high frequency regions.

In addition, we tested the ability of the model to reproduce a new sentence not present in either training condition ("Today I'm really happy" in Swedish). Now the imitation was much better when the model was trained with the more variable material (training 2) than when trained on only three sentences (training 1). This shows that although much repetition improves performance on the trained speech sequences, better generalization is obtained with a more natural and varied training material. Interestingly, the prosody of the novel sentence is imitated in both conditions although the one trained with the story (training 2) performs better.

7. Experiment 3

In experiment 3, the machine had never heard any speech sounds. It had only heard a C major scale of sine tones. It was fully capable of imitating the Swedish sentences, but the imitation was now a cascade of flute-like tones. Fig. 8 shows a very specialized SOM organization with eight partitions and most SOM elements within a partition are very similar (due to no variation in the rigid sine tones). Training on piano tones would also have shown eight partitions, but they would not have been so distinct

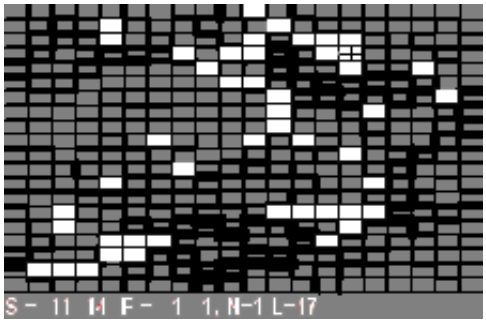

FIGURE 9: *The SSOM has been trained with with a Chopin piano piece. The imitation is not very speech-like – it is more piano-like. But sounds reminding of speech can actually be heard.*

since the spectral complexity and time variation of the tones are higher.

8. Experiment 4

In the final experiment, we tested the ability of the model to use a more complex auditory material to produce speech. The machine was nurtured with a Chopin piano piece repeted many times. Here the imitation is of course not very speech-like – it is more piano-like, but some sounds resembling speech can actually be heard. Fig. 9 shows the SOM organization resulting from repeated listening to a Chopin piano piece. This illustrates that the model is able to use whatever training it has received in an attempt to reproduce speech.

9. Discussion

The Double Cone Model is a model of how information is processed in the brain from sensory organs to motor output. The model was applied to speech processing and it was shown that it is able to reduce the continuous speech input to a sequence of speech categories that can be recognized and reproduced.

Comparing the results of experiment 1 and 2 shows that initial training on speech signals is necessary to produce intelligible speech output. In the language acquisition literature, it has often been suggested that humans must be born with an innate ability to recognize and produce the sounds of human languages (Gibson and Spelke, 1983). Although this may be the case, it is also possible that the sound of the mother's voice has been learned prior to birth, but as in the model, some tuning toward speech sounds is necessary before recognizable speech can be produced. It has been suggested that infants are initially sensitive to all the speech sounds of the world but gradually tune in to the sounds of their specific language (Werker and Tees, 1999).

Experiments 3 and 4 show that the model is able to make use of whatever sound it has been trained on to attempt to produce speech. The categories formed in the input cone are templates that are matched to a spectral representation of the incoming speech. It was recently shown that template based methods can be as good as a human listener in recognizing speech sounds (Hillenbrand and Houde, 2003). When trained on sinusoidal tones or piano music it will attempt to use those sounds to imitate human speech. This result is relevant to the difficulties most of us face in learning a second language. Instead of adapting our speech to the new language, the sounds learned from our first language are used but in new combinations.

Although the model is able to imitate speech, there are two main limitations. The first is that the sequence mechanism added to the original SOM model is too simple for a larger amount of speech material. In the current version of the model, each SOM element can only be included in a single sequence. This was sufficient for the experiments performed, but this mechanism must be extended in the future with contextual control of the learned sequences (Grossberg, 1986, Balkenius and Morén, 2000).

The second limitation is the speech synthesis part. The use of the inverse Fourier transform made it easy to investigate the model, but did not produce very good speech. The main reason for this is that the sequence of spectra produced were not smoothly joined which resulted in audible clicks between the segments. Although it would be possible to overcome this limitation, we have instead aimed at a more interesting solution. We are now developing a muscle controlled speech organ – a virtual tongue. The backend speech synthesizer is the Klatt formant synthesizer (Klatt, 1980). It is controlled by 40 independent parameters that should be updated at least each 5th ms. In this type of synthesizer a wide-spectrum excitation signal, that can be voiced with a fundamental frequency or unvoiced noise, excites a set of resonators (formants). These formants reflect the resonators created within the human speech organ and its parameters (e.g. center frequency) vary with the tongue position, openness of the mouth and so on (Ladefoged 1962, 2001). Although the intelligibility of this type of formant synthesis is not as good as methods based on concatenation of speech segments (Venkatagiri, 2003), it allows for much easier modulation of prosody and emotional content (Carlson, 1991, Braezeal, 2001).

The imitatation will be performed by a Double Cone Model consisting of one sensory SSOM and one motor SSOM controlling the speech organ (See Fig. 3). In this extended model, it will be necessary for the model to convert between a sensory and a motor code thus producing a more realistic phonological loop (cf Morasso, Sanguineti and Frisone, 2001). The model will be trained by letting it listen to its own

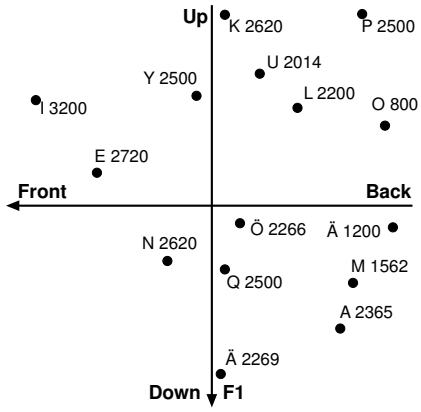

FIGURE 10: *The tongue position diagram for the virtual tongue. The numbers represent the frequency of the third formant.*

spontaneously produced speech sounds and tune it to perceived speech (cf. Holmes and Pearce, 1990).

We will introduce constraints on the 40 parameters that will produce human speech sounds and hopefully babbling when set randomly (Stevens and Bickley, 1991). With all 40 parameters, most settings do not produce speech sounds. A muscular interface (the dynamic muscle model) will be placed in front of the Klatt synthesizer to obtain a speech organ with physical and timing constraints similar to the human speech organ. Among the parameters are tongue positions: front-back, high-low and in our model they approximately represent the first two formants F1 and F2. Other important parameters are the amplitudes, frequency and proportions of the voiced/unvoiced source. This approach agrees with the view that the phonemes of natural languages are based on simple (and possibly binary) motor parameters (Chomsky and Halle, 1968, Stevens, 2002), though the innateness of such features has been questioned (Oudeyer, 2002).

The third formant, F3, is important for speech intelligibility, but we have found no simple and obvious way to control it programmatically. Also higher formants, and the relative amplitudes and the bandwidths for the formants are important to increase speech quality. We have chosen to interpolate them from tongue position. This is obtained by calibrating the tongue position diagram (Fig. 10) with these values for all vowels (and also for many consonants). For example, if we move the tongue from the sound I to E, F3 is interpolated from their calibrated values. This interpolation will expand the few, and constrained, outputs from the dynamic muscle model to the many parameters that control the final sound-producing Klatt synthesizer. This idea is related to the suggestion that speech sounds can be coded in terms of acoustic landmarks (Stevens, 2002).

Our hope is that this speech organ will be of high quality, if we succeed in controlling it accurately. A main problem with formant synthesizers is to control the parameters well and sufficiently often. Our challenge will be to get the motor SSOM to produce appropriate muscle control signals by learning this through imitation (and more imitation to improve the motor programs).

We have tested the virtual tongue by interactive control of the parameters. Trimming this control has shown that intelligibility is much increased. The parameters we programmatically control are: fundamental frequency, amplitude and proportions of voiced/unvoiced source, tongue position, nasal formant, speech segment time and interpolation time.

With zero interpolation time, the tongue is indefinitely fast and the sound is not very good. By requiring a time to move from one position to another, the quality increases. A distinctness in the speech appears. We obtain coarticulation and have also been able to produce Swedish diphthongs and triphthongs. By controlling the glottis frequency, F0, over a word we have also increased its intelligibility. This implies that better control can yield very good speech quality with the Klatt synthesizer. However, a general mechanism to produce the sequence of muscle control parameters instead of hand-coded control is needed. In the near future, we will do this with the two SSOM Double Cone Model (Fig. 3).

The inner simulation feedback (Fig. 1) will eventually be used for short term memory, and to convert between short term and long term memory. The double cone will probably be expanded with more feedback paths and selectors, e.g. the T-selector will consist of many parallel selectors with different control signals.

References

Balkenius, C. Morén, J. (2000) A Computational Model of Context Processing, In J-A Meyer, A Berthoz, D Floreano, H. L. Roitblat, S. W. Wilson, (Eds.) *From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behaviour*, (pp. 256-265), Cambridge, MA: The MIT Press.

de Boer, B. (2000) Emergence of vowel systems through self-organisation. *AI Communications 13*, 27-39

Breazeal, C. (2001). Emotive Qualities in Robot Speech, *In Proc of the international conference on intelligent robots and systems*, Maui, Hawaii.

Breidegard, B. (2000). *En datorexekverbar modell för lärande.* Licentiate Thesis. Certec, LTH. NR 1:2000.

Carlson, R. (1991), Synthesis: modeling variability and constraints, *Keynote paper at European Con-*

ference on Speech Communication and Technology 91, Genova, Italy.

Chomsky, N. and Halle, M. (1968) *The Sound Pattern of English*. New York: Harper and Row

Doupe, A. J. and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms*Annu. Rev. Neurosci., 22*, 567–631.

Fuster, J. M. (1995). *Memory in the cerebral cortex*, Cambridge, MA: MIT Press.

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2002). *Cognitive Neuroscience*, New York: W. W. Norton & Company.

Gibson, E. J., and Spelke, E. S. (1983). The development of perception. In J. H. Flavell and E. M. Markman (Eds.), *Handbook of child psychology: Cognitive Development*. New York: Wiley.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: speech, language, and motor control. In Schwab, E. C. and Nusbaum, H. C. (Eds.) *Pattern recognition by humans and machines, vol 1: Speech perception*, Academic Press.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception, *Trends in Cognitive Sciences, (6), 6*, 242–247.

Hillenbrand, J. M.and Houde, R. A. (2003) A narrow band pattern-matching moddel of vowel perception. *Journal of the Acoustic Society of America, 113 (2)*, 1044-1055.

Holmes, W. J. and Pearce, D. J. B. (1990) Automatic derivation of Proc segment models for synthesis-by-rule. *ESCA Workshop on Speech Synthesis*, Autrans, France.

Klatt , D. H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. AM.*, Vol 67, 1980, pp. 971-995.

Kohonen, T. (1997). *Self-Organizing Maps*. Second Edition. Berlin: Springer-Verlag.

Kohonen, T. (1988). The 'neural' phonetic typewriter, *IEEE Computer Magazine, 21*, 11-22.

Ladefoged, P. (1962). *Elements of Acoustic Phonetics*. Second Edition. The University of Chicago Press.

Ladefoged, P. (2001). *A Course in Phonetics*. Fourth Edition. Heinle & Heinle.

Morasso, P., Sanguineti, V. and Frisone, F. (2001). Cortical maps as topology-representing neural networks applied to motor control: articulatory speech synthesis. In Mastebroek, H. A. K. and Vos, J. E. (Eds.) *Plausible neural networks for biological modelling*. Dordrecht: Kluwer.

Murray, I. and Arnott, L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, *J Acoust Society America, 93, (2)*, 1097-1108.

Nobili, R, Mammano F and Ashmore. J.F. (1998). How well do we understand the cochlea? *Trends in Neural Science, 21*, 159-167.

Oudeyer, P-y. (2002). Phonemic coding might result from sensory-motor coupling dynamics. In Hallam, B., Floreano, D., Hallam, J., Hayes, G., and Meyer, J-A. (Eds) *From animals to animats 7*. Cambridge, MA: MIT Press.

Piaget, J. (1950). *The Psychology of Intelligence*. London: Routledge & Kegan Paul.

Roberts, D. (2002). *Signals and Perception: The Fundamentals of Human Sensation*. Palgrave–Macmillan.

Stevens, K.N. and C.A. Bickley (1991) Constraints among parameters simplify control of Klatt formant synthesizer, *J Phonetics 19*, 161-174.

Stevens, K. N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features, *Journal of the Acoustic Society of America, 111 (4)*, 1972-1891.

Tallal, P., Merzenich, M. M., Miller, S. and Jenkins, W. (1998). Language learning impairments: integrating basic science, technology, and remediation. *Exp Brain Res 123*, 210–219

Varshavskaya, P. (2002) Behavior-Based Early Language Development on a Humanoid Robot, In Prince , C., G., Demiris, Y., Marom, Y., Kozima, H., Balkenius, C. (Eds.) *Proceedings of the Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Lund University Cognitive Studies, 94.

Venkatagiri, H. S. (2003) Segmental intelligibility of four currently used text-to-speech synthesis methods, *Journal of the Acoustic Society of America, 113 (4)*, 2095-2104.

Werker, J. F. and Tees, R. C. (1999) Influences on infant speech processing: toward a new synthesis, *Annu. Rev. Psychol. 50*, 509–535.