A Quantitative Model for Formant Dynamics and Contextually Assimilated Reduction in Fluent Speech

Li Deng, Dong Yu, and Alex Acero

Microsoft Research
One Microsoft Way, Redmond, WA 98052, U.S.A.
{deng,dongyu,alexac}@microsoft.com

Abstract

A quantitative model of coarticulation is presented that accurately predicts formant dynamics in fluent speech using the prior information of resonance targets in the phone sequence, in absence of actual acoustic data. Realistic formant undershoot (reduction) and "static" sound confusion is produced naturally from the model for fast-rate speech in a contextually assimilated manner. The model developed is capable of resolving the confusion with dynamic speech specification. As a source of a-priori knowledge about the speech structure, the model is a central component of our Bayesian generative modeling approach to automatic recognition of conversational speech, where varying degrees of sound reduction abound due to the free-varying speaking style and rate. We present details of the model simulation that demonstrates quantitative effects of speaking rate and segment duration on the magnitude of reduction, agreeing closely with experimental measurement results in the acoustic-phonetic literature. The model simulation also gives quantitative effects of varying the "stiffness' parameter in the model.

1. Introduction

Dynamic patterns for the spectral prominences or formants in fluent speech, especially in vowel sounds, have been a subject of intensive research in phonetics and in speech synthesis for many years [1, 2, 3, 4, 5, 6, 7, 8]. The research has been focusing on the central issue that the same formant values taken from the middle portion of a speech sound from its dynamic pattern can correspond to different sound classes specified solely in static terms. This inherent "static" confusion of speech classes without dynamic aspects of speech sound specification is believed to be one significant factor impeding current HMM-based speech recognition for casual-style, conversational speech.

In this paper, we present a modeling approach to dynamic specification of speech sounds, where the observed dynamic pattern of speech is the result of an interaction among phonetic context, speaking rate/duration, and spectral rate of change as related to speaking style [4]. The quantitative model that we have developed and will be presented in this paper assumes that each speech sound is specified by a largely context-independent target distribution in the vocal tract resonance (formant) space, together with a dynamic, *stiffness* parameter specifying how formant trajectories may be formed in any specific phonetic and prosodic environment. In the implementation of the model, the stiffness parameter is used to control temporal filtering of the sequentially arranged, statistically sampled formant targets, and is dependent on a range of prosodic factors, speaking style in particular. The result of the temporal filtering, in both forward and backward directions, gives rise to the phonetically realized dynamic formant patterns. A direct consequence of this filtering operation is as follows: the shorter a segment is, the greater the difference becomes between the filter's input (target formant values) and the output (observed formant values).[1] Therefore, our model naturally simulates the target-undershooting, or reduction phenomenon [1, 4, 6]. Because the input to the filter is the phonetically composed, discontinuous target sequence, which is smoothed by the filter resulting in continuous, "reduced" trajectories, our filter-based model represents the reduction phenomenon in a contextually assimilated manner. This is similar to the mechanism suggested in [1, 4].

In the remaining of this paper, details of the model construction are presented in Section 2. A series of model simulation results are provided in Section 3 and conclusions drawn in Section 4.

2. Model of formant dynamics and reduction

The model presented here for formant dynamics directly exhibits contextually assimilated reduction. The model is constructed using a (slowly time-varying) finite impulse response (FIR) filter characterized by the following non-causal impulse response function:

$$h_s(k) = \begin{cases} C\gamma_{s(k)}^{-k} & -D < k < 0 \\ C & k = 0 \\ C\gamma_{s(k)}^{k} & 0 < k < D \end{cases} \quad (1)$$

where k represents time frame, typically with a length of 10 msec each. $\gamma_{s(k)}$ is the "stiffness" parameter, and is positive and real-valued, ranging between zero and one.[2] The subscript $s(k)$ in $\gamma_{s(k)}$ indicates that the stiffness parameter is dependent on the segment state $s(k)$ on a moment-by-moment, time-varying basis. D in (1) is the unidirectional length of the impulse response. It represents the temporal extent of *coarticulation*, assumed for simplicity to be equal in length for the forward direction (anticipatory coarticulation) and for the backward direction (regressive coarticulation).

In (1), C is the normalization constant to ensure that the filter weights add up to one. This is essential for the model to produce target "undershoot", instead of "overshoot". To determine C, we note first:

$$\sum_{k=-D}^{D} h_s(k) = C \sum_{k=-D}^{D} \gamma_{s(k)}^{|k|} = 1. \quad (2)$$

[1] The precise difference also depends on the stiffness parameter associated with speaking style, in addition to the dependency on the filter input.

[2] In this paper, γ is treated as a deterministic quantity for simplicity purposes. In the more comprehensive version of the model, γ_s is a Gaussian random vector characterized by the (automatically learned) mean vector and covariance matrix.

For simplicity, we make the assumption that over the temporal span of $-D \leq k \leq D$, the stiffness parameter's value stays approximately constant

$$\gamma_{s(k)} \approx \gamma$$

That is, the adjacent segments within the temporal span of $2D+1$ in length which contribute to the coarticulated home segment have a similar stiffness parameter value to that of the home segment. Under this assumption, we simplify (2) to

$$C \sum_{k=-D}^{D} \gamma_{s(k)}^{|k|} \approx C[1+2(\gamma+\gamma^2+...+\gamma^D)] = C\frac{1+\gamma-2\gamma^{D+1}}{1-\gamma}.$$

Thus,

$$C(\gamma) \approx \frac{1-\gamma}{1+\gamma-2\gamma^{D+1}}. \quad (3)$$

The input to the system is the target sequence (discontinuous function) represented as a sequence of step-wise functions with variable durations and heights:

$$T(k) = \sum_{i=1}^{P}[u(k-k_{s_i}^l) - u(k-k_{s_i}^r)] \times T_{s_i}, \quad (4)$$

where $u(k)$ is the unit step function, $k_s^r, s = s_1, s_2, ..., s_P$ are the right boundary sequence of the segments (P in total) in the utterance, and $k_s^l, s = s_1, s_2, ..., s_P$ are the left boundary sequence. The difference of the two gives the duration sequence. $T_s, s = s_1, s_2, ..., s_P$ are the target values for the segments.[3]

In the work presented in this paper, we assume that both left and right boundaries (and hence the durations) of all the segments in an utterance are known (e.g., those provided in TIMIT database). However, in general cases where the current model is used to predict the formant trajectories as the FIR's filter's output, the boundaries in the target sequence input to the filter are not given. They either need to come from a recognizer's forced alignment results, or to be learned automatically using algorithms such as those described in [10].

Given the filter's impulse response and the input to the filter as described above, the filter's output as the model's prediction for the formant trajectories is the convolution between these two signals. The result of the convolution within the boundaries of the home segment s is

$$g_s(k) = h_{s(k)} * T(k) = \sum_{\tau=k-D}^{k+D} C(\gamma_{s(\tau)}) T_{s(\tau)} \gamma_{s(\tau)}^{|k-\tau|}, \quad (5)$$

where the input target value and the filter's stiffness parameter value may take not only those associated with the current home segment, but also those associated with the adjacent segments. The latter case happens when the time τ in (6) goes beyond the home segment's boundaries; i.e., when the segment $s(\tau)$ occupied at time τ switches from the home segment of an adjacent one.

A sequential concatenation of all outputs $g_s(k), s = s_1, s_2, ..., s_P$ in (5), each corresponding to a single segment in the utterance, constitutes the model prediction of formant trajectories for the entire utterance:

$$g(k) = \sum_{i=1}^{P}[u(k-k_{s_i}^l) - u(k-k_{s_i}^r)] \times g_{s_i}(k) \quad (6)$$

[3] In a more comprehensive version of the model, the target values are drawn from a statistical distribution, whose parameters (e.g., means and variances) are automatically learned in a manner similar to [9].

Note that the convolution operation carried out by the filter in the model guarantees continuity of the trajectories at each junction of two adjacent segments, contrasting the discontinuous jump in the input to the filter at the same junction. This continuity applies to all classes of speech sounds including consonantal closure.

3. Results of model prediction and formant measurements

In this section, we present the model simulation results demonstrating contextually assimilated reduction, and compare the results with the corresponding results from direct formant measurements in the literature.

To illustrate formant undershoot, we first show the spectrogram of three renditions of a three-segment /iy aa iy/ (uttered by one author of this paper) in Fig. 1. From left to right, the speaking rate increases (and speaking effort decreases), with the durations of the /aa/'s decreasing from approximately 230 msec to 130 msec. Formant target undershoots for F1 and F2 are clearly shown in the spectrogram (where automatically tracked formants are superimposed).

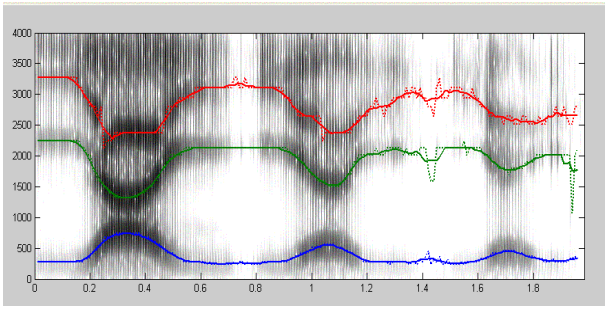

Figure 1: Spectrogram of three renditions of /iy aa iy/ by one author, with an increasingly higher speaking rate and increasingly lower speaking efforts.

3.1. Effects of stiffness parameter on reduction

The same kind of target undershoot for F1 and F2 as in Fig. 1 is exhibited in the model prediction, shown in Fig. 2, where we also illustrate the effects of the FIR filter's stiffness parameter on the magnitude of formant undershoot or reduction. The model prediction is the FIR filter's output for F1/F2 according to $g(k)$ in (6). Fig. 2a, b, and c correspond to the use of the stiffness parameter value set at $\gamma = 0.85, 0.75$ and 0.65, respectively, where in each plot the slower /iy aa iy/ sounds (with the duration of /aa/ set at 230 msec or 23 frames) are followed by the faster /iy aa iy/ sounds (with the duration of /aa/ set at 130 msec or 13 frames). F1 and F2 targets for /iy/ and /aa/ are set appropriately in the model also. Comparing the three plots, we have the model's quantitative prediction for the magnitude of reduction during the faster /aa/ that is decreasing as the γ value decreases.

In Fig. 3a, b, and c, we show the same model prediction as in Fig. 2 but for different sounds /iy eh iy/, where the targets for /eh/ are much closer to those of the adjacent sound /iy/ than in the previous case for /aa/. As such, the absolute amount of reduction becomes smaller. However, the same effect of the filter parameter's value on the size of reduction is shown as for the previous sounds /iy aa iy/.

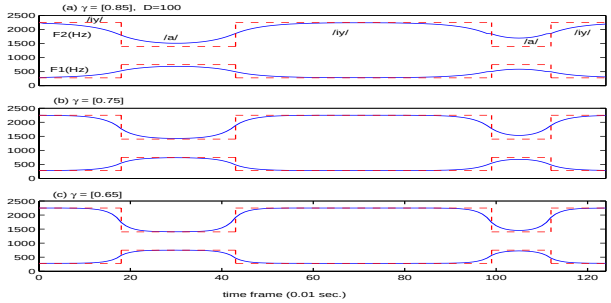

Figure 2: F1 and F2 formant trajectories produced from the model ($g(k)$ in (6)) for a slow /iy aa iy/ followed by a fast /iy aa iy/. (a), (b), and (c) correspond to the use of the stiffness parameter values of $\gamma = 0.85, 0.75$ and 0.65, respectively. The amount of formant undershoot or reduction during the fast /aa/ is decreasing as the γ value decreases. The dashed lines indicate the formant target values and their switch at the segment boundaries.

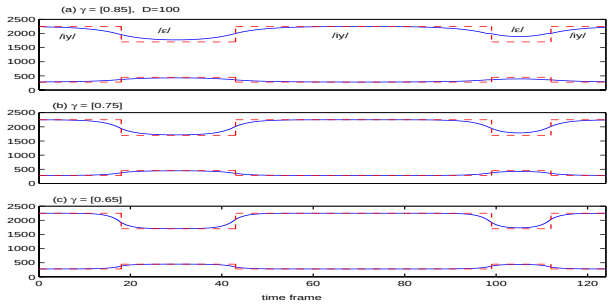

Figure 3: Same as Figure 2 except for the /iy eh iy/ sounds. Note that the F1/F2 target values for /eh/ are closer to /iy/ than those for /aa/.

3.2. Effects of speaking rate on reduction

In Fig. 4a, b, and c, we show the effects of speaking rate, measured as the inverse of the sound segment's duration, on the magnitude of formant undershoot. (a), (b), and (c) correspond to three decreasing durations of the sound /aa/ in the /iy aa iy/ sound sequence. The plots illustrate an increasing amount of the reduction with the decreasing duration or increasing speaking rate. Symbol 'x' in Fig. 4 indicates the F1/F2 formant values at the central portions of vowels /aa/, which are predicted from the model and are used to quantify the magnitude of reduction. These values (separately for F1 and F2) for /aa/ are plotted against the inversed duration in Fig. 5, together with the corresponding values for /eh/ (i.e. ϵ) in the /iy eh iy/ sound sequence. The most interesting observation is that as the speaking rate increases, the distinction between vowels /aa/ and /eh/ gradually diminishes if their static formant values extracted from the dynamic patterns are used as the sole measure for the difference between the sounds. We refer to this phenomenon as the "static" sound confusion induced by increased speaking rate (or/and by a greater degree of sloppiness in speaking).

3.3. Comparisons with formant measurement data

The "static" sound confusion between /aa/ and /eh/ quantitatively predicted by the model as shown in Fig. 5 is consistent with the

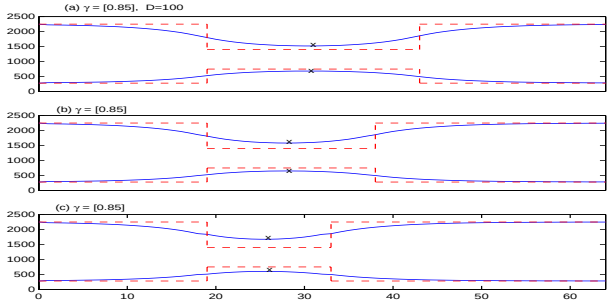

Figure 4: F1 and F2 formant trajectories produced from the model for three different durations of /aa/ in the /iy aa iy/ sounds: (a) 25 frames (250 ms), (b) 20 frames, and (c) 15 frames. The same γ value of 0.85 is used. The amount of target undershoot increases as the duration is shortened or the speaking rate is increased. Symbol 'x' indicates the F1/F2 formant values at the central portions of vowels of /aa/.

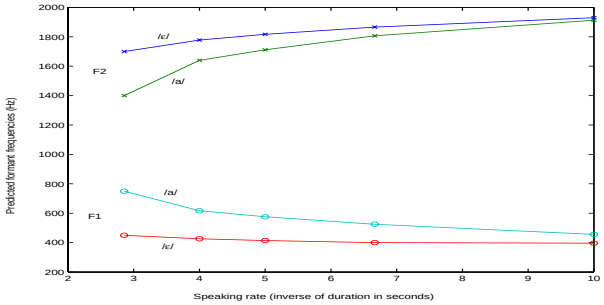

Figure 5: Relationship, based on model prediction, between the F1/F2 formant values at the central portions of vowels and the speaking rate. Vowel /aa/ is in the carry-phrase /iy aa iy/, and vowel /eh/ in /iy eh iy/. Note that as the speaking rate increases, the distinction between vowels /aa/ and /eh/ measured by the difference between their static formant values gradually diminishes. The same γ value of 0.9 is used in generating all points in the figure.

formant measurement data published in [6], where thousands of natural sound tokens were used to investigate the relationship between the degree of formant undershoot and speaking rate.[4] We re-organized and re-plotted the raw data from [6] in Fig. 6, in the same format as Fig. 5. While the measures of speaking rate differ between the measurement data and model prediction, the general trend for the greater degree of "static" sound confusion as speaking rate increases is clear from both the data and prediction.

3.4. Model prediction of formant trajectories for natural speech utterances

We have used the model presented in this paper to predict actual formant trajectories for natural speech utterances. Only the phone identities and their boundaries are input to the model for the prediction, and no use is made of speech acoustics (as is the case in formant tracking).

Given the phone sequence in any utterance, we first broke up the compound phones (affricates and diphthongs) into their con-

[4] We are grateful to Dr. M. Pitermann for providing us with the raw data of formant measurements published in [6] and for useful discussions.

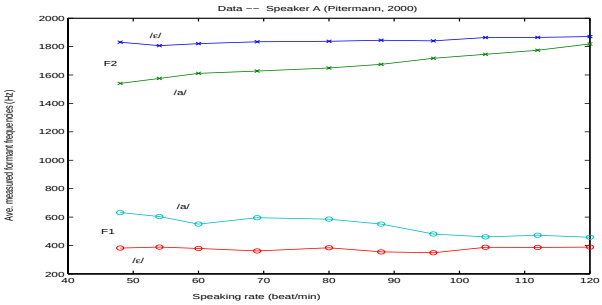

Figure 6: The formant measurement data from [6] are re-organized and plotted, showing similar trends to the model prediction in Figure 5.

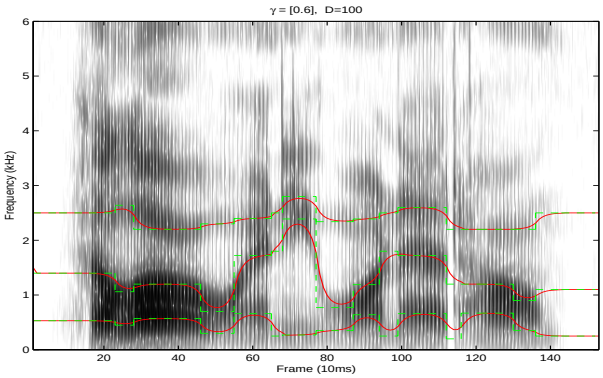

Figure 7: Fitting the F1/F2/F3 formant trajectories generated from the model to a typical fluent speech utterance (male) taken from TIMIT database. The utterance is "*hello, anyone at home?*"

stituents. Then we obtained the formant target values[5] based on limited context dependency by table lookup (see details in Chapter 13 of [11]). These target values, together with the phone boundaries provided (such as in TIMIT database), form the input to the FIR filter in the model and the output of the filter gives the predicted formant trajectories.

One example is provided in Fig. 7 for a TIMIT utterance. The step-wise dashed lines (F1/F2/F3) are the input to the filter, and the continuous lines (F1/F2/F3) are the output of the filter as the predicted formant trajectories. To facilitate assessment of the accuracy in the prediction, the input and output are superimposed on the spectrogram of the utterance, where the true formants are shown as the dark bands. For the majority of the frames, the filter's output either coincides or is close to the true formants, even though no acoustic information is used. Also, comparing the input and output of the filter, we observe only a rather mild degree of formant undershoot or reduction.[6]

[5]These target values are provided not only to vowels, but also to consonants for which the resonance frequency targets are used with weak or no acoustic manifestation.

[6]When we do the same kind of model prediction for formant trajectories for the casual speech data (such as those in the Switchboard database), a greater degree of reduction is observed.

4. Conclusions

We have presented a quantitative model for predicting formant or vocal tract resonance dynamics, and for the related reduction and "static" speech sound confusion phenomena. The prediction requires only the phone sequence and the individual phone boundaries, and requires no acoustic information. The development of the model is motivated by phonetic theories and experiments on sound reduction in free-style speech. We intend to use the model as one source of a-priori knowledge about the speech structure for automatic recognition of conversational speech. We have accumulated evidence that the strong reduction and "static" sound confusion in this mixed style of speech, ranging widely in the hyper-hypo speaking continuum, are responsible for many recognition errors by state-of-the-art automatic systems.

Our current research in this direction involves extension of the formant trajectory prediction discussed in this paper to the prediction of trajectories of cepstra and of other types of acoustic features through an additional nonlinear mapping function. This will enable automatic learning of the input and filter parameters (targets, boundaries, and stiffness parameter) directly from these reliable acoustic measurements, instead of from automatically tracked formants which may be prone to errors.

5. References

[1] B. Lindblom. "Spectrographic study of vowel reduction," J. Acoust. Soc. Am., Vol. 35, 1963, pp. 1773-1781.

[2] J. van Santen. "Contextual effects on vowel reduction," Speech Communications, Vol. 11, 1992, pp. 513-546.

[3] D. van Bergem. "Acoustic vowel reduction as a function of sentence accent, word stress and word class," Speech Communications, Vol. 12, 1993, pp. 1-12.

[4] S. Moon and B. Lindblom. "Interaction between duration, context, and speaking style in English stressed vowels," J. Acoust. Soc. Am., Vol. 96, 1994, pp. 40-55.

[5] L. Pols. "Psycho-acoustics and speech perception," in *Computational Models of Speech Pattern Processing*, (K. Ponting Ed.), Berlin: Springer, pp. 10-17.

[6] M. Pitermann. "Effect of speaking rate and contrastive stress on formant dynamics and vowel perception," J. Acoust. Soc. Am., Vol. 107, 2000, pp. 3425-3437.

[7] S. Hertz. "Streams, phones, and transitions: Towards a new phonological and phonetic model of formant timing," J. Phonetics, Vol. 19, 1991, pp. 91-109.

[8] J. Wouters and M. Macon. "Control of spectral dynamics in concatenative speech synthesis," IEEE Trans. Speech and Audio Proc., Vol. 9, 2001, pp. 30-38.

[9] L. Deng. "Computational models for speech production," in *Computational Models of Speech Pattern Processing*, (K. Ponting Ed.), Berlin: Springer, 1999, pp. 199-213.

[10] J. Ma and L. Deng. "Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model for vocal-tract-resonance dynamics," IEEE Trans. Speech and Audio Proc., Vol.11, 2003, pp. 590-602.

[11] L. Deng and D. O'Shaughnessy. *SPEECH PROCESSING — A Dynamic and Optimization-Oriented Approach,* Marcel Dekker: New York, NY, 2003.