Abstract for the11th Conference on Motor Speech: Motor Speech Disorders and Speech Motor Control [Abstract], Williamsburg, 2002.
Christian Geng & Christine Mooshammer

Speaker normalization of stressed and unstressed vowels in articulatory and formant spaces

One of the goals of phonetic investigations is to find strategies for vowel production independent of speaker-specific vocal-tract anatomies and individual biomechanic properties. In this paper we apply a technique for speaker normalization to formant spaces, lingual vowel target positions and palate midlines. The general goal is to extract speaker-independent strategies for stress production in vowels and relate speaker-dependent strategies to their specific anatomy. The method used here, termed Procrustes Analysis, is a method for translating, rotating and scaling objects to obtain one consensus configuration -in our case a consensus formant space, a consensus articulatory space or a consensus palate. In addition to these consensus configuration the analyses return speaker-dependent coordinates and transformation matrices necessary to transform the original speaker-dependent data into their modeled cognates. Procrustes Analysis comes in two guises: One, the orthogonal version, is based on Euclidean geometry, i. e. only transformations are allowed which preserve angles between measured landmarks or objects. A second version, Generalized (or oblique) Procrustes Analysis is based on affine geometry, which means that angles between corresponding landmarks are not necessarily preserved. The advantage of this method is that it can be applied to different kinds of data. Most acoustical speaker normalization procedures cannot be applied to articulatory data and vice versa. Furthermore a comparison of individual models of palate shapes and articulatory vowel spaces can be made to achieve an abstraction from individual biomechanical properties.

Data acquisition: Six native speakers of German were recorded by means of Electromagnetic Midsagittal Articulography (AG100, Carstens). The corpus consisted of nonsense words containing /tVt/ syllables with nuclei V=/i,ɪ,y,ʏ,e,ɛ,ɛː,ø,œ,aː,a,u,ʊ,o,ɔ/ in stressed and unstressed positions. Stress alternations were fixed by morphologically conditioned word stress and contrastive stress. Each CVC sequence was embedded in the carrier phrase "Ich habe ’tVte, nicht tV’tal gesagt." (I

said _, not _) and each of the 15 sentences were recorded six (three speakers) and ten (three speakers) times. Four sensors were attached to the tongue, one on the lower lip and one on the lower incisors. The analyses reported are limited to the four transducers on the tongue. Simultaneously, the speech signal was recorded by a DAT recorder. Articulatory target configurations were extracted in the mid of the acoustically defined vowel. The data were averaged over the available repetitions of each vowel. Formant values of the first two formants were extracted and palate midlines were measured by sliding calipers on the EPG palates.

Results: The amount of necessary affine transformation of speaker-dependent data compared to the consensus object is shown by means of so called "strain-crosses", a visualization technique for the behaviour of the eigenvalues of the transformation matrices applied. If these eigenvalues are identical, then the shape change from original to modeled data is isotropic and the angles are not uniquely defined. As a consequence, the orthogonal version of the model is appropriate and objects can be alligned using a combination of rigid rotations and scalings. This was sufficient to achieve a consensus model in the case of the formant data. Speaker-dependent variance was reduced to 15 percent of the raw data (Lobanov: 47%). In contrast, the lingual tongue positions during vowels had to be affinely stretched or shrunk to a much greater degree than the formant values.

The modeled articulatory data provide us with the possibility to visually compare the (modeled) stressed and unstressed target configurations of different speakers in one single consensus coordinate system. These projections suggested that all unstressed vowels were produced by a greater degree of coarticulation.

One drawback of the procedure applied is that of a possible "overnormalization": With respect to the modeled articulatory data not only variance due to anatomical differences, but also substantial articulatory variability is removed. In order to assess the severity of such potential artifacts,canonical correlations between the deformation matrices of the midsagittal palate traces and the deformation matrices of our articulatory vowel target data were calculated. The first canonical correlation accounted for more than 90% the total variance, which suggests that these artifacts are present but do not in principle endanger our interpretations with respect to word stress.

Discussion: The applied methods provide a useful means for modeling speaker-independent strategies of stress production. Not all speaker-dependent differences could be attributed to different palate shapes. This result can either be interpreted as differential behavior independent of biomechanical properties or that the palate shape does not account for all variance in the data, e.g. other structures such as the pharynx shape might also play an important role.