Concatenative Speech Synthesis with Articulatory Kinematics obtained via Three-Dimensional Electro-Magnetic Articulography

Hartmut R. Pfitzinger

Inst. of Phonetics and Speech Communication, Ludwig-Maximilians-Univ., 80799 Munich, Germany, Email: hpt@phonetik.uni-muenchen.de

Introduction

Currently, commercial speech synthesis systems are based on concatenation of pre-recorded stretches of speech. This technique yields a highly satisfying speech quality and therefore seems to be an optimal solution for the problem of automatically generating utterances. Nevertheless, the central question of distance and similarity between consecutive stretches of speech still poses problems in automatic segment selection, even though it has been extensively investigated from the acoustic-perceptual perspective. Our idea is that kinematic measurements by electro-magnetic articulography (EMA) are adequate to quantify the amount of discontinuity at unit boundaries since articulatory kinematics in natural speech are always continuous, and reflect dynamic and coarticulatory effects. Therefore they might also be appropriate to predict perceptual distance which is more important than the acoustic distance in concatenative speech synthesis [3].

Method

This hypothesis is tested by replacing the acoustically-based join cost function with our new articulatory-based version in a standard concatenative speech synthesizer. An articulatory and acoustic spoken language resource was needed. The tongue as well as the velum are very important articulators, and only few methods are available to reliably track their movements during speech. EMA with a high frame-rate of 200 Hz was favoured because the current frame-rate of real-time MRI is still too small to appropriately capture the kinematics of natural speech, and x-ray has adverse health effects, especially when recording a large single-speaker database.

Additionally there is a number of obvious advantages of the 3D-EMA system (Fig. 1): *i)* the subject has the freedom to move the head and even the body, *ii)* no helmet is necessary to be worn, and *iii)* the orientation of the sensors is not restricted. Less obvious benefits are: *i)* it detects movements of the tongue which do not change its shape and are invisible to cine-MRI (when there is no muscle texture), *ii)* it provides information on sensor orientation (two of the three possible angles), *iii)* it allows to attach sensors out of the mid-sagittal plane to the left or right to measure the lateral tongue activity (this also avoids collision of two sensors, e.g. in the case of upper/lower lip sensors or velum/tongue-back sensors), and *iv)* lip activity can also easily be recorded since pellets can be attached e.g. to the corners of the mouth.

A New Articulatory Speech Synthesis Database

The text to be read was designed considering the question of how to evaluate final synthesis results. Therefore an already existing speech database containing 200.000 word tokens, 20.000 word types, and 23 hours of continuously read speech of a professional speaker served as a reference. The new synthesis system was intended to be able to generate any utterance of this corpus on the basis of a Greedy condensed subset thus enabling the comparison of synthesized and originally spoken utterances. The Greedy algorithm was tuned to control coverage of phonetically rich words and rely on getting high frequency words as a by-product.

In September 2003 we recorded a large single-speaker read-speech corpus using the new three-dimensional articulograph *AG-500* (Carstens Medizinelektronik) [6]. We used 12 sensors (Fig. 2), a sampling rate of 200 Hz, and the transmitter frequencies 7.5, 8.7, 9.9, 11.1, 12.3 and 13.5 kHz. The acoustic speech signal was recorded with a *Sennheiser MKH-20* microphone (which is almost not influenced by electro-magnetic fields) in a distance of 40 cm to the mouth, and a *John Hardy M-1* high-quality pre-amplifier (F_S=48 kHz, 16 Bit). The transmitter frequencies are audible and consequently acoustically recorded in any quiet environment. To remove them from the speech signal a time-subtraction method was applied which is based on an artifact-free hum-removal technique [2].

In a single recording session of two hours, 40 minutes of running speech and articulatory data have been recorded. An additional recording session is not feasible since it is not possible to glue all sensors at exactly the same positions and ori-

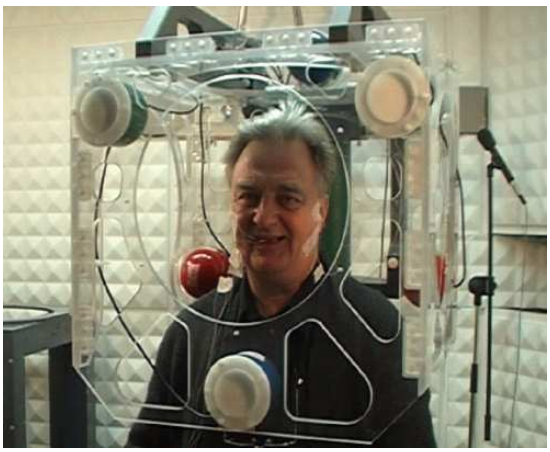

Figure 1: The three-dimensional articulograph *AG-500*

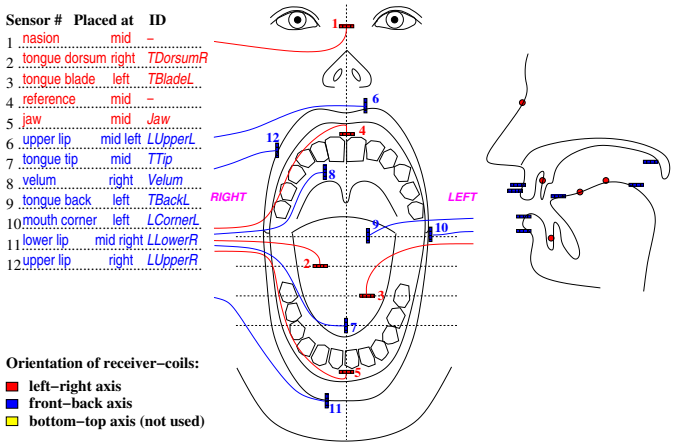

Figure 2: Sensor positions, orientations, and identifiers.

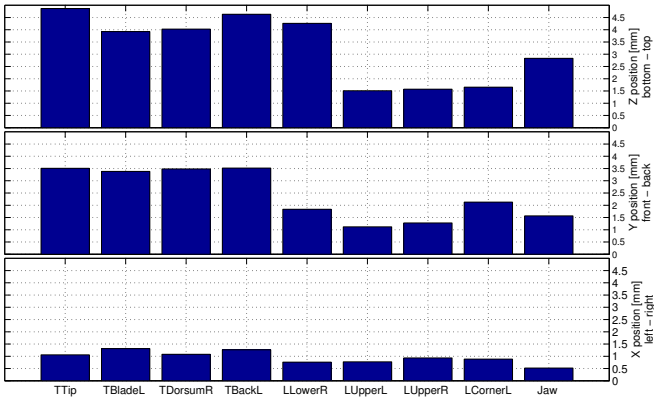

Figure 3: Standard deviations of the sensor X-, Y-, and Z-positions.

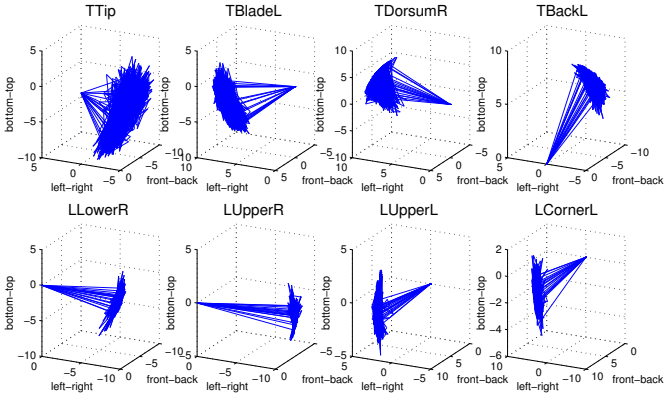

Figure 4: Variations of the orientations of the tongue and lip sensors.

entations as in the first session, which would be necessary to obtain compatible measurements.

The articulatory synthesis database contains 6.400 word tokens, 1.300 word types, 30.000 phone tokens, and 11.400 syllable tokens. Two of the 12 sensors were needed as reference positions for subtracting the head movements (sensors 1 and 4, Fig. 2). Unfortunately, sensor 8 which was attached to the velum came off very soon so that only a minute of velum data was reliably recorded. Thus nine sensors remained to further analysis. First synthesis results were presented in 2003 [4].

First Experiment

The question motivating the first experiment was: What is the impact of acoustic and perceptual discontinuities introduced at unit boundaries by concatenative speech synthesis on the corresponding articulatory kinematics? To answer this question the kinematic variability of adjacent speech units in natural speech has been measured at unit boundaries and described by Gaussian distributions of positions, velocities, and accelerations of articulators. Fig. 3 shows the standard deviations of the X-, Y-, and Z-positions of all sensors while Fig. 4 shows the variability of sensor orientations.

Then, a standard concatenative speech synthesis system [3] was used to systematically create new utterances with more or less obvious acoustic and perceptual discontinuities. Finally, the mismatches between articulatory trajectories of concatenated units were quantified and compared with the Gaussian distributions. The results indicate that trajectory mismatches deteriorate the acoustic and the perceptual mismatch much more than a deviation of a concatenated trajectory from its mean. Presumably, a deviation of a continuous trajectory from the target trajectory is easier to accept than a mismatch between trajectory ends of adjacent concatenation units.

Second Experiment

The second experiment is concerned with the following question: Is the articulatory domain well-suited to predict the amount of acoustic or perceptual mismatch? In other words: How good is the speech quality of a concatenative synthesizer using a purely articulatory-based unit join cost function?

The results of the first experiment were applied in an articulatory-based join cost which estimates the Euclidean distance between all articulatory trajectories which meet at the unit boundary, thus minimizing the position mismatches. Former experiments based on EMMA data [3] already indicated that velocities and accelerations are less important than positions.

The original spectral distance measure of the synthesis system was replaced with this purely articulatory-based join cost function to synthesize 50 utterances. 12 subjects participated in a MOS test and assessed their quality. The result was that the articulatory-based join cost function is equally suited to predict perceptual mismatch as the acoustic join cost function. A further and very encouraging outcome was that the suprasegmental structure of the generated speech was maintained without explicitly using any prosodic features.

Discussion

Concatenative speech synthesis is feasible without taking into account any acoustic measurements. It is surprising that even F0 contours become more continuous without applying any pitch detection methods. A probable explanation for this observation could be that F0 changes are correlated with variations of the larynx height which in turn causes vocal tract deformations. These deformations are superimposed on the phonemically induced vocal tract deformations and thus are also represented in the articulatory kinematics.

The Euclidean distance-based join cost function uses invariant weights. However, this is only a simple approximation of what is necessary to more adequately process articulatory kinematics. The most interesting challenge is to cope with the *Sensitivity Functions* [1] of the vocal tract which are closely related to the *Quantal Nature of Speech* [5]. The simple join cost function should be replaced with a more sophisticated distance function which takes into account the variable positioning precision of certain articulators when approaching particular regions or degrees of constriction in the vocal tract.

References

[1] Boë, L.-J.; Perrier, P. (1990). Comments on "Distinctive regions and modes: A new theory of speech production" by M. Mrayati, R. Carré and B. Guérin. *Speech Communication*, 9(3): 217–230.

[2] Pfitzinger, H. R. (2000). Removing hum from spoken language resources. In *Proc. of ICSLP 2000*, vol. 3, pp. 618–621, Beijing.

[3] Pfitzinger, H. R. (2002). Fleshpoint measurements of articulatory kinematics in concatenative speech synthesis. Forschungsberichte (FIPKM) 39, pp. 17–24, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München.

[4] Pfitzinger, H. R. (2003). Using two- and three-dimensional fleshpoint measurements of articulatory kinematics in concatenative speech synthesis. In *6th Int. Seminar on Speech Production. Programme and Abstracts*, pp. 51, Manly, Australia.

[5] Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In Denes, P. B.; David Jr., E. E., eds., *Human communication: A unified view*, chap. 3, pp. 51–66. McGraw-Hill, New York.

[6] Zierdt, A.; Hoole, P.; Tillmann, H. G. (1999). Development of a system for three-dimensional fleshpoint measurement of speech movements. In *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 1, pp. 73–75, San Francisco.