ACOUSTICS VS. ARTICULATION IN ARTICULATORY SPEECH SYNTHESIS: ONE VOCAL TRACT TARGET CONFIGURATION HAS MORE THAN ONE SOUND

Eva Lasarcyk

Department of Phonetics and Phonology, Saarland University
evaly@coli.uni-saarland.de

Abstract

Articulatory speech synthesis can be used for speech production research to gain insight into articulation patterns and their acoustic counterparts, the speech sounds. It can be used e.g. to conduct perception experiments that study the relationship between articulation and fine phonetic detail in the acoustic domain. In a case study, we focus on articulatory details in German vowels. Results indicate that the transcription of vowel quality changes depending on the acoustic settings used. The goal of this contribution is to illustrate the importance of these acoustic settings in articulatory synthesis, and to increase awareness regarding these settings. From a user's perspective, the selection of a specific synthesis strategy entails certain acoustic settings. They define the details of how a geometric vocal tract target configuration is rendered into a vocal tract area function. Again applying certain acoustic settings, this area function is then used in an aerodynamic-acoustic simulation to produce the speech signal. Depending on the acoustic settings, one underlying vocal tract target configuration can in the end produce several different sounds.[1]

1 Introduction

We are all experts in producing speech and use our speech organs to execute numerous coordinated and fast movements in a very small space in our vocal tracts. While it is known that in some regions of the vocal tract, small deviations in the position of an articulator can cause distinct acoustic differences [10], it is not yet fully understood how exactly the speech organs articulate to create sounds. In speech production research, studies aim at furthering the understanding of the details of articulation in the vocal apparatus.

In the study presented here, we use articulatory speech synthesis to investigate articulatory details (and fine phonetic details) of vowel production. We evaluate the articulatory hypotheses in the acoustic domain. The step of going from a geometrically defined vocal tract shape – the vocal tract target configuration – to the acoustic speech signal will be referred to as acoustic synthesis in this paper. We argue that the acoustic settings selected for the synthesis step are of major importance since they can influence the vowel sounds to a considerable degree, even when the underlying vocal tract target configuration remains the same.

Various methods can be applied to investigate the articulatory movements in humans during speech production (cf. [8] for an overview). One type of methods deal with the acquisition of articulatory data which are then processed to access the movements of the speech articulators such as the tongue, the lips, the jaw, or the velum. Examples of these methods are MRI, EMA, X-ray and X-ray microbeam.

[1]Special thanks go to Peter Birkholz for providing the software and for comments and support for this paper.

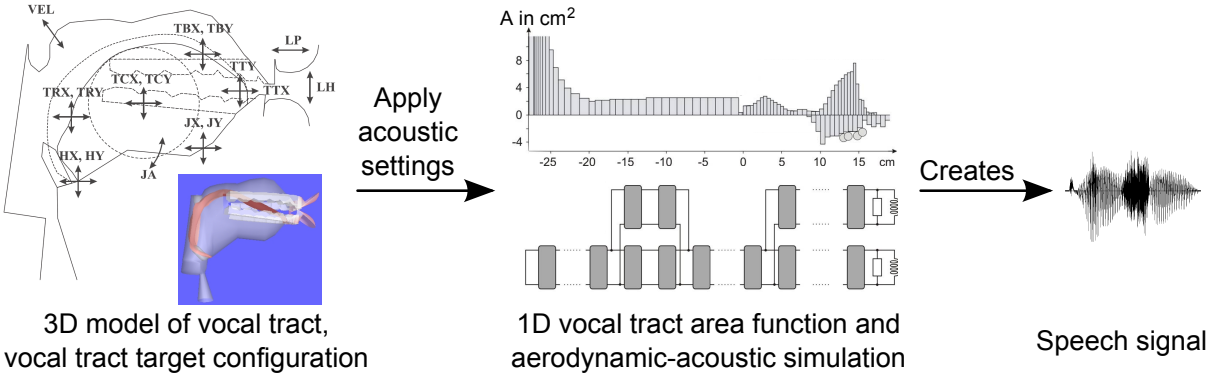

Figure 1 - Basic concept of articulatory speech synthesis. Illustration items adapted from [2, 3].

While it is essential to record data from real speakers, a complementary type of method for speech production research can be articulatory speech synthesis. It can be based on articulatory and anatomical data of real speakers, acquired with the methods above but then go one step further in imitating the movements in virtual models of the vocal tract. If no articulatory data is available to evaluate the hypothesized articulatory movements, an evaluation can also take place in the acoustic domain. This gives indications about possible solutions to articulatory tasks such as the production of a vowel. The virtual vocal tract has the advantage of making the speech production process transparent and every parameter can be controlled separately.

Articulatory synthesis is capable of imitating or modeling fine phonetic detail and has been applied to various aspects of articulation, such as e.g. nasal coupling and vowel height [5], displacement of the tongue body center and formant values [9], acoustic and perceptual characteristics of smiled vowels [7], or impressions of body size vs. emotions [13].

Articulatory synthesis can be used as a discovery tool for speech production research by following a copy-synthesis paradigm and comparing the synthetic stimuli with acoustic (or articulatory) data from a corpus. We select a speech phenomenon from a corpus and create hypotheses about its articulation. These hypotheses can then be transferred into vocal tract target configurations and gestural scores. Vocal tract target configurations describe the shape of a virtual vocal tract by defining the position of the articulators for a given sound with a list of vocal tract parameters. They are the general basis for vocal tract area functions. The gestural scores define the temporal relations and transitions from one shape to the next (articulatory movements). These articulatory descriptions are used to execute an aerodynamic-acoustic simulation, in the end of which we obtain synthetic versions of the original phenomenon in the corpus. To complete the "discovery" process, the acoustic stimuli can be used for auditive and acoustic evaluation to determine e.g. the influence of certain articulatory parameters on the acoustic outcome.

This technique is useful because it makes the whole speech production process transparent, and relations of articulation to acoustics can be studied in an idealized vocal tract. In addition, it is relatively low-cost, and experiments are, in principle, reproducable because the program can simply be executed again. There are, however, limitations to this technique. First of all, an articulatory synthesizer can only *model* a vocal tract and is not a real speaker. Issues such as implausible articulatory shapes and movements have to be monitored by the experimenter. However, this is not the focus of this contribution. We concentrate on single vowel target configurations because even with plausible articulation assumptions it is possible that the synthetic speech is not representing exactly what we assumed it would do. This is due to the acoustic settings used to synthesize a sound based on a geometrically defined vocal tract shape. Since these settings can influence the acoustic outcome to a considerable degree, they are the focus of this contribution.

In the following sections, we exemplify the issue of acoustic settings with one particular articulatory synthesis software, and report on a case study which illustrates the effects the acoustic settings can have if they are not monitored very closely. We conclude with an appeal for more awareness of acoustic settings in articulatory synthesis when using it for research in speech science.

2 Acoustic Settings

During the process of aerodynamic-acoustic simulation, a geometric vocal tract target configuration is used as the general basis to generate a sound (cf. Fig. 1). Since it is computationally very costly to base the acoustic simulation on complete 3D models, synthesis frameworks usually use 1D models instead ([2], ch. 3). The main simplification which is made in that case is that the movement of fluids takes places in only one dimension. The result is a representation of the vocal tract in terms of a 1D area function (see Fig. 1).

From this area function, the actual aerodynamic-acoustic simulation is calculated. It is hard to obtain the optimal speech outcome due to constraints e.g. in the numeric simulations that have to be applied to simulate all the necessary acoustic effects, and there does not seem to be a consensus about how to deal with all the relevant factors (cf. [11, 12, 1]). Factors can be e.g. the degree of nasal coupling, the kind of voice excitation, the losses that are computed, how the radiation impedance is considered, and how subglottal and glottal coupling are integrated.

In general, one can choose between two basic simulation strategies, frequency-domain simulation (FDS) and time-domain simulation (TDS). Each simulation faces different constraints when simulating the acoustic effects from a given area function and presents the user with different options on how to render the acoustic signal. The possible acoustic settings, described in Table 1 for the software VocalTractLab [2], reflect the choices that can be made for each synthesis strategy. On the one hand, the acoustic settings comprise the core-technical settings for each synthesis strategy (FDS or TDS). On the other hand, they also comprise additional synthesis settings such as e.g. the velic aperture. From a user's perspective, they appear as an acoustic setting because they are controlled outside of the plain vocal tract target configuration. They may thus silently alter the vocal tract area function used for the acoustic-aerodynamic simulation if the user does not pay explicit attention to these settings.

The important point to note here is that articulatory synthesis can create more than one sound from one underlying vocal tract target configuration. Firstly, some acoustic settings influence the way the vocal tract target configuration is transferred into a corresponding vocal tract area function, yielding different area functions from one single target configuration. Secondly, one area function can produce different speech signals depending on the acoustic settings selected for the aerodynamic-acoustic simulation. It is thus not completely reliable to state what a particular vowel target configuration *per se* sounds like because the sound of the vowel very much depends on the strategy of the acoustic synthesis.

The two rightmost columns of Table 1 list default values for each synthesis strategy in VocalTractLab [2]. They will be called synthesis profiles in the following. The FDS profile is used e.g. to generate a single vowel sound directly from a vowel vocal tract target configuration. In default mode, it entails e.g. the generation of a high-low (98–120–77 Hz) intonation contour to utter the vowel. The TDS profile is used when generating speech from a gestural score. For each time step of the simulation, the shape of the vocal tract is computed and used for time-domain synthesis. Our default gestural score was programmed with certain default values including velic aperture (0.23), glottal area (0.24) and intonation (level contour at 109 Hz).

The following case study reports on the *overall* auditive impacts these synthesis profiles can have on listeners who classify different vowels.

Table 1 - Acoustic parameters and their possible values for acoustic synthesis in the synthesis framework VocalTractLab [2], as they can be found in the settings menus for FDS and TDS settings. Not all parameters are applicable to both synthesis strategies.

Parameter	**Description**	**Possible values**	**Default for FDS (FDS profile)**	**Default for TDS (TDS profile)**
Glottal excitation	Model used for voiced excitation	FDS: Liljencrants-Fant (LF) model; TDS: Titze model	LF model	Titze model
Glottal area	Degree of glottal abduction	0–1	0.00	0.24
Radiation impedance	Kind of approximation that is selected for radiation impedance	FDS: 0 (ideal soft opening)/Piston in sphere/ Piston in wall/Parallel circuit of R and L	Parallel circuit of R and L	
Energy losses	Which kinds of energy losses in the vocal tract tube are considered	Yes/no; FDS: Boundary layer resistances, Heat conduction losses, Absorbing/soft walls Hagen-Poiseuille resistance; TDS: Fluid dynamic losses, Soft walls, Sound radiation from skin	Boundary layer resistances, Absorbing/soft walls	Fluid dynamic losses, Soft walls
Signal resistances	Whether extra small signal resistances are used	FDS: yes/no	Yes	
Nose sinuses	Whether nose sinuses are considered or not	FDS: yes/no	Yes	
Piriform fossa	Coupling of the sinus piriformis to the vocal tract	Yes/no	No	No
Lumped elements	Whether lumped elements in T-sections are considered	FDS: yes/no	Yes	
Velic aperture	Degree of velic opening (nasal coupling)	0–1	Defined in vocal tract target configuration for each vowel (0.00)	Defined in gestural score (0.23 in our case)
Intonation	Contour of the fundamental frequency for an utterance	Variable	High-low (98–120–77 Hz)	Defined in gestural score (constant 109 Hz in our case)

3 Case Study

In this section, we illustrate the effects and therefore importance of carefully selected acoustic settings when investigating fine phonetic detail with articulatory synthesis. We analyze fine phonetic details of vowel articulation with respect to vowel height, horizontal tongue body position, and lip rounding. Since these articulatory dimensions correlate with the position of the formants of a vowel, the importance of the acoustic settings is evident: If some settings exert a strong influence on the formants, an auditive evaluation may go as far as to report different underlying phonemes for the same vocal tract target configurations.

This may not be so evident when only prototypical vowels are analyzed. However, in this study, we also analyzed regionally accented vowels which are not prototypical in their sound characteristics. Knowing about the influence of acoustic settings, it is crucial for speech accent evaluation to pay attention to these synthesis settings because they may acoustically manipulate exactly those articulatory dimensions that are being evaluated.

3.1 Data and method

From a study which is work in progress, we use vocal tract target configurations of the German vowels /o, ø, y, i, e, e_{acc}/, where the last one is a regionally accented vowel. Some of the vocal tract target configurations are very similar but not identical. So the question is whether the choice of synthesis profile causes the same or different auditive results. Since the articulatory configurations differ only slightly, a precise acoustic rendering is desirable.

We synthesized each vowel (i.e. each vocal tract target configuration) once with the FDS synthesis profile and once with the TDS profile as shown in Table 1. One version of the vowel therefore belongs to vowel set I (using the FDS profile), the other one belongs to vowel set II (TDS profile). The duration of a vowel is around 650 ms, they are standardized for intensity to around 80 dB. Their formant frequencies are shown in Table 2, measured with Praat [4] (maximum formant = 5000 Hz; number of formants = 5; window length = 0.02 s).

Complementary to the acoustic analysis, we conducted a classification task in order to obtain an auditive characterization of the influence of the acoustic settings. This can give an indication about whether acoustic changes are noticed by listeners or not. Six phonetically trained test subjects classified the vowels. For playback, we used the same pair of headphones for each participant to ensure consistency in the playback quality. Firstly, each participant listened to vowel set II, and then, after a break, to vowel set I. They could listen to a stimulus as often as they wanted. We presented a visual transcription guideline, including the IPA vowel chart and a set of diacritics to describe deviations from the base phoneme regarding vowel height, lip rounding, and horizontal tongue body position.

3.2 Results

Table 2 shows the transcriptions of the vowels of this case study. In some cases, the transcribers put down several transcriptions for a sound, which are indicated by a forward slash /. The sequence of transcriptions reflects the contributions from each transcriber: In each cell, the first result stems from the first transcriber, the second one from the second transcriber etc. This way, individual differences in transcription between the two vowel sets can be traced back.

Although transcribing at this level is not free of a certain subjectivity, we find an overall interlabeler agreement in the transcriptions with respect to base phonemes and vowel height. In the following, we describe the results of comparing the transcriptions of each vowel stimulus in Set I with its counterpart in Set II. The changes across vowel sets include vowel height, horizontal tongue position, lip rounding, and nasality.

Table 2 - Left: Formant frequency values (means taken over the middle 30 % of each vowel, in Hz) of the vowels presented in the case study. Right: IPA transcriptions of the two vowel sets, obtained from 6 trained phoneticians. Diacritics used (from top to bottom): centralized [e̽], nasalized [ẽ], more rounded [e̹] or less rounded [e̜], retracted [e̠] or advanced [e̟], raised [e̝] or lowered [e̞].

Vowel	Set I (FDS)			Set II (TDS)			Set I (FDS)	Set II (TDS)
	F_1	F_2	F_3	F_1	F_2	F_3		
/o/	361	732	2517	336	679	2617	o o̞ o̜ o o̹ o	o̹ o̠ o̹/o̜ o̜ u̟
/ø/	347	1488	2229	335	1490	2410	ø ø ø̠ ø ø ø	œ œ̞ œ̠ œ̝/œ̜ ø̠ ø̞
/y/	274	1732	2214	315	1802	2424	y y ʏ ʏ/ʉ̟ y̆ y/y̆	y̹ ø̽/y̆ ø̠ ø ʏ y̠
/i/	288	2232	3069	315	2366	3098	i i̟ ɪ i̠ i̽ i/e̝	i e̠ e i/e̟ e e̝
/e/	336	2236	2799	363	2315	2798	e e e̠ e e/e̝ e	e̽/ẽ e̠ e ɛ̃ e̠ e
/e$_{acc}$/	358	2068	2739	413	2076	2779	e/e̽ e e̠ e e̹ e/e	ɛ̃ ɛ̠ e̹ ɛ̃ ɛ̠ e̟

An obvious shift takes places in perception of vowel height. It generally decreases from Set I to II by one level in the IPA vowel chart for all vowel catecories. In Set I, the vowels mostly receive their intended symbols [o, ø, y, i, e]. However, in Set II, vowels are often perceived as more open than was intended with the vocal tract target configuration: [ɔ, œ, ø, e, ɛ].

The place of articulation (horizontal tongue position) often shifted to a more retracted auditive impression of articulation, mostly tied to a lower vowel height, e.g. [o̞, ø, ʏ, i̟, e] → [ɔ̠, ø̠, ø̠, e̠, e̠]. However, there are also examples of frontings, such as [e̠] → [e̹], thus the perceptual influence of synthesis profile on horizontal tongue body position is not homogeneous.

The articulatory dimension of lip rounding was affected only slightly by the mode of synthesis, and in an inhomogenous way. A few cases indicate a decrease in lip rounding from Set I to Set II, [ø, y, e̹] → [œ̠, y̠, ɛ̠], while others, [o̜, y̽, e̟, e̠] → [ɔ̟, y̠, ɛ̃, e̹], show indications into the opposite direction, towards more rounded vowels.

Lastly, the overall transcribed amount of nasality increases from Set I to Set II. This can be expected since the velar opening was one of the acoustic parameters, which was changed from closed in Set I to slightly open in Set II.

In an extended vowel set, we also analyzed non-prototypical (i.e. regionally accented) vowels. An example, an accented version of [e], is displayed in the last row of Table 2. The results are similar to the ones depicted above, but in some aspects more distinct, e.g. concerning phonemic changes due to vowel height perception. In the example, accented versions of [e] predominantly receive a different phoneme transcription from Set I to Set II. A prototypical, standard articulation of [e] however is not affected so strongly by changes in synthesis strategy (see Table 2) and the transcribers rather use diacritics than switch to a more open phoneme.

3.3 Discussion and conclusions

The transcription results indicate that the choice of synthesis profile can influence the sound of a vowel to a considerable degree as summarized in Table 3, especially with regard to perceived vowel height. The impact can vary for different vowels, and it also depends on whether a vocal tract target configuration defines a prototypical vowel or a regionally accented vowel (non-prototypical sound). The general tendency in the results is that the acoustic settings used for the TDS vowel set favors a perception towards more open vowels. Horizontal tongue body position and lip rounding are also influenced, but to a lesser degree and not homogeneously. Additionally, these vowels induced a nasalized impression, mainly due to the increased velic

Table 3 - Summary of main acoustic effects of the two synthesis profiles used in this case study.

Synthesis strategy	**I (FDS default)**	**II (TDS default)**
Overall sound	Crisp and clear	Muffled and slightly unclear
Vowel height	As intended	Lowered by about 1 level in IPA vowel chart
Horizontal tongue body position		Often slightly retracted
Lip rounding		Only slightly affected
Nasality	None	Tendency to sound nasalized

aperture in synthesis profile II (TDS). For non-prototypical vowels, the transcriptions reveal even more distinct auditive differences and show phonemic changes where prototypical vowels would still be perceived as phonemically identical.

The main finding of changes in vowel height due to synthesis profile may be explained by the parameter of velic aperture. Since the TDS profile uses substantially more nasal coupling than the FDS profile, some formant changes may have been interpreted as vowel height changes instead of nasalization, or as a combination of both. Support for this explanation can be found in [5] who report that non-contextually nasalized vowels are perceived as more open in vowel quality than oral vowels. Although velic aperture is also defined in the vocal tract target configuration itself, it is treated as an acoustic parameter here because its value is overridden by a parameter on the gestural score. From a user's perspective, this happens silently when switching from TDS, which uses gestural scores, to FDS, which uses solely the vocal tract target configuration. Thus, the resulting vocal tract area functions and the simulated speech signals vary while the vocal tract target configuration remains unchanged.

The more distinct results of accented vs. regular /e/ stimuli could be explained by the acoustic differences between the two pairs of vowels /e/ and /e$_{acc}$/. For the regionally accented vowel, the difference in F_1 is much greater than for the prototypical vowel when comparing FDS vs. TDS versions (see Table 2). Categorical perception could be another explanation because the vocal tract target configuration of /e$_{acc}$/ is non-prototypical for the phonetically trained subjects who all have a native or near-native German language background and are not familiar with that particular regional accent. As the Perceptual Magnet Effect [6] suggests, the participants are more sensitive to acoustic changes in the area of /e$_{acc}$/ than of /e/, and may document the increased discriminative ability by using different phonemic symbols.

4 General Conclusion

We introduced VocalTractLab [2] as a synthesis tool which can be used in speech production research. Coming from a user's perspective, we focused on the acoustic settings that are applied to a given vocal tract target configuration during acoustic synthesis. In a case study, we illustrated the effects that acoustic settings can have on vowel quality if they are not monitored closely enough. Since one and the same underlying vocal tract target configuration even received different vowel phoneme transcriptions (and not only within-phoneme changes), it seems essential to be aware of the acoustic settings used. If, due to unawareness, these are not consistent, this may result in low intra-study consistency and low inter-study comparability.

The possible effects and importance of the acoustic settings were illustrated using one specific synthesis software. However, the general issue probably applies to articulatory synthesis in general due to the common underlying principles of creating an acoustic signal from a geometric shape by using certain assumptions and coping with certain constraints e.g. in numeric simulations.

To make a point, if a phonetic study aims at imitating fine phonetic details and characterizes synthesized articulatory patterns by auditive assessment and transcription, it is essential to as-

sure that the experimental variables are not influenced unexpectedly by settings of the acoustic rendering of the vocal tract target configuration. Otherwise, this may foil a study and hide links between articulation and acoustics that otherwise would have been transparent.

Finally, increased awareness regarding acoustic settings in articulatory synthesis when using it for research in speech science is only the first step. It could be facilitated by extra-transparent software design which allows for direct and more systematic control over the different acoustic settings.

References

[1] BADIN, P. and G. FANT: *Notes on Vocal Tract Computation*. STL-QPSR, 25(2-3):53–108, 1984.

[2] BIRKHOLZ, P.: *3D-Artikulatorische Sprachsynthese*. Berlin: Logos, 2006.

[3] BIRKHOLZ, P.: *Control of an Articulatory Speech Synthesizer Based on Dynamic Approximation of Spatial Articulatory Targets*. In *Proceedings of 8th Interspeech 2007*, pp. 2865–2868, Antwerp, August 2007.

[4] BOERSMA, P. and D. WEENINK: *Praat: Doing Phonetics by Computer [Computer Program]*. Retrieved July 19, 2009, from http://www.praat.org/, 2009.

[5] KRAKOW, R. A., P. S. BEDDOR, L. M. GOLDSTEIN and C. A. FOWLER: *Coarticulatory Influences on the Perceived Height of Nasal Vowels*. J. Acoust. Soc. Am., 83(3):1146–1158, 1988.

[6] KUHL, P. K.: *Learning and Representation in Speech and Language*. Current Opinion in Neurobiology, 4(6):812 – 822, 1994.

[7] LASARCYK, E. and J. TROUVAIN: *Spread Lips + Raised Larynx + Higher F0 = Smiled Speech? - An Articulatory Synthesis Approach*. In *Proceedings 8th International Speech Production Seminar (ISSP)*, pp. 345–348, Strasbourg, December 2008.

[8] MAEDA, S., M.-O. BERGER, O. ENGWALL, Y. LAPRIE, P. MARAGOS, B. POTARD and J. SCHOENTGEN: *Deliverable D1.1 Technology Inventory of Audiovisual-To-Articulatory Inversion*. http://aspi.loria.fr/Save/survey-1.pdf, 2008.

[9] PERKELL, J. S. and W. L. NELSON: *Variability in Production of the Vowels /i/ and /a/*. J. Acoust. Soc. Am., 77(5):1889–1895, 1985.

[10] STEVENS, K. N.: *On the Quantal Nature of Speech*. Journal of Phonetics, 17:3–45, 1989.

[11] SUNDBERG, J., B. LINDBLOM and J. LILJENCRANTS: *Formant Frequency Estimates for Abruptly Changing Area Functions: A Comparison Between Calculations and Measurements*. J. Acoust. Soc. Am., 91(6):3478–3482, 1992.

[12] WAKITA, H. and G. FANT: *Toward a Better Vocal Tract Model*. STL-QPSR, 19(1):9–29, 1978.

[13] XU, Y. and S. CHUENWATTANAPRANITHI: *Perceiving Anger and Joy in Speech Through the Size Code*. In *Proceedings 16th International Congress of Phonetic Sciences (ICPhS)*, pp. 2105–2108, Saarbrücken, August 2007.