

Prospects for Articulatory Synthesis: A Position Paper

Christine H. Shadle and Robert I. Damper

Image, Speech and Intelligent Systems Research Group,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK.

Abstract

Concatenative synthesis is currently the favoured approach to text-to-speech synthesis, yet it has fundamental limitations. In the longer-term, articulatory synthesis has much greater potential. Different approaches to articulatory synthesis are discussed in terms of the choices made concerning the articulatory processes modelled, the simplifying assumptions, and the data collected.

1. Introduction

Concatenative synthesis is now the leading approach in speech synthesis, based on numbers of researchers pursuing that approach and numbers of commercial speech synthesizers using it. Many have also asserted that it delivers more natural speech than formant synthesis (though some would debate that assessment). Yet authors of *The Bell Labs Approach* state: “We use concatenative synthesis because that is currently the best available method to produce synthetic speech of consistently high quality ... however, at the same time we also believe that in the long run concatenative synthesis is not the answer” [1, p. 3].

The very first fully automatic text-to-speech synthesizer was based on an articulatory synthesizer [2], and articulatory synthesizers continue to be developed, primarily as tools for research into fundamental aspects of speech production. However, the first synthesizers that were intelligible enough to be commercially useful were based on formant synthesis, and more recently, even formant synthesis has given way to concatenative synthesis. This has led to a generalisation about “knowledge-based” versus “ignorance-based” systems which mirrors that used in the field of speech recognition: whereas a knowledge-based system (articulatory or formant synthesis) would seem to be a better approach, the ignorance-based systems (concatenative) appear to be more successful. One might conclude that knowledge is all very well for understanding the process of speech production, and necessary for studying, for instance, disordered speech, but a hindrance for synthesis. Is this because we do not yet possess enough knowledge, or because an attempt to mimic the human method of producing something so

complex is doomed? Bailly goes so far as to write: “One could claim that this lack of theoretical background is the key to the success of sub-symbolic approaches in speech technology” [3, p. 160].

Synthetic speech of “consistently high quality” may not possess all the attributes we are interested in. The ideal synthesizer should:

1. be as intelligible as a human being.
2. sound natural.
3. be able to sound like many different speaker types: male, female; old, young, inbetween; low or high voice.
4. be able to speak in any language.
5. be able to sound like a specific speaker, not just a generic type.
6. be able to sound like an extraordinary speaker, e.g. a singer with a seven-octave voice range, or someone with disordered speech, or an alien with extra sinuses.
7. be able to change to another speaker type, or alter the voice quality of a given speaker, without having to go through as much effort as required for the first voice.
8. have parameter domains that can be conceptualised, so that if it sounds wrong, intuition is useful in fixing it.
9. teach us something and provide opportunities to learn more as we work to produce a commercially usable system.

A concatenative synthesizer with a sufficiently large database can satisfy the first four attributes, and can outperform a formant synthesizer on the third. Both types of synthesizer could satisfy the fifth if that specific speaker’s voice had been incorporated. An articulatory synthesizer is clearly the way to achieve attributes 6, 7, and 8. Regarding 9, although any synthesis approach teaches us

something, we know less about parameters and processes that are more internal and therefore harder to observe, and are thus likely to learn more through attempting articulatory synthesis.

This last point implies that *knowledge* and *ignorance*, used to classify approaches to synthesis, form a false dichotomy. As Scully stated: “Any attempt to produce synthetic speech is based on some view or model of how natural speech is made. Diphone synthesis, for example, exploits the fact that some portions of the acoustic signal change less rapidly and are less influenced by phonetic context than others. Acoustic-based synthesis has generally assumed that there are sources of sound which are separable from filter shapes and that the resonance processes linking them are linear” [4, p. 151]. The basic problems in designing a synthesizer can then be described as threefold: 1) choose the level of the basic parameter set; 2) choose the set of simplifying assumptions: what will be included, what ignored? 3) decide what kind of data are needed, and how the data should be used to build a database, create rules, and/or develop the model.

The choice made regarding the level of the basic parameter set will determine the dimensionality of the system, and also how easy it will be to gather data [5]. The choice of simplifying assumptions will determine the degree to which one can conceptualise the inner workings of the synthesizer, its degree of match to the real world, and also how easy it is to know what to fix when the synthetic speech sounds unintelligible or unnatural. It will also affect the degree of flexibility and the extent to which the synthesizer can be generalised. The data to be obtained follow as a consequence of the other choices made. For concatenative synthesis, the synthesizer is to a large extent as flexible and as natural as its database. For formant or articulatory synthesis, the data gathered chiefly serve to refine the models.

In this paper, we consider the three problems defined above in terms of articulatory synthesis, describing the progress made and the merits of various choices in each case. We then summarise what seem to be the most fruitful directions, and consider a few aspects that have received scant attention by those interested in speech synthesis. Finally, we conclude.

2. Approaches to Articulatory Synthesis

In this section, we do not attempt a full review of articulatory synthesis, but rather explore the choices that have been made and the problems encountered as a result.

2.1. Choice of basic parameter set

Articulatory parameters have an advantage in that they describe the system that produces the sound rather than the result of that process. Thus, extrapolating parameters beyond normal bounds should still produce a “correct”

sound (e.g. the way an eight-foot tall human would sound), whereas extrapolating the acoustic results may create an unnatural effect (e.g. lowering a front-cavity resonance without allowing for the sudden change when a sublingual cavity comes into existence). In concatenative synthesis, the same principle is at work, as exemplified by the burbling sounds created by abutting segments. In formant synthesis, interpolation in the acoustic domain amounts to a model of coarticulation that is fundamentally flawed. Articulatory synthesis should handle coarticulation correctly, since the articulators themselves are objects following the laws of physics. But because there is an essentially endless sequence of cause and effect, one can always find some mechanism that has not been itself modelled.

Within the domain of articulatory parameters, there are still choices to be made regarding the level of the basic parameter set. The line analogue model, as used in, for example, the DAVO synthesizer [6], uses cross-sectional area along the tract to determine the transmission-line parameters. Since it differs very little from a terminal analogue model, in which the resonances are modelled, the line analogue could be considered to be at a level closest to formant synthesis. In the line analogue model, however, there is a clear correlate within the model to distance along and area across the vocal tract; noise sources are inserted between sections and thus also have a clear spatial referent.

While many synthesizers use an area function-to-transfer function module that converts the vocal tract shape to an acoustic filter, the area function is generated in a variety of ways. In some cases, vocal tract measurements are used directly to generate the area function (e.g. [7]). In other cases, articulatory models generate a midsagittal profile that is then converted to an area function. These differ in terms of the parameters used (although all of them use approximately nine), the theoretical basis for that definition, and the degree to which interaction between structures is modelled [8, 9, 10].

At a further remove from formants, EMG signals were used as a basis for articulatory modelling. The initial assumption was that a correspondence existed between the phoneme and the EMG signals. According to Harris [11], this was a mistake. EMG signals are related to the distance and time needed for a muscle to move—not to the phoneme. Subsequent EMG studies also revealed large variation in the signals from token to token. Though EMG proved not to be a useful level at which to define a parameter set, that work combined with recent advances in medical imaging have led to recent approaches in which muscles, rather than articulator positions, are modelled. Dang and Honda [12] have modelled the tongue tissue, “roughly replicating the fiber orientation of the genioglossus muscle”; each section of the tongue model is driven by muscle activation patterns.

Wilhelms-Tricarico [13] is generating an exact model of every muscle, bone, and other anatomic structures, based on the Visible Human data sets. The resulting synthesized speech should be an exquisite example of attribute 5, depiction of a specific speaker; it remains to be seen how generalisable it is to other speakers (i.e. attribute 7).

While research effort on complete articulatory models has waxed and waned, that on articulatory models of vocal fold oscillation has been much more steady. Models have varied in level: the one-mass [14] and the two-mass model [15] both capture the gross mechanical and aerodynamic characteristics of vocal-fold oscillation. Multi-mass models have been explored (e.g. [16]) that could vibrate in other modal patterns observed in real vocal folds, but with a great increase in complexity (and reduction in conceptual power). Beam models with three degrees of freedom in their motion [17] and models that incorporate two-dimensional positioning of the vocal folds [18] are more recent approaches that generate more complex patterns of vibration, without substantially increasing the number of parameters.

Noise sources have received much less attention. Unlike the voice source, where the mechanical, aerodynamic and acoustic interaction clearly must be incorporated in the model, noise sources have often been constructed separately and inserted into an articulatory model that is otherwise limited to acoustics. Fant [7] took this approach, and more recent work based on mechanical models [19, 20] has refined the choice of ‘off-the-shelf’ noise sources available. Flanagan and Ishizaka [21] instead developed a model that generated its own noise sources wherever the local Reynolds number rose above a threshold. Scully’s synthesizer [4] likewise determined in the aerodynamic module where sources should be, and their strength, for use in the acoustic module. Recently, Sinder, Krane and Flanagan [22, 23] have developed a model that computes its own noise sources. While it is still limited to very simple tract shapes, it is much more flexible and should generalise far better than any parametric approach.

Finally, the aerodynamics of the tract as a whole has been modelled in various ways. Solutions to the Navier-Stokes equation were attempted when computing power rose to a point where that became feasible (e.g. [24, 25]), but this approach has largely been abandoned. The method used by Krane et al. [23] of modelling vortices and their sound production mechanisms seems promising.

2.2. Choice of simplifying assumptions

Consideration of noise-source modelling above is difficult to separate from issues of the simplifying assumptions made in a given system. In concatenative synthesis a modular approach is used; each module is visited only once [1]. While it is true that the process of

sound generation has already occurred, and the results are contained in the speech units in the database, the extraction of the excitation, its modification and subsequent reassembly embody an assumption of source-filter separation. In articulatory synthesis, likewise, the assumption is often made that aerodynamics can be separated from the acoustics.

Clearly this assumption of separation cannot be used for vocal fold oscillation, nor for trills; vocal fold models were discussed above, and trills have been elegantly modelled [26]. But even when source-filter interaction is not essential, it may still be important. For instance, it is possible to put a vocal-fold module that includes mechanical-aerodynamic-acoustic interaction into a system that excludes interaction between vocal fold oscillation and vocal tract impedance. Yet doing so affects the naturalness of the resulting synthetic speech [27]. Likewise, generating noise sources parametrically incorporates aerodynamic information, but certainly is less flexible and can result in woefully inadequate fricatives. Even a relatively subtle effect, the modulation of the noise source in voiced fricatives, appears to occur by means of a combined acoustic and aeroacoustic mechanism that could not easily be recreated in a model where aerodynamic and acoustic modules are separate [28].

Fant established that the area function, rather than the detailed cross-sectional shape of the tract, was sufficient to model speech sounds [7]. At the time of his work, and for many years after, it was difficult even to obtain an accurate area function: much research went into deriving the area from sagittal distances obtained by X-ray. Now that it is possible to obtain the area function accurately via magnetic resonance imaging (MRI), the first priority has been to rerun the models predicting the acoustic output from the area function, on the assumption that the articulatory data was the weakest link.

But the three-dimensional high-resolution MRI data are not being fully exploited. Tract asymmetries are revealed in normal speakers; some of these appear not to matter acoustically, but the extreme asymmetries that were already known to exist in, for example, cleft-palate speakers do matter. How can we decide when the area alone is not enough? Clearly the details of the shape downstream of the constriction affect noise generation. Evidence also exists that regions of large articulatory variability (e.g. palate vault depth) are handled differently, adaptively, by speakers in order to produce smaller acoustic variability [29]. In such a case, should the articulatory variability be included in a model, or should a typical palate shape and tongue contact pattern be adopted, with the assumption that for normal speakers, on the average, the acoustic result will be the same? The answer may well depend on whether we want to model a particular speaker, disordered speech, or simply generic normal speakers. Whatever the answer in a particular

case, we no longer need to, and can no longer afford to, make the blanket assumption that an area function is a sufficient description of the vocal tract shape.

Many algorithms exist for converting the area function to a transfer function. They differ in whether they operate in the time or frequency domain, or both; in the models of loss used, whether yielding walls are incorporated and how, and how side branches are treated. Most use a set of assumptions implicit in the representation of an acoustic system as an electrical analogue. One system that relaxes some of these assumptions, VOAC [30], also uses hydraulic radius in addition to area function, thus incorporating some information about the cross-sectional shape.

Particular studies have addressed the acoustic effect of the bend in the tract and the consequences of an elliptical versus circular cross-section [31]. Traditionally, cross-modes have been neglected, assuring that the acoustic output will be less accurate above 5 kHz. But recently they have been incorporated via parallel transmission lines [32], though the conceptual power of the electrical analogue is then considerably reduced.

2.3. Data collection

With formant synthesis, control of prosody seems to be the weakest link. With articulatory synthesis, vocal tract shape data seemed to be one of the chief problems, but with the advent of MRI, that is less true. We now have high-resolution three-dimensional shape data that very few acoustic models use fully. Tract-internal acoustic and aerodynamic signals are nearly as difficult to acquire as 3D shape data, but can serve as useful intermediate signals with which to check the performance of the synthesizer (as was done by Scully [4]).

Although X-rays can no longer be condoned on safety grounds, other types of vocal tract imaging exist. MRI was a big advance, despite the drawbacks of long image acquisition times and the difficulty of getting bone, especially the teeth, to appear different from air. Methods for outlining the teeth have had some success, as have fusing the images with those of dental impressions. The acquisition times keep improving, and different techniques have been developed to allow images taken of moving articulators to be recombined to yield much shorter effective frame rates [33, 34, 35]. The numerous repetitions required, however, mean that the speech cannot be said to be natural.

Other imaging techniques do allow for more natural speech: articulography and X-ray microbeam, which image midsagittal points on the tongue, electropalatography, which images tongue contact on the palate, and ultrasound of the tongue, which images any given plane of the tongue. While none of these has high spatial resolution or includes the complete vocal tract, together they have provided essential information about timing,

control, and tongue movement.

In vivo measurements of aerodynamic parameters are far more primitive. The Rothenberg mask and intraoral pressure measurements are useful and relatively noninvasive. Subglottal pressure and lung volume have been measured, mostly for studies of respiration and of singing, by invasive and/or less accurate methods. Hot wires have been used in the tract to measure flow velocity, but it is a hostile environment for hot wires; they are much better used in mechanical models where they can be calibrated and positioned accurately.

A variety of methods has been developed for measuring various vocal fold parameters, ranging from the noninvasive laryngoscope to high-speed filming of the vocal folds. While MRI cannot capture the moving vocal folds, it has been used to explore larynx movement in relation to f_0 control.

3. Other Points to Consider

Does the chosen form of the input limit the synthesizer's flexibility? For text-to-speech synthesis, the input is invariably some form of text (e.g. a phonemic string with stress markings). Yet this automatically rules out non-speech utterances that may be not only possible but natural (e.g. "Mm-hmm, unh-unh," yawning), less natural but possible utterances such as a place-changing continuum from /s/ to /ʃ/, and natural-sounding but impossible utterances such as a seven-octave sung scale, or a vowel produced while the vocal tract steadily lengthens. These could be straightforwardly specified in terms of articulatory parameters, so the input to the synthesis 'module' should not be so text-oriented as to preclude them.

While so many avenues are being explored, evaluation of systems is not straightforward. If the goal is commercial applications, then the synthesizer must be able to produce complete sentences, and the output should be evaluated on customer acceptance, intelligibility, etc. But articulatory synthesis is still far from that goal. A system that produces worse-quality synthesis, but is designed in such a way as to make it more flexible, may well be the better system overall. It is important to consider all of the attributes of an ideal system, even if only some of them are of importance for the current application. Likewise, it is important to allow some steps to produce worse results in order to explore paths that may allow for better synthesis in the end (cf. the argument of Boutilier, Hermansky and Morgan [36] in the context of speech recognition). Finally, systems based on an entirely different set of assumptions may be very difficult to compare. Varying one module while leaving others in a standard state (e.g. using different articulatory models to produce an area function, and a single model to produce a transfer function) may not always be a valid method.

It is fruitful to consider music synthesis. First, many

of the stages of the research have been similar. The initial stages focussed on imitation of existing acoustic instruments. The better a synthetic violin, the more control was demonstrated. Then, new instruments were synthesized that are not physically realisable, or would take too long to learn to play, or could be made but with great labour.

Second, the acoustic problems are harder, while the ‘articulatory’ problems are easier (most musical instruments, with the exception of the singing voice, have a constant shape). There are instruments with harmonic, and some with inharmonic, partials. Nonlinear acoustics is included in models much more often, since many musical instruments require it. Since each instrument demands that the type of model and the set of appropriate assumptions be approached anew, there is a greater variety in modelling. One result is that when singing is synthesized, the assumptions commonly made in speech synthesis are not necessarily assumed to hold.

Third, in speech synthesis the priorities are first to make the synthesizer intelligible, and then make it sound more natural. In music, the ‘information content’ (pitch, duration, amplitude, as specified in the score) is simple to control with a synthesizer, and the expressiveness is relatively more important: it is permissible to use vibrato, ritardando, even to sing a different vowel from that in the libretto, if these departures are done artistically. We can benefit from research done on how musicians, particularly singers, achieve such expressiveness. For example, singers allow subglottal pressure to vary more widely, and larynx position to vary less than in speech, controlling f_0 and amplitude in ways that produce a more beautiful and sustainable tone (voice quality) [37]. We can also consider baffling aspects of speech synthesis from a musical point of view: could intonation on a paragraph level be treated as having a musical structure? Studies on different singing styles and expressiveness may be usefully tapped for speech synthesis with a range of voice qualities and emotional states.

4. Conclusions

The detailed consideration of the attributes of an ideal speech synthesizer, and of the current state of articulatory models and synthesis, developed in this paper lead us to the following conclusions.

- Ultimately, concatenative synthesis is not the answer. In the long term, articulatory synthesis has more potential, not only for extending our knowledge of speech science, but for high-quality speech synthesis.
- Choosing the optimal level on which an articulatory synthesizer would operate is one of the major problems.

- Although a pipelined set of independent modules makes evaluation much more straightforward, a realistic articulatory synthesizer demands interconnections and feedback as well as feedforward paths. Choosing the interconnections to include is one of the major problems.
- Lousy speech can be a good thing. That is to say, insisting that all changes must lead to immediate improvement may prevent momentary steps backwards that allow for major improvement.
- We need more knowledge about many fields: motor control, vocal tract imaging and images of the tract during speech, aeroacoustics, speech disorders, singing, emotional states and their effect on speech.

In the long run, efforts devoted to concatenative synthesis may be efforts wasted.

5. References

- [1] J. van Santen and R. Sproat. Introduction. In R. Sproat, editor, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, pages 1–6. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
- [2] N. Umeda, E. Matsui, T. Suzuki, and H. Omura. Synthesis of fairy tales using an analog vocal tract. In *Proceedings of 6th International Congress on Acoustics*, pages B159–162, Tokyo, Japan, 1968.
- [3] G. Bailly. Introduction to Part III: Prosody in Speech Synthesis. In Y. Sagisaki, N. Campbell, and N. Higuchi, editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pages 157–164. Springer, New York, NY, 1997.
- [4] C. Scully. Articulatory synthesis. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 151–186. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.
- [5] R. Togneri, M. D. Alder, and Y. Attikiouzel. Dimension and structure of the speech space. *IEE Proceedings. I: Communications, Speech and Vision*, 139(2):123–127, 1992.
- [6] G. Rosen. A dynamic analog speech synthesizer. *Journal of the Acoustical Society of America*, 30:201–209, 1958.
- [7] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, The Netherlands, 1960.
- [8] C. H. Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64:452–460, 1976.
- [9] U. Goldstein. *An Articulatory Model for the Vocal Tracts of Growing Children*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980.
- [10] P. Rubin, E. Saltzman, L. Goldstein, R. S. McGowan, M. Tiede, and C. Browman. CASY and extensions to the task-dynamic model. In *Proceedings of 1st ESCA ETRW on Speech Production Modelling and 4th Speech*

- Production Seminar*, pages 125–128, Autrans, France, 1996.
- [11] K. S. Harris. Speech research from acoustic transmission to gestural analysis. *Journal of the Acoustical Society of America*, 103(5(2)):3024, 1998. Abstract.
- [12] J. Dang and K. Honda. Estimation of vocal tract shape from speech sounds via a physiological articulatory model. In *Proceedings of 5th Seminar on Speech Production: Models and Data*, pages 233–236, Kloster Seeon, Germany, 2000.
- [13] R. Wilhelms-Tricarico. Development of a tongue and mouth floor model for normalization and biomechanical modeling. In *Proceedings of 5th Seminar on Speech Production: Models and Data*, pages 141–144, Kloster Seeon, Germany, 2000.
- [14] J. L. Flanagan and L. L. Landgraf. Self-oscillating source for vocal-tract synthesizers. *IEEE Transactions on Audio and Electroacoustics*, AU-16:57–64, 1968.
- [15] K. Ishizaka and J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Systems Technical Journal*, 50:1223–1268, 1972.
- [16] I. Titze and D. T. Talkin. A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *Journal of the Acoustical Society of America*, 66(1):60–74, 1979.
- [17] J. Liljencrants. A translating and rotating mass model of the vocal folds. *Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology (KTH), Stockholm*, STL-QPSR 1/1991:1–18, 1991.
- [18] I. Titze. Biomechanical modeling of vocal fold posturing. In *Proceedings of 5th Seminar on Speech Production: Models and Data*, pages 193–196, Kloster Seeon, Germany, 2000.
- [19] C. H. Shadle. Articulatory-acoustic relationships in fricative consonants. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 187–209. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.
- [20] S. Narayanan and A. Alwan. Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing*, 8(3):328–344, 2000.
- [21] J. L. Flanagan and K. Ishizaka. Automatic generation of voiceless excitation in a vocal cord-vocal tract speech synthesizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24:163–170, 1976.
- [22] D. J. Sinder, M. H. Krane, and J. L. Flanagan. Synthesis of fricative sounds using an aeroacoustic noise generation model. In *Proceedings of 16th International Congress Acoustics*, volume 1, pages 249–250, Seattle, WA, 1998.
- [23] M. H. Krane, D. J. Sinder, and J. L. Flanagan. Aeroacoustic modeling of speech sound production. In *Proceedings of 5th Seminar on Speech Production: Models and Data*, pages 177–180, Kloster Seeon, Germany, 2000.
- [24] T. J. Thomas. *An Articulatory Model of Speech Production Including Turbulence*. PhD thesis, University of Cambridge, Cambridge, UK, 1985.
- [25] J. Liljencrants. Numerical simulations of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology (KTH), Stockholm*, STL-QPSR 1/1991:69–74, 1989.
- [26] R. S. McGowan. Tongue-tip trills and vocal-tract wall compliance. *Journal of the Acoustical Society of America*, 91(5):2903–2910, 1992.
- [27] M. Rothenberg. Acoustic interaction between the glottal source and the vocal tract. In K. N. Stevens and M. Hirano, editors, *Vocal Fold Physiology*, pages 305–328. University of Tokyo Press, Tokyo, Japan, 1981.
- [28] P. J. B. Jackson and C. H. Shadle. Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *Journal of the Acoustical Society of America*, 108(4):1421–1434, 2000.
- [29] J. Perkell, M. Matthies, R. Wilhelms-Tricarico, H. Lane, and J. Wozniak. Speech motor control: Phonemic goals and the use of feedback. In *Proceedings of 1st ESCA ETRW on Speech Production Modelling and 4th Speech Production Seminar*, pages 133–136, Autrans, France, 1996.
- [30] P. A. O. L. Davies, R. S. McGowan, and C. H. Shadle. Practical flow duct acoustics applied to the vocal tract. In I. R. Titze, editor, *Vocal Fold Physiology: Frontiers in Basic Science*, pages 93–142, San Diego, CA, 1993. Singular Publishing.
- [31] H. Matsuzaki and K. Motoki. FEM analysis of 3-D vocal tract model with asymmetrical shape. In *Proceedings of 5th Seminar on Speech Production: Models and Data*, pages 329–332, Kloster Seeon, Germany, 2000.
- [32] K. Motoki, P. Badin, X. Pelorson, and H. Matsuzaki. A modal parametric method for computing acoustic characteristics of three-dimensional vocal tract models. In *Proceedings of 5th Seminar on Speech Production: Models and Data*, pages 325–328, Kloster Seeon, Germany, 2000.
- [33] M. Stone, A. Lundberg, E. Davis, R. Gullapalli, and M. NessAiver. Three-dimensional coarticulatory strategies of tongue movement. In *Proceedings of 5th European Conference on Speech Communication and Technology, Eurospeech'97*, volume 1, pages 31–34, Rhodes, Greece, 1997.
- [34] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, and Y. Nakamura. MRI observation of dynamic articulatory movements using a synchronized sampling method. *Journal of the Acoustical Society of America*, 102(5(2)):3166, 1997. abstract.
- [35] C. H. Shadle, M. Mohammad, J. N. Carter, and P. J. B. Jackson. Multi-planar dynamic magnetic resonance imaging: New tools for speech research. In *Proceedings of the XIVth International Congress of Phonetic Sciences, ICPhS'99*, volume 1, pages 623–626, San Francisco, CA, 1999.
- [36] H. Bourlard, H. Hermansky, and N. Morgan. Towards increasing speech recognition error rates. *Speech Communication*, 18(3):205–231, 1996.
- [37] J. Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, DeKalb, IL, 1987.