Articulatory Speech Re-synthesis: Profiting from Natural Acoustic Speech Data

Dominik Bauer, Jim Kannampuzha, and Bernd J. Kröger

Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Aachen, Germany
{dobauer,jkannampuzha,bkroeger}@ukaachen.de

Abstract. The quality of static phones (e.g. vowels, fricatives, nasals, laterals) generated by articulatory speech synthesizers has reached a high level in the last years. Our goal is to expand this high quality to dynamic speech, i.e. whole syllables, words, and utterances by re-synthesizing natural acoustic speech data. Re-synthesis means that vocal tract action units or articulatory gestures, describing the succession of speech movements, are adapted spatio-temporally with respect to a natural speech signal produced by a natural "model speaker" of Standard German. This adaptation is performed using the software tool SAGA (Sound and Articulatory Gesture Alignment) that is currently under development in our lab. The resulting action unit scores are stored in a database and serve as input for our articulatory speech synthesizer. This technique is designed to be the basis for a unit selection articulatory speech synthesis in the future.

Keywords: speech, articulatory speech synthesis, articulation, re-synthesis, vocal tract action units.

1 Introduction

Articulatory speech synthesizers are able to generate an acoustic speech signal from an articulatory description. (Kröger 1993, Birkholz 2005, Birkholz et al. 2007, Engwall 2005, Badin et al. 2002). Our articulatory speech synthesizer (Birkholz et al. 2007, Kröger and Birkholz 2007) is controlled by a set of temporally coordinated vocal tract actions (action-based control concept) where the complete acticulatory control information is stored in the "action unit score". Typical examples for vocal tract actions are bilabial closing actions, vocalic tract forming actions, consonantal closing actions, glottal opening and velopharyngeal (or velic) opening actions (Kröger and Birkholz 2007). If it is the goal to integrate an articulatory speech synthesizer into a TTS-system the system must be able to generate vocal tract action scores automatically. In our TTS approach the dual-route phonetic encoding idea (Levelt and Wheeldon 1994 and Levelt 1999) is integrated. It says that action scores for frequent syllables are stored as one object, whereas the action scores for non-frequent syllables are generated by rule. The complete TTS-system would then consist of a database, where all articulatory action scores of frequent syllables are saved and it will also have a rule-based mechanism to create non-frequent syllables not included in the database. The rules in addition have to account for prosodic and paralinguistic variation.

A. Esposito and R. Vích (Eds.): Cross-Modal Analysis, LNAI 5641, pp. 344–355, 2009.

The extraction of articulatory movement parameters from the acoustic speech signal is often solved in a speaker-dependent way by using acoustic-to-articulatory inversion procedures (e.g. Dang and Honda 2002). In our approach many spatio-temporal parameters of vocal tract action units are specified with respect to speaker-independent mean values stemming from articulatory measurement data from speakers of different Indo-European languages (Draper et al. 1959, Moll and Daniloff 1971, Löfqvist and Yoshioka 1980, Yoshioka et al. 1981, Adams et al. 1993, Löfqvist and Gracco 1997, Wrench 1999, Löfquist 2005, Birkholz et al. 2007, Deterding and Nolan 2007). This leads to important constraints on the level of the articulatory control model (on the level of the action unit score). The remaining vocal tract action parameters can be estimated easily from the acoustic signal.

2 Control of an Articulatory Speech Synthesizer

In our control concept each syllable realization is considered to consist of one or more vocal tract action units distributed over gestural tiers. These tiers are named 'vocalic', 'consonantal', 'velic', 'glottal' and 'subglottal pressure' (Kröger and Birkholz 2007 and Birkholz et al. 2006, and see Fig. 1). All actions are realized as goal- or target-directed movements. In the case of *vocalic actions*, which occur on the vocalic tier, the goal is to reach a vowel specific vocal tract shape. In the case of *consonantal closing actions*, which occur on the consonantal tier, the goal is to reach a consonant specific *full-closure* for realizing plosives or nasals or to reach a consonantal *near-closure* in order to realize a fricative. Since some phoneme realizations have the same kind of oral constriction, a single consonantal action unit can correspond with more than one phone. The disambiguation is done by combination with other action units on the velic tier or on the glottal tier. For example an apical full-closing action on the consonantal tier can be combined with a *glottal opening action* (opgl) for producing a voiceless plosive. For example a labial full-closing action (clla) on the consonantal tier can be combined with glottal closing action (clgl) on the glottal tier to result in a fully voiced bilabial plosive or with glottal closing action and velopharyngeal or *velic opening action* to produce a bilabial nasal (Note that clgl is a default gesture and thus not indicated in action scores, see Kröger and Birkholz 2007). The set of consonantal action units consists of full-closing actions (cl) for the production of stops and near-closing actions (nc) for the production of fricatives. Beside the manner of articulation, the consonantal action units also contain information about the place of articulation. Full closings can be labial (clla), apical (clap) and dorsal (cldo). Near closings can be labio-dental (ncld), alveolar (ncal) and postalveolar (ncpo) (Kröger and Birkholz 2007).

All action units comprise an onset, steady-state, and offset time interval (Fig. 1 and Kröger et al. 1995). The steady-state time interval is often nearly zero, since in real articulation behavior, no steady states can be found although perception suggests a succession of (steady state) phones. Steady states are used in our control model mainly for defining the time interval of full-closure or near-closure time intervals for plosives, nasals, and fricatives. During action onset time intervals, the goal-directed movement of the articulator towards the action target (e.g. a labial, apical, or dorsal full- or near-closure, a wide velopharyngeal or glottal opening, a specific vocalic vocal tract shape) is performed, while during offset the articulator-dependent rest

position is approached if no further action unit is using that particular articulator at the same time. Thus the time interval of action onset represents the time interval from the start of activation of an action until the time point at which the target or goal of the action is reached. The articulatory targets of actions were estimated for a speaker of Standard German by adaptation of MRI data (Birkholz and Kröger 2006).

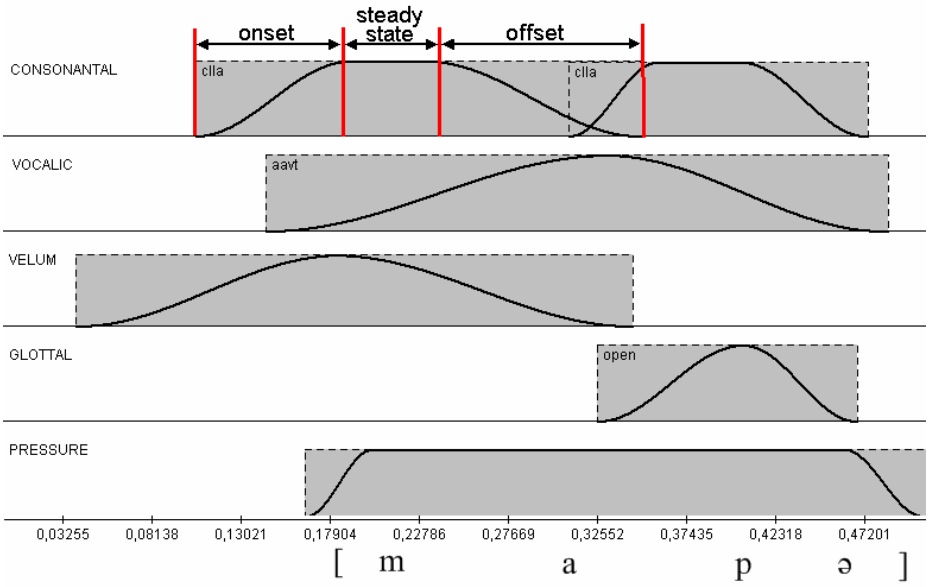

Fig. 1. Action score of the German word "Mappe" (see also Fig. 3 and Fig. 4). Onset, steady state, and offset time intervals are marked for the first consonantal closing gesture. The time points 'begin of onset' and 'end of onset' (BON and EON) and 'begin-' and 'end of offset' (BOF and EOF) are labeled. Amplitude (in bold black lines) indicates the degree of realization of each action. The transcription of the word is given below the action score. Note that the second vocalic action in "Mappe" for realizing the schwa-sound is a default action (Kröger and Birkholz 2007) and thus not explicitly shown in the action score.

3 Re-synthesis Method

Our re-synthesis approach is basically an analysis by synthesis method. The natural acoustic signal of each speech item produced by our reference speaker of Standard German is collected in an acoustic database (Fig. 2). Each item is transcribed manually first. A phonological action score consisting of all relevant vocal tract action units is generated next. At this level no exact temporal alignment is provided. After a first re-synthesis the resulting (synthetic) acoustic wave file is compared with the natural wave file and a temporal alignment of action units can now be done (see below). Temporal alignment is mainly done for matching discrete landmarks in the acoustic signal between natural and synthetic wave file (i.e. onset and release time points of consonantal full-closures or near-closures and begin or end of voicing). It is not possible to match the formants of the natural wave file exactly with that of the synthetic wave file, but the general tendencies of formant movements are tried to be

matched as well. In this way a manual fine tuning with respect to acoustic land marks and with respect to formant trajectories is done. After this tuning, the current action score is stored in the action score database (Fig. 2). The re-synthesis is done by using the articulatory speech synthesizer described by Birkholz (2005) and Birkholz et al. (2007). Note that in our re-synthesis method the pitch trajectories are copied from the original utterance to ensure that the perception of the synthetic signal is not affected by intonation artifacts. Although our action-based control model in principle is capable of generating intonation patterns by rule, the natural intonation contour is copied in this study in order to concentrate on supralaryngeal articulation.

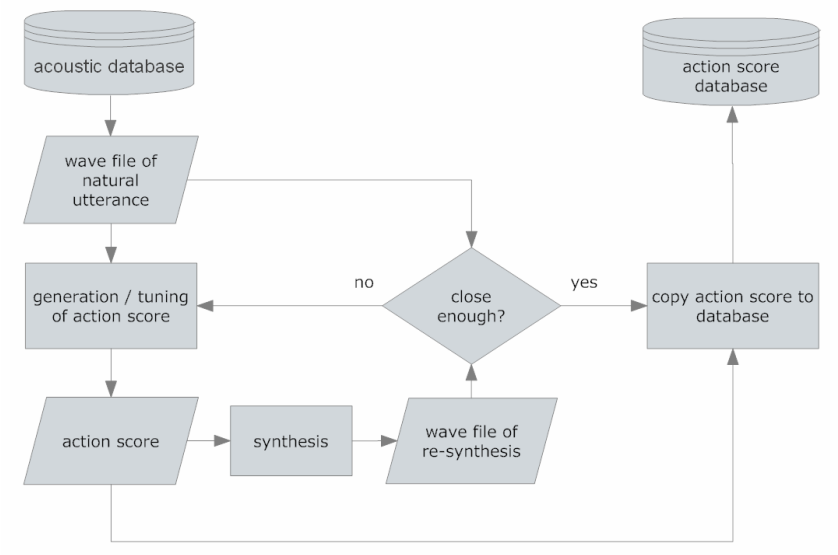

Fig. 2. Schematic view of the re-synthesis workflow

The fitting of natural and synthetic acoustic wave files is performed with the software tool "SAGA". This software comprises the possibility of synchronous display of oscillogram and spectrogram of natural and synthetic signal. Furthermore the software allows the synchronous display of the vocal tract action score for performing action specific temporal alignments (Fig 3). This temporal alignment of onset, steady state, and offset time interval of action units can be done manually in order to match the acoustic landmarks in the acoustic speech signal. In addition, the program is able to show and to overlay synchronously intensity, pitch, and first three formant trajectories for both signals (natural and synthetic).

4 Acoustic Data Corpus

Re-synthesis is done in this study for words, pseudo-words and syllables. The main problem with the collection of an acoustic data corpus is that either a set of (phonotactic correct) pseudo-words with a systematic variation of their syllable structure or a set of

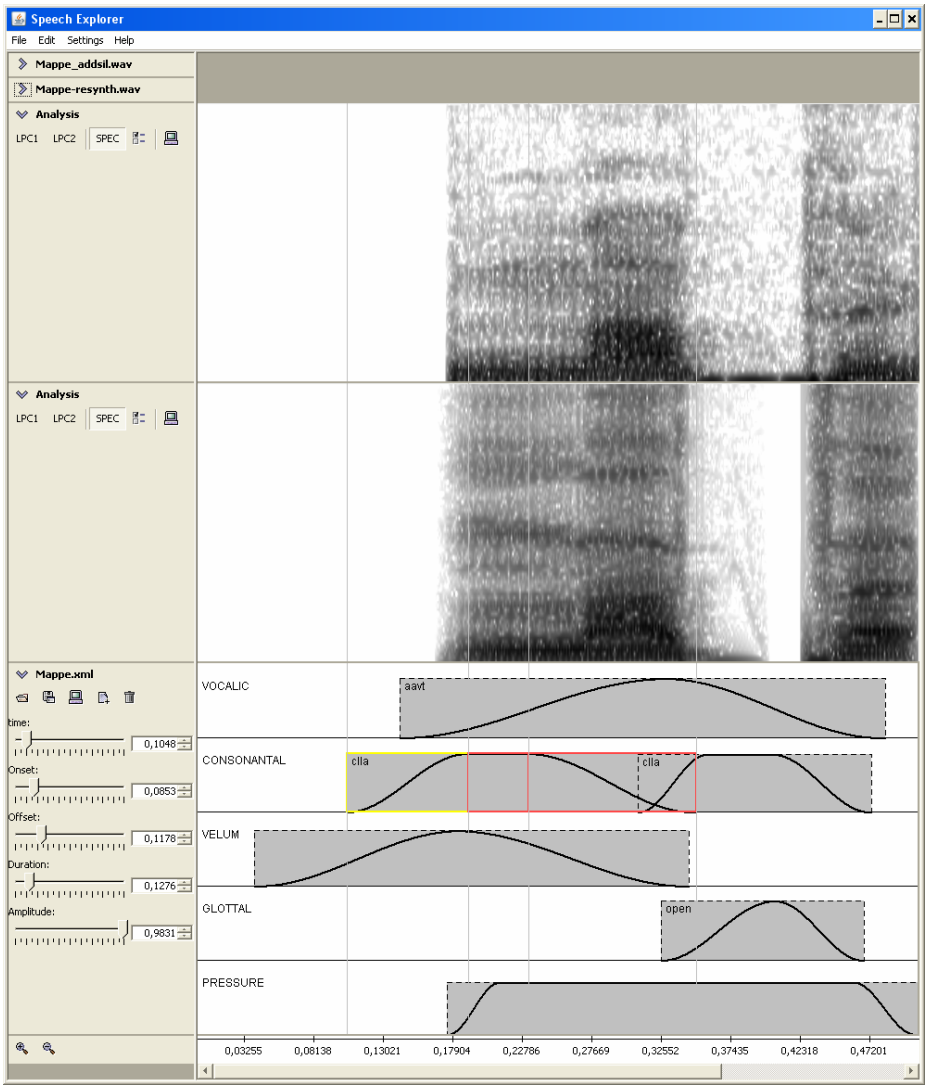

Fig. 3. Screenshot of SAGA showing the spectrograms of synthetic and natural acoustic signals and of the appropriate action score of the German word "Mappe" (see also Fig. 1 and Fig. 4). The lower spectrogram shows the synthetic, the upper the natural signal. The action score in the lower part contains the different tiers for action units: vocalic, consonantal, velum, glottal and subglottal pressure. The temporal alignment labels are shown for onset and offset of the first consonantal closing action (vertical lines). Rules for temporal alignment in the case of this action see section 5.1.

words which really exist in Standard German can be chosen. In the latter case, gaps for certain phone combinations often occur. For this reason we decided to have two corpora with different focus: A pseudo-word corpus and a real-word corpus.

Table 1. Corpus of mono-syllabic words with CVC structure in Standard German. The vowel can be a German tense or lax vowel or a diphthong. Rows indicate different initial consonants. Columns indicate different final consonants.

	[p]	[t]	[k]	[m]	[n]	[ŋ]	[l]
[b]	[bɔp] [buːp]	[bɛt] [boːt]	[bɔk] [boːk]	[baʊm]	[baːn]		[baɪl] [bɑl]
[d]	[diːp]		[dɪk] [dɔk]	[dɑm]	[dɑn]	[dɪŋ]	[dɔl]
[g]	[gaːp] [gɪp]	[gɔt] [guːt]				[gɑŋ] [gɔŋ]	[gaʊl]
[p]	[pɔp]	[pɑt]	[pak]				[paʊl] [poːl]
[t]	[taʊp] [tɪp]	[taːt] [toːt]	[taːk] [taɪk]		[toːn]	[tɑŋ]	[taːl] [taɪl] [tɔl]
[k]	[kɑp]	[kɪt] [koːt]		[kɑm]	[kɪn]		[kaːl] [koːl] [kaɪl] [kyːl]
[m]	[mɔp]	[maːt]	[maːk]	[mʊm]	[mɑn] [moːn]		[maʊl] [meːl] [mɔl] [mʏl]
[n]	[nɛp]	[naːt] [naɪt] [nɛt] [noːt]		[nɪm]	[naɪn] [nɔɪn]		[niːl] [nʊl]
[l]	[laʊp] [laɪp]	[laʊt] [laɪt] [liːt] [loːt]	[lɛk] [lɔk] [laːk]	[laːm] [lɑm] [leːm] [laɪm]	[loːn]	[lɑŋ]	[lal]

The *pseudo-word corpus* contains CV syllables with all voiced and voiceless plosives, nasals and a lateral (/b, d, g/, /p, t, k/, /m, n/, /l/) combined with all 5 long vowels in Standard German (/i/, /e/, /a/, /o/, /u/) (→ 45 items) and CCV syllables with plosives as the first consonant and the lateral as the second consonant (/bl/, /gl/, /pl/, /kl/) (→ 10 items). The CCV syllables match the phonotactic constraints of Standard German. For this reason the corpus does not contain syllables like */dli/ or */tla/. These 55 items were recorded four times by a speaker of Standard German.

The *real-word corpus* comprises 85 mono-syllabic words of Standard German with CVC structure (Tab. 1) and currently 21 bi-syllabic words of Standard German with syllable structure CV-CV with a variation of the first consonant (Kanne, Panne, Tanne) and with a variation of the second consonant (Macke, Mappe, Matte). In this

constellation the second consonant is ambisyllabic when the first vowel is lax (Tab. 2). In addition four words with CVm-pV structure were integrated in order to re-synthesize nasal-plosive successions (Tab. 2). All words were recorded in the environment of the carrier sentence "Ich habe xxx gesagt" (Where 'xxx' is replaced by the actual word). In this sentence position the words are always stressed. Most of the recorded words showed a strong coarticulation effect from the velar constriction of the following word "gesagt". Thus this action unit (cldo) was included during the re-synthesis even if it is not an integral part of the target word.

Table 2. Corpus of bi-syllabic words with CV-CV or CVC-CV structure in Standard German. The first vowel is always a German lax vowel, the second vowel is always schwa. Rows indicate different initial consonants. Columns indicate different medial consonants and consonant clusters.

	[p]	[t]	[k]	[m]	[n]	nas + plos
[p]	[pɑpə]		[pɑkə]		[pɑnə]	[pʊmpə] [pɑmpə]
[t]	[tɑpə]	[tɪtə]			[tɑnə] [tɔnə]	[tantə]
[k]	[kɑpə]		[kɑkə]	[kɔmə]	[kɑnə]	[kantə]
[m]	[mɑpə]	[mɑtə]	[mɑkə] [mykə]			
[n]		[nɛtə]			[nɔnə]	

5 General Alignment Rules

Three different types of temporal alignment rules or temporal coordination rules exist. i) Alignment of onset or offset time points of an action with specific acoustic landmarks like begin or release of consonantal closure or begin of voicing (*acoustic-to-action alignment*, circles in Fig. 4); ii) Alignment of time points of onset or offset of an action with respect to time points of onset or offset of an other gesture (*inter-action alignment*, diamonds in Fig. 4); iii) Alignment of duration of onset or offset of an action (*intra-action alignment*, squares in Fig. 4). These different types of alignment occur for all types of action units, i.e. consonantal, vocalic, glottal, velic, and subglottal pressure action units (see below and see Fig. 4). In addition vocalic targets can be adjusted with respect to formant trajectory matching.

5.1 Alignment of Consonantal Action Units

In our approach, the manual temporal alignment of action scores starts with the alignment of consonantal action units realizing the syllable initial and syllable final consonants. The durations of onset and offset intervals of consonantal full-closing and near-closing actions are taken from articulatory data (Tab. 3). That leads to an intra-action alignment for these action units (see time interval labels 1 to 4 in Fig. 4). From

the acoustic signal consonantal full-closure or near-closure intervals can be easily detected in most cases for plosives, nasals, fricatives, and laterals. This leads to a temporal alignment of EON and BOF for the appropriate consonantal closing vocal tract action units (see time point label 5 to 8 in Fig. 4 for example word "Mappe"). Since the temporal alignment of the acoustic landmarks for begin and end of consonantal closure or consonantal constriction have to be slightly before EON and slightly after BOF (see Fig. 4 and Fig. 1, rule: acoustic begin of closure or constriction coincides with time point at 3/4 of onset interval; acoustic end of closure or constriction coincides with time point of 1/4 of offset interval) these time intervals must be set before the temporal alignment of EON and BOF can be done for consonantal full-closing or near-closing actions.

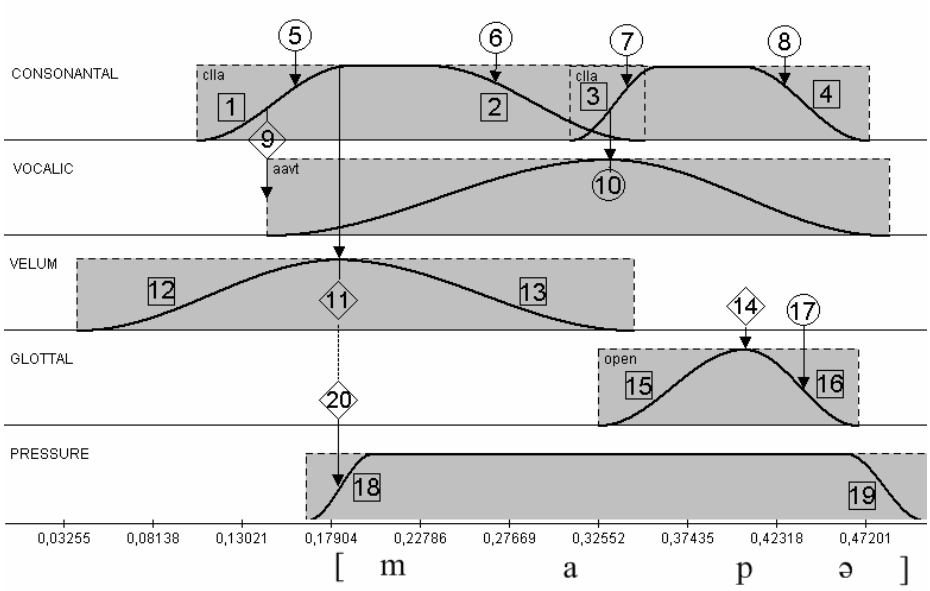

Fig. 4. Action score of the German word "Mappe" (see also Fig. 1 and Fig. 3). The time point labels indicated by circles indicate time points which are aligned with respect to the natural acoustic data (acoustic-action alignment). The time point labels indicated by diamonds indicate time points which are aligned with respect to other actions (inter-action alignment). The time interval labels indicated by squares indicate duration of onset or offset time intervals of actions which are set with respect to articulatory data (intra-action alignment).

Onset and offset durations of consonantal action units are shorter than those of vocalic action units. The onset duration ranges between 75 and 90 ms, the offset duration between 95 and 120 ms for consonants (Tab. 3). Plosives have slightly longer offsets than fricatives and nasals. The temporal coordination of these already aligned consonantal action units with glottal action units and velic action units in the case of the production of voiceless plosives, fricatives and nasals is given below (section 5.3 and 5.4).

Table 3. Onset and offset durations of consonantal closing actions and reference

action unit	onset	offset	reference data
clla (plosive)	90	110-120	Löfqvist 2005, Fig. 10
			Löfqvist and Gracco 1997, Fig. 3
clla (nasal)	75-85	95-120	Löfqvist 2005, Fig. 11
clap	90-120	100-120	Wrench 1999, Fig 2
	100	100	Adams et al. 1993, Fig. 6
cldo	90-120	100-140	Wrench 1999, Fig 2

5.2 Alignment of Vocalic Action Units

The preparation of the vocal tract to produce a vowel starts long before the vowel becomes audible. With an acoustic based alignment method it is not possible to determine the starting time exactly because it is covered by the preceding consonants. The EMA data acquired by Wrench (1999) show that the preparation is done during the constriction phase of the preceding consonants and it can still go on after the release of the closure. Thus the temporal coordination of vocalic actions can be done with respect to the already aligned preceding and following consonantal closing actions. Begin of vocalic onset starts in the middle of the onset interval of the preceding consonantal gesture (time point label 9 in Fig. 4). End of vocalic onset coincides with the middle of the onset interval of the following consonantal gesture (time point label 10, Fig. 4). The offset interval of vocalic gesture is synchronous to the onset interval of the following vocalic gesture, i.e. the offset starts in the middle of the onset of the following consonant and ends in the middle of the next consonant onset interval (not illustrated in Fig. 4).

5.3 Alignment of Velic Action Units

EMA analyses of the velum movement in nasals indicate that the velopharyngeal opening is at its maximum at the end of onset interval of the appropriate consonantal closing gesture (Wrench 1999). Thus the end of onset interval of a velic action coincides with the end of onset of the appropriate consonantal closing action (time point label 11 in Fig. 4). The onset movement is relatively slow and starts already at the beginning of the preceding vowel. The offset movement of the velic action unit often begins during the appropriate consonantal closing action unit. The same finding was reported by Moll and Daniloff (1971). The length of onset and offset interval is about 200 ms (Horiguchi and Bell-Berti 1987) and can be used for specifying the begin of onset time interval and the end of offset time interval for the velic action (time interval label 12 and 13 in Fig. 4). When the nasal is followed by a plosive, the velum raises much faster in order to prevent a pressure loss during the constriction phase of the following plosive. That leads to a much shorter offset interval for the velic gesture. This is in accordance also with the EMA data given by Wrench (1999). The duration of onset and offset time interval of a velic action unit during the production of nasals ranges between 140 ms and 250 ms (see above), but can be shortened up to 100 ms when a plosive follows (not shown in Fig. 4).

5.4 Alignment of Glottal Action Units

Glottal opening action units (opgl) occur in voiceless fricatives and voiceless plosives. In voiceless plosives the glottal opening reaches its maximum in the middle of the offset of the consonantal full-closing gesture in order to ensure a strong noise burst (aspiration) at the release of the consonantal constriction (Löfqvist and Yoshioka 1980, and see time point label 14 in Fig. 4 and see Fig. 5 left side). For fricatives the glottal opening reaches its maximum in the middle of the steady state portion of the appropriate consonantal near-closing action (ibid., and see Fig. 5 right side). The duration of onset and offset time interval vary between 90 ms and 120 ms for onset and between 100 ms to 130 ms for offset (ibid.). Thus the beginning of onset and end of offset interval of glottal opening actions is determined by these durations (time interval label 15 and 16 in Fig. 4). In addition the exact duration of onset and offset in the case of plosives is limited by the fact that begin of onset coincides with the begin of the steady state portion of the appropriate consonantal full-closing gesture in order to prevent pre-aspiration and the middle of offset time interval coincides with the begin of phonation for the following vowel (time point label 17 in Fig. 4).

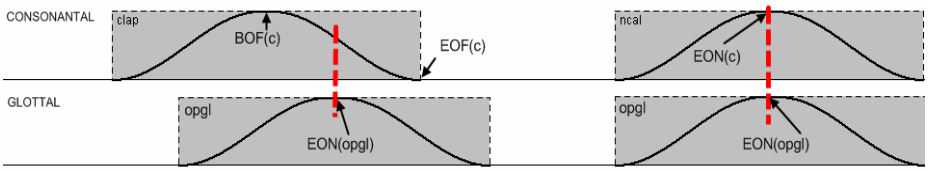

Fig. 5. Temporal coordination of glottal opening actions with respect to the appropriate consonantal closing action for plosives (left side) and for fricatives (right side)

5.5 Alignment of Subglottal Action Units

Each utterance is produced on the basis of one single subglottal pressure action. A fast pressure built-up and pressure fall is assumed (i.e. short pressure onset and offset intervals (around 30 ms; time interval labels 18 and 19 in Fig. 4). The goal of pulmonic actions is that the subglottal pressure is roughly constant over the complete utterance (Draper et al. 1959).

It can be assumed that the constriction of the first consonant of an utterance coincides with the middle of onset interval for subglottal pressure built up (time point label 20 in Fig. 4). That rule ensures a correct production of the first consonant without an unwanted insertion of prevocalic schwa which would occur, if subglottal pressure onset starts earlier. If subglottal pressure onset starts later than defined by this rule, the intra-oral pressure built-up during the consonantal closure of obstruents is not strong enough or the voicing of sonorants would start too late.

The temporal coordination of the offset of the subglottal pressure action coincides with the offset interval of the consonantal gesture if the utterance ends with a consonant. The offset interval of the subglottal pressure action coincides with the offset of the vocalic gestures if the utterance ends with a vowel. In the case of our corpus, the offset of the subglottal pressure action is always temporally coordinated

with the offset of the consonantal closing gesture of the /g/-realization, which is part of the first syllable of the last part of the carrier sentence "gesagt" (see section 4). This closing action is not shown in Fig. 1, 3, and 4 in order not to achieve simple and understandable figures.

6 Results and Discussion

Re-synthesis trials were done for 22 out of 85 items of the mono-syllabic real-word corpus and for 11 out of 21 items of the bi-syllabic real-word corpus (see section 4). The re-synthesis procedure indicates that the control model including the acoustic-to-action alignment rules and the inter-action and intra-action alignment rules on the one hand leaves still enough flexibility for fitting the natural speech signals of our speaker of Standard German but on the other hand delivers enough constraints for specifying all action parameters for the complete action unit score (i.e. location of onset, steady state, and offset for all vocal tract action units of the score).

A preliminary perceptual evaluation of the synthesized utterances indicates a high degree of naturalness and intelligibility. Especially the smooth transitions of the articulation movements resulting from the temporally overlapping vocal tract action units as result from our control model lead to promising results. The action-based control concept is capable to handle all major phonetic effects for producing a high quality acoustic signal.

The method described in this paper is just a preliminary approach to generate TTS-applicable control information for articulatory speech synthesis. But the procedure can also be used for basic phonetic research on articulatory processes and is also the basis for constructing an action score corpus for frequent syllables in Standard German. It is planned to complete the set of re-synthesized action scores for the whole corpus outlined in section 4 and to define rules for an automatic generation of action scores for infrequent syllables.

Acknowledgments. This work was supported in part by the German Research Council DFG grant Kr 1439/13-1 and grant Kr 1439/15-1.

References

Adams, S.G., Weismer, G., Kent, R.D.: Speaking Rate and Speech Movement Velocity Profiles. Journal of Speech and Hearing Research 36, 41–54 (1993)

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C.: Three-Dimensional Linear Articulatory Modeling of Tongue, Lips and Face, Based on MRI and Video Images. Journal of Phonetics 30, 533–553 (2002)

Birkholz, P.: 3D Artikulatorische Sprachsynthese. Ph.D Thesis, Rostock (2005)

Birkholz, P., Kröger, B.J.: Vocal Tract Model Adaptation Using Magnetic Resonance Imaging. In: Proceedings of the 7th International Seminar on Speech Production, Belo Horizonte, Brazil, pp. 493–500 (2006)

Birkholz, P., Jackel, D., Kröger, B.J.: Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15, 1218–1225 (2007)

Birkholz, P., Jackèl, D., Kröger, B.J.: Construction and Control of a Three-Dimensional Vocal Tract Model. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), Toulouse, France, pp. 873–876 (2006)

Birkholz, P., Steiner, I., Breuer, S.: Control Concepts for Articulatory Speech Synthesis. In: Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 5–10 (2007)

Dang, J., Honda, K.: Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. Journal of Phonetics 30, 511–532 (2002)

Deterding, D., Nolan, F.: Aspiration and Voicing of Chinese and English Plosives. In: Proceedings of the ICPhS XVI, Saarbrücken, pp. 385–388 (2007)

Draper, M.H., Ladefoged, P., Whiteridge, D.: Respiratory Muscles in Speech. Journal of Speech and Hearing Research 2, 16–27 (1959)

Engwall, O.: Articulatory Synthesis Using Corpus-Based Estimation of Line Spectrum Pairs. In: Proceedings of Interspeech, Lisbon, Portugal (2005)

Horiguchi, S., Bell-Berti, F.: The Velotrace: A Device for Monitoring Velar Position. Cleft Palate Journal 24(2), 104–111 (1987)

Kröger, B.J.: A gestural production model and its application to reduction in German. Phonetica 50, 213–233 (1993)

Kröger, B.J., Birkholz, P.: A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)

Kröger, B.J., Schröder, G., Opgen-Rhein, C.: A gesture-based dynamic mo¬del describing articulatory movement data. Journal of the Acoustical Society of America 98, 1878–1889 (1995)

Levelt, W.J.M., Roelofs, A., Meyer, A.S.: A Theory of Lexical Access in Speech Production. Behav. Brain Sci. 22, 1–38 (1999)

Levelt, W.J.M., Wheeldon, L.: Do Speakers Have Access to a Mental Syllabary? Cognition 50, 239–269 (1994)

Löfqvist, A.: Lip Kinematics in Long and Short Stop and Fricative Consonants. J. Acoust. Soc. A. 117(2), 858–878 (2005)

Löfqvist, A., Gracco, V.L.: Lip and Jaw Kinematics in Bilabial Stop Consonant Production. Journal of Speech, Language, and Hearing Research 40, 877–893 (1997)

Löfqvist, A., Yoshioka, H.: Laryngeal Activity in Swedish Obstruent Clusters. J. Acoust. Soc. Am. 68(3), 792–801 (1980)

Moll, K.L., Daniloff, R.G.: Investigation of the Timinig of Velar Movements during Speech. JASA 50(2), 678–684 (1971)

Wrench, A.: An Investigation of Sagittal Velar Movements and its Correlation with Lip, Tongue and Jaw Movement. In: Proceedings of the ICPhS, San Francisco, pp. 435–438 (1999)

Yoshioka, H., Löfqvist, A., Hirose, H.: Laryngeal adjustments in the production of consonant clusters and geminates in American English. J. Acoust. Soc. Am. 70(6), 1615–1623 (1981)