A Hybrid Acoustic-Articulatory Model of the Speech Spectrum

J. Schoentgen(*), A. Kacha, F. Grenez
L.I.S.T., CP 165/51, Université Libre de Bruxelles, Brussels, Belgium
(*) National Fund for Scientific Research, Belgium
E-mail: jschoent@ulb.ac.be, akacha@ulb.ac.be, fgrenez@ulb.ac.be

Abstract

1 Introduction

The framework of the present article is spectrum-to-area mapping. The area function is the cross-section of the vocal tract as a function of the distance from the glottis. The goal is to infer an area function model from observed spectra so that the model transfer function approximates the observed spectral contour as accurately as possible. This objective is more ambitious than mapping from formant frequencies because spectra do not only depend on the shape of the main tract cavity, but also on side-cavities and their connectivities as well as acoustic sources and losses.

Earlier experiments, which are not reported here, with area function models have shown that observed spectral contours can be recovered only if no extra formant & anti-formant pairs (owing to side cavities) are observed in the spectrum and the description of the spectral contour is rough (i.e. typically 6 spectral intervals). Even then the recovered tract shapes may be biased owing to a priori assumptions with regard to the vocal tract losses or sound sources that may not agree with the speaker's loss and source characteristics.

This presentation reports preliminary results of a mapping method that takes into account that vocal tract losses, sound sources and sizes and connectivities of side cavities cannot be known a priori and that their models cannot be adapted to any speaker. The proposal consists in turning the spectrum→area problem into a formant→area problem via a hybrid model that combines a quasi-lossless area function model, which generates formant frequencies, with a spectral model that simulates other features such as extra formant & anti-formant pairs, formant bandwidths as well as the spectral contour of the speech sound source. The goal is to examine whether a) such an approach may enable inferring hybrid models the transfer functions of which fit observed spectra accurately and b) under what conditions the affiliated tract shape models may be anatomically plausible.

2 Models

2.1 Area function model

The sound propagation in a quasi-lossless truncated cone may be described by means of the relation between input and output acoustic pressures and volume velocities. This relation is given by a 2×2 transfer matrix that spectrally describes the wave propagation, which is assumed to be planar [3]. To simulate a vocal tract area function, several truncated cones may be concatenated. The transfer matrix of the concatenation is obtained by multiplying the individual transfer matrices, which yields a 2×2 matrix that relates the acoustic pressures and volume velocities at the lips and glottis [1].

The eigenmode conditions are that the volume velocity at the glottis and the acoustic pressure at the lips are zero. Inserting these requirements into the transfer matrix gives rise to the condition that the bot-

tom right matrix element is 0. In practice, the eigenmode frequencies are therefore found by numerically searching for all frequency values that zero that matrix element.

2.2 Spectral model

The spectral model is a conventional linear pole & zero model, which in the spectral domain is represented by a ratio of two polynomials, the roots of which are either real or paired complex conjugate. The roots of the numerator fix the spectral zeros and the roots of the denominator the spectral poles.

2.3 Hybrid model

The hybrid model combines the area function model with the spectral model. For the experiments reported here, the number of real poles and real zeros is set to 2. They simulate the glottal formant and spectral slope owing to the speech sound source(s). The number of excess formants and anti-formants is also set to 2, because the poles and zeros affiliated with side-cavities appear pair-wise [4] and one observes up to two anti-formants in the spectra of some of the sounds that have been analysed.

3 Methods

3.1 Corpus

The corpus comprises V_1V_2 and V_1CV_2 disyllables spoken by one French male speaker. Vowels in V_1V_2 tokens are the qualities [a] [ɛ] [e] [i] [u] [o] [ɔ] [y] [ø] and [œ]. Vowels in V_1CV_2 are the qualities [a] [i] [u] and consonants C are [p] [t] [k] [f] [s] and [ʃ]. The total number of tokens is 89. Before hybrid modelling they are down-sampled to $8kHz$.

3.2 Speech analysis

Discrete Fourier spectra are computed for each frame. The analysis window is a Hamming window that is hopped by $10ms$. If required, the spectral contour is estimated via cepstral analysis, which consists in computing the inverse Fourier transform of the log-magnitude spectrum, zeroing all cepstral samples above $1.875ms$ and re-computing the Fourier transform.

3.3 Analysis-by-synthesis

The hybrid model is fitted by minimising a cost function. The optimiser is the Differential Evolution Algorithm, which belongs to the family of evolutionary algorithms that involve stochastically evolving populations of solutions. One feature of Differential Evolution Algorithms is that the model parameters that evolve are encoded numerically as floating-point numbers (instead of binary numbers) [2]. Here, the size of the evolving population is equal to $3\times$ the number of parameters that are being adjusted. The so-called cross-over probability is equal to 1 and the so-called differential mutation is in the interval $0.25...0.75$. The algorithm halts when the value of the cost function or the variance of the population fitness is smaller than a threshold.

3.4 Cost function, hard and soft constraints

The cost function that is minimised is the sum of the squared differences between the observed log-magnitude spectrum and the log-magnitude transfer function of the hybrid model, which are z-normalized. When the target is a spectral contour, it is down-sampled into 30 spectral bins.

Hard constraints guarantee that the pole and zero frequencies and bandwidths as well as the tract cross-sections remain within plausible intervals by rejecting model parameters that do not comply.

Soft constraints guide the optimiser towards solutions that are considered desirable. One soft constraint requests that the vocal tract volume stays within $\pm 10\%$ of a reference volume of $77cm^3$. A second soft constraint limits the relative frame-to-frame distance between cross-sections. That is, the cross-sections are expected to evolve evenly over a time interval of $10ms$, which is the analysis frame hop.

3.5 Experiments

The results section reports two preliminary experiments. The first involves the fitting of the hybrid model to the spectral contour estimated via cepstral smoothing. The hybrid model includes 6 formants, 4 are assigned to the area function that models the main vocal tract cavity and the 2 remaining formants are allocated to the 2 anti-formants with which they constitute 2 formant $\&$ anti-formant pairs. The other model components are as described in section 2. The analysis window length is $50ms$. The cross-sections are soft-constrained to evolve by less than 10% from one frame to the next.

The second experiment involves a hybrid model that comprises 4 formants only, which are exclusively assigned to the area function model. The

Table 1: *Relative* dB *(p 1).*

%	20	40	60	80	100
Min	-28.4	-27.5	-27.0	-26.5	-24.5
1. Quart.	-25.5	-24.8	-23.6	-22.4	-19.1
Median	-23.1	-21.3	-20.7	-19.6	-17.3
3. Quart.	-19.8	-18.6	-18.0	-17.1	-14.0
Max	-13.9	-12.3	-11.3	-9.4	-4.8

2 supernumerary formants that are theoretically requested for the 2 formant & anti-formant pairs are assumed to be fixed and to have a bandwidth that is so large that they do not contribute to the model transfer function. In practice, they can therefore be omitted from the model. The model is $directly$ fitted to the spectrum (the estimation of the spectral contour is omitted). The analysis frame length is $25ms$. The cross-sections are soft-constrained to evolve by less than 5% from one frame to the next.

The total number of analysis frames is 4053 for 89 speech tokens. For each frame, one obtains the modelling error in dB, the relative frame-to-frame distance between cross-sections and the relative vocal tract volume.

For the second experiment, the ratio of the spectral flatness of the unprocessed spectrum and residue spectrum are reported for a sub-set of speech tokens. The residue spectrum is the difference between the log-magnitude model transfer function and the unprocessed spectrum. The spectral flatness is the geometric average of the magnitude spectrum divided by the arithmetic average. The reason for calculating the spectral flatness is that it reports indirectly on the capability of the hybrid model to recover the spectral contour. Indeed, when one disregards the harmonics and noise, the contour of the residue spectrum is ideally horizontal.

4 Results

4.1 Experiment 1

For all Tables, the columns labelled 20, 40, $etc.$ report the minimum, first quartile, median, third quartile and maximum below which 20%, 40%, $etc.$ of the cue values of all the tokens can be found. Table 1 reports in dB the square root of the ratio of the residual energy divided by the spectral energy. Small dB values therefore correspond to small modelling errors.

Table 2 reports the relative frame-to-frame distance between cross-sections. One observes that, except for the last row, the inter-frame distance is smaller than 10%, as requested.

Table 2: *Relative frame-to-frame distance (p 1).*

%	20	40	60	80	100
Min	0.05	0.05	0.06	0.07	0.08
1. Quart.	0.09	0.10	0.10	0.10	0.10
Median	0.10	0.10	0.10	0.10	0.10
3. Quart.	0.10	0.10	0.10	0.10	0.10
Max	0.35	0.38	0.43	0.47	0.65

Table 3: *Relative (p 1).*

%	20	40	60	80	100
Min	0.90	0.90	0.90	0.90	0.90
1. Quart.	0.90	0.90	0.90	0.90	0.92
Median	0.91	0.92	0.92	0.93	0.96
3. Quart.	0.94	0.95	0.96	0.97	1.01
Max	1.02	1.04	1.06	1.08	1.10

Table 3 finally shows the relative vocal tract volume. One observes that the volume remains within $\pm 10\%$ of a reference volume, as requested.

4.2 Experiment 2

Table 4 reports the square root of the ratio of the residual energy and the spectral energy in dB. Compared to Table 1, the relative errors are larger because the target has been the whole spectrum including the harmonics or noise, which are not simulated by the hybrid model.

An alternative description of the faithfulness of the modelling of the contour is the spectral flatness of the residue. Indeed, the flatter the residue the more faithful the modelling is. Table 5 shows the ratio of the spectral flatness of the residue and unprocessed magnitude spectra, which is a measure of the relative increase of the spectral flatness of the residue. One sees that the residue is always flatter than the unprocessed magnitude spectrum. Visual inspection confirms that the model transfer functions take on the shape of the spectral contour.

Table 6, finally, reports the relative frame-to-frame distance between cross-sections. It shows that, except for the last row, the frame-to-frame distance between cross-sections has been less than 5%. Distances $> 5\%$ are due to the initialisation of the analysis-by-synthesis via a quasi-uniform tract. Visual inspection confirms that for token-internal frames, the frame-to-frame distance has always been below 5%, as requested. Finally, the relative tract

Table 4: *R* dB (*p*2).

%	20	40	60	80	100
Min	-7.5	-6.9	-6.4	-5.9	-4.8
1. Quart.	-5.0	-4.4	-4.1	-3.8	-3
Median	-3.8	-3.5	-3.2	-2.8	-2.1
3. Quart.	-2.9	-2.5	-2.3	-2.0	-0.1
Max	-1.2	-0.7	-0.4	0.3	1.8

Table 5: *R* (*p*2).

%	20	40	60	80	100
Max	6.2	12	18.7	72.3	247.2
3. Quart.	5.4	7.5	12.0	15.7	31.0
Median	4.4	5.6	6.5	10.4	13.9
1. Quart.	3.7	4.2	4.7	6.1	11.1
Min	3.3	3.5	3.7	4.0	4.5

volume has always been within $\pm 10\%$ of the reference volume, as requested. Volume results are not listed here due to lack of space.

5 Discussion and conclusion

The inter-frame distances in the last row of Table 2, which are larger than 10% have two distinct causes. The first is that the recovered tract shape for the first analysis frame of any token is compared to a quasi-uniform tract shape that launches the analysis-by synthesis.

The second cause is token-internal. Indeed, one observes that for isolated frames, the cross-sections evolve from one frame to the next by an anatomically unlikely large distance. For these frames, the evolving tract shapes are anatomically unlikely or unrealisable. A possible explanation is a sudden re-affiliation between formants and tract cavities. Indeed, in the framework of Experiment 1, $four$ formants have been actually affiliated with the main tract cavity, and two have been virtually affiliated with side-cavities. The hard and soft constraints appear not to be able to prevent occasional re-affiliations between formants and cavities.

Therefore, in Experiment 2 supernumerary formants have been left out and all formants that are time-variable are affiliated with the main cavity. This has a beneficial effect on the anatomical plausibility of the inferred shapes. Visual inspection of the results as well as Tables 4 to 6 confirm that the frame-to-frame distance between cross-sections is less than 5% (token-boundary effects put aside), the volume

Table 6: *R* (*p*2).

%	20	40.	60	80.	100
Min	0.034	0.038	0.041	0.043	0.048
1. Quart.	0.048	0.049	0.049	0.049	0.050
Median	0.049	0.049	0.050	0.05	0.050
3. Quart.	0.050	0.050	0.050	0.05	0.050
Max	0.090	0.091	0.171	0.257	0.489

of the main tract cavity remains within $\pm 10\%$ of the volume of the neutral tract cavity and the spectral flatness of the residue spectrum is $5-10$ times larger than the spectral flatness of the unprocessed spectrum. Visual inspection confirms that spectral contours are recovered faithfully.

One may therefore conclude that Experiment 2 demonstrates that the hybrid model, together with the signal processing options and the hard and soft constraints, enables recovering tract shapes that satisfy necessary conditions for anatomical plausibility.

Anatomical likeness per se has not been tested because the original tract shapes have not been available. Also, the area function model that has been used is highly stylised.

At present, the study of feasibility that is reported here shows that hybrid models are flexible enough to combine the ability of simulating any spectral contour with the possibility to infer plausible tract shapes for a subset of phonetic segments.

Acknowledgements

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324.

References

[1] A. Kacha, F. Grenez, and J. Schoentgen. Anatomical plausibility of area functions inferred by analytic formant-to-area mapping. In *P* pages 897–900, 2007.

[2] K. Price, R. Storn, and J. Lampinen. *D* Springer, 2005.

[3] G. P. Scavone. *A* PhD thesis, Standford: Standford University, 1997.

[4] K. N. Stevens. *A* The MIT Press, Cambridge, 1998.