MAGNITUDE-ONLY ESTIMATION OF HANDSET NONLINEARITY WITH APPLICATION TO SPEAKER RECOGNITION*

T.F. Quatieri, D.A. Reynolds, and G.C. O'Leary

Lincoln Laboratory, MIT, Lexington, MA, USA
Email: {tfq,dar,gco}@sst.ll.mit.edu

ABSTRACT

A method is described for estimating telephone handset nonlinearity by matching the *spectral magnitude* of the distorted signal to the output of a nonlinear channel model, driven by an undistorted reference. This "magnitude-only" representation allows the model to directly match unwanted speech formants that arise over nonlinear channels and that are a potential source of degradation in speaker and speech recognition algorithms. As such, the method is particularly suited to algorithms that use only spectral magnitude information. The distortion model consists of a memoryless polynomial nonlinearity sandwiched between two finite-length linear filters. Minimization of a mean-squared spectral magnitude error, with respect to model parameters, relies on iterative estimation via a gradient descent technique, using a Jacobian in the iterative correction term with gradients calculated by finite-element approximation. Initial work has demonstrated the algorithm's usefulness in speaker recognition over telephone channels by reducing mismatch between high- and low-quality handset conditions.

1 INTRODUCTION

A major source of performance loss in speaker recognition systems is telephone handset mismatch between training and testing data [3, 4]. Although linear compensation techniques improve recognition performance, such methods address only part of the problem, not accounting for the nonlinear distortion component [4]. Telephone handset nonlinearity often introduces spurious resonances that are not present in the original speech spectrum. An example showing such spurious resonances, which we shall refer to as "phantom formants," is given in Figure 1 where a comparison of all-pole spectra from a TIMIT waveform and its counterpart carbon-button microphone version from HTIMIT [5] are shown. Phantom formants, occurring at multiples, sums, and differences of original formants, as well as two other spectral distortions of bandwidth widening and spectral flattening seen in Figure 1, have been consistently observed not only in HTIMIT, but also in other databases with dual wideband/telephone recordings such as the wideband/narrowband King and the TIMIT/NTIMIT databases. For example, in a TIMIT/NTIMIT comparison, using a formant tracking algorithm, we have measured for male and female speakers, respectively, on the order of 10% and 56% phantom formants between F1 and F2, and 16% and 18% phantom formants between F2 and F3 [2].

In this paper, a nonlinear model is first hypothesized to account for the observed handset distortion. With dual waveform recordings before and after the handset, a method is then described for estimating the handset model parameters using spectral magnitude only, thus directly matching any phantom formants, bandwidth widening, and spectral flattening due to nonlinearity. The technique, therefore, aims at matching distortion in the feature domain used in speaker recognition and other speech processing tasks. This method of estimation provides a powerful alternative to time-domain-based matching techniques that do not directly match spectral magnitude and thus do not explicitly account for the presence of phantom formants due to nonlinear distortion. Because a goal of this work is to eliminate handset mismatch between training and testing data, we next extend our approach to modeling and estimation of a handset *mapper*, in contrast to a handset itself. Specifically, the method is applied to the mapping of high-quality (e.g. electret) to low-quality (e.g., carbon-button) handsets, and, to a lesser extent, low-quality to high-quality inversion. Using the mappings in conjunction with a handset classifier [5], we can improve consistency between training and testing datasets. These mappings have resulted in significant improvement in automatic speaker recognition using a Gaussian mixture-model-based speaker recognition system [6].

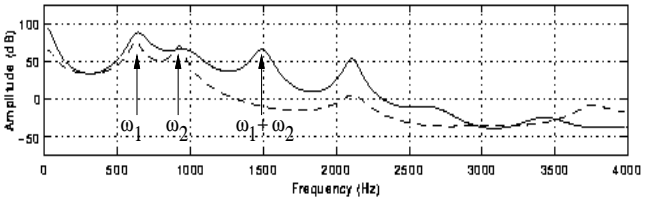

Figure 1: Illustration of phantom formants, comparing all-pole spectra from wideband TIMIT (dashed) and carbon-button HTIMIT (solid) recordings. Location of first phantom formant ($\omega_1+\omega_2$) is roughly equal to sum of locations of first two original formants (ω_1 and ω_2).

2 MODEL

At the core of our hypothesized telephone handset model is a memoryless polynomial nonlinearity whose selection is based on the observation that raising a resonant response to an integer power corresponds to adding new phantom formants at sums, differences, and multiplies of the original formants, and a broadening of resonant bandwidths[1]. This simple nonlinear operation is thus consistent with observed

*THIS WORK WAS SUPPORTED BY THE DEPARTMENT OF THE AIR FORCE. OPINIONS, INTERPRETATIONS, CONCLUSIONS, AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND ARE NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

[1] For periodic signals, these forms of spectral distortion can be approximately determined from the original formants of the underlying response. Therefore, although harmonics of the periodic signal add to one another in a complex fashion, the resulting spectral envelope can be predicted from the original resonances.

spectral distortion. The nonlinearity is sandwiched between a prefilter and a postfilter, both of which are assumed FIR. The primary purpose of the prefilter is to provide a scaling and dispersion[2], while the postfilter provides some additional spectral shaping.

In discrete time, let $x(n)$ denote the undistorted signal, to be referred to as the *reference signal*. The output of the nonlinear handset model is given by

$$y(n) = Q[g(n) * x(n)] * h(n) \tag{1}$$

where $g(n)$ is an Mth-order FIR prefilter, $h(n)$ is an Nth-order FIR postfilter, and Q is a Pth-order polynomial nonlinear operator, which for an input value z has an output

$$Q[z] = q_0 + q_1 z + q_2 z^2 + \ ... \ q_{P-1} z^P \tag{2}$$

Observe that we can consider our model (1) as a special case of the Volterra series representation of nonlinear systems [8] which has the advantage of being linear in the unknown parameters. For our problem, however, this series expansion significantly increases the number of model parameters, further complicates the internal workings of the model, and lacks convergence with "hard" constraints such as a saturating element that we will later add to our model.

To formulate the estimation problem, we define a vector of model parameters $\underline{a} = [\underline{g}, \underline{q}, \underline{h}]$, where $\underline{g} = [g(0), g(1), ...g(M-1)]$, $\underline{q} = [q_0, q_1, ...q_P]$, and $\underline{h} = [h(0), h(1), ...h(N-1)]$ so that the goal is to estimate the vector $\underline{a}$. A time-domain approach is to minimize an error criterion based on waveform matching, such as $\sum_n [s(n) - y(n;\underline{a})]^2$, where $s(n)$ is the *measurement signal* and where we have included the parameter vector $\underline{a}$ as an argument in $y(n)$. One technique for parameter estimation is through the Volterra series that yields a linear estimation problem. Because of the aforementioned problems of this series, as well as sensitivity of waveform matching to phase dispersion and delay (e.g., typical misalignment between model output and measurement), we have found this approach not to be feasible for our application. An alternative approach is to define the error in the frequency domain using the spectral magnitude.

3 SPECTRAL MAGNITUDE MATCHING

We begin by defining an error between the spectral magnitude of the measurement and nonlinearly distorted model output. Because a speech signal is nonstationary, the error function uses the spectral magnitude over multiple short frames and is given by

$$E(\underline{a}) = \sum_{k=0}^{K-1} \int_0^{\pi} [\,|S(\omega;k)| - |Y(\omega;k;\underline{a})|\,]^2 d\omega \tag{3}$$

where $S(\omega;k)$ and $Y(\omega;k;\underline{a})$ are the short-time Fourier transforms of $s(n)$ and $y(n;\underline{a})$, respectively, over an observation interval and where k refers to the frame index. An important advantage of the use of spectral magnitude is that the error is defined in the domain of interest for the speaker recognition system, providing a "direct" spectral match to phantom formants, bandwidth widening, and spectral flattening. Furthermore, there is robustness to dispersion and delay. Our goal is to minimize $E(\underline{a})$ with respect to the unknown model coefficients $\underline{a}$. This is a highly nonlinear problem with no obvious closed-form solution. An approach to parameter estimation is solution by iteration, one in particular being the generalized Newton method [1].

To formulate an iterative solution, we first discretize the continuous Fourier transform, i.e.,

$$E(\underline{a}) = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} [\,|S(\omega_l;k)| - |Y(\omega_l;k;\underline{a})|\,]^2 \tag{4}$$

where L is the discrete Fourier transform length. We then define the *residual vector* $\underline{f}(\underline{a})$ by

$$\underline{f}(\underline{a}) = [\underline{f}^0(\underline{a}), \underline{f}^1(\underline{a}), ...\underline{f}^K(\underline{a})] \tag{5}$$

where $\underline{f}^k(\underline{a}) = [\,|S(\omega_l;k)| - |Y(\omega_l;k;\underline{a})|\,] \quad l = 0, 1, ...L-1$. The error in (4) is a scalar function that can be interpreted as the sum of squared residuals over all frames, i.e.,

$$E(\underline{a}) = \underline{f}^T \underline{f}(\underline{a}) \tag{6}$$

where T denotes matrix transpose. The gradient of $E(\underline{a})$ is given by $\nabla E = 2\mathbf{J}^T \underline{f}$ where $\mathbf{J}$ is the Jacobian matrix of first derivatives of the residual equations, i.e., the elements of $\mathbf{J}$ are given by

$$J_{ij} = \frac{\partial f_i}{\partial a_j} \tag{7}$$

where f_i is the ith element of $\underline{f}(\underline{a})$. The generalized[3] Newton iteration [1], motivated by the first term of a Taylor series expansion of $\underline{f}(\underline{a})$, is given by adding to an approximation of $\underline{a}$ at each iteration a correction term, i.e.,

$$\underline{a}_{m+1} = \underline{a}_m + \mu \Delta \underline{a}_m \tag{8}$$

where $\Delta \underline{a}_m = -(\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T \underline{f}(\underline{a}_m)$ with $\underline{f}$ evaluated at the current iterate $\underline{a}_m$, and where the factor μ scales the correction term to control convergence.

One of the most important computational elements of the iterative solution is the calculation of the Jacobian $\mathbf{J}$, and thus the partial derivatives in (7). For the residual definition of (5), there is not a closed-form expression for the gradients. Nevertheless, an approximate gradient can be calculated by finite-element approximations. It follows that one algorithm is given by the steps (Figure 2)

(1) Initiate with a parameter vector $\underline{a}_0$, representing no nonlinear distortion (straight line) and impulses for the linearities, i.e., an identity. Compute the short-time Fourier transform magnitude $|S(\omega_l;k)|$ of the measurement $s(n)$.

(2) Compute the short-time Fourier transform magnitude $|Y(\omega_l;k;\underline{a}_m|$ of the synthesized model output based on the current $\underline{a}_m$, i.e., of the signal

$$y_m(n;\underline{a}_m) = Q_m[g_m(n) * x(n)] * h_m(n)$$

and form the residual vector $\underline{f}(\underline{a}_m)$.

(3) Compute an estimate of the partial derivative of the elements of the residual vector $\underline{f}(\underline{a})$ evaluated at the current value of $\underline{a} = \underline{a}_m$ with respect to each element

[2]The dispersion provides memory to the nonlinearity. This is in lieu of an actual handset model which might introduce a more complex process such as hysteresis.

[3]When the number of equations equals the number of unknowns, the generalized Newton method reduces to using a correction $\Delta u = -\mathbf{J}^{-1}\underline{f}$ which is the standard Newton method.

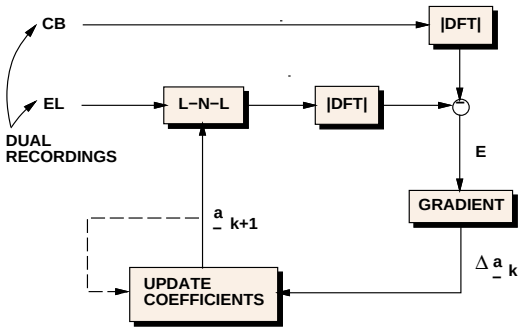

Figure 2: Iterative magnitude-only estimation

of $\underline{a}$. This requires recalculating $y(n;\underline{a}_m)$ for each perturbed component of $\underline{a}_m$ and computing its short-time Fourier transform magnitude. Using a first backward difference, the partial derivative estimate of (7) for each element of $\underline{a}$ is given by

$$\frac{\partial f_i}{\partial a_j} \approx \frac{|Y(\omega_i; a_0, a_1, ...a_j + \epsilon...)| - |Y(\omega_i; a_0, a_1, ...a_j...)|}{\epsilon}$$

where ϵ is a small perturbation.

(4) Based on step (3), form the Jacobian, compute the correction term, and update the parameter vector with

$$\underline{a}_{m+1} = \underline{a}_m + \mu\Delta\underline{a}_m$$

In general, there will not be a unique solution in fitting the representation (1) to a measurement, even if the measurement fits the model exactly. One problem is the ambiguity of scale. For example, the coefficients of the polynomial nonlinearity can always account for an input scaling by c, i.e., $y(n) = \sum_{k=0}^{P} q_k[cx(n)]^k = \sum_{k=0}^{P} \hat{q}_k x^k(n)$ where $\hat{q}_k = q_k c^k$. To remove this particular ambiguity, we invoke constrained optimization. Under the assumption that all handsets eventually saturate, for the purpose of removing solution ambiguity, as well as controlling the size of the residual vector and thus yielding more consistent results across an utterance, limiting is introduced to the output of the nonlinear operator by $Q[z] = sgn[z]$ for an operator input $|z| > 1$. In addition, three boundary constraints[4] are imposed given by $y(0) = 0$, $y(1) = 1$, and $y(2) = -1$, thus reducing the number of free variables by three.

In Figure 2 we have written the measurement as "CB", denoting a carbon-button handset output, and the undistorted reference as "EL" denoting an electret handset output. This was done because our ultimate goal is not necessarily a handset model but rather a handset *mapper*[5] for the purpose of reducing handset mismatch between high- and low-quality handsets. In particular, our reference signal is the highest-quality electret (EL1) and the measurement the lowest-quality carbon (CB3) from HTIMIT [5]. We will refer to this transformation as a "forward mapper", having occasion to also invoke an "inverse mapper" from the low-quality carbon to high-quality electret.

[4]These boundary constraints, as well as the limiting operation, are conjectures of the underlying handset mechanism and are currently being refined.

[5]Because we assume distortion introduced by a high-quality electret handset is linear, our model of a handset, i.e., a nonlinearity sandwiched bewteen two linear filters, is also used for the handset mapper.

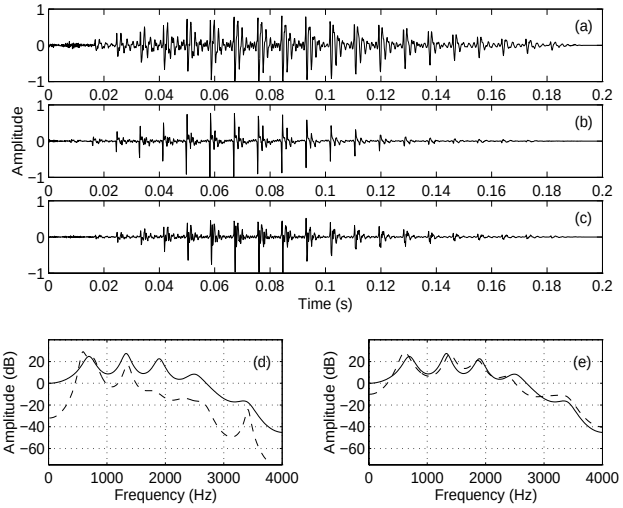

Figure 3: Example of electret-to-carbon button mapping: (a) electret waveform output; (b) carbon-button waveform output; (c) electret-to-carbon mapped waveform; (d) comparison of all-pole spectra from (a) (dashed) and (b) (solid); (e) comparison of all-pole spectra from (b) (solid) and (c) (dashed) .

4 EXAMPLE HANDSET MAPPER

An example of mapping a EL1 handset to a CB3 handset output is shown in Figure 3. The "training" data consists of 1.5 seconds of a male speaker from HTIMIT, analyzed with a 20ms Hamming window at a 5ms frame interval. The prefilter and postfilter are both of length 5 and the polynomial nonlinearity of order 7. Figures 3a and 3b show particular time slices of the original electret and carbon-button outputs, while Figure 3d shows the disparity in their all-pole spectra, manifested in phantom formants, bandwidth widening, and spectral flattening. Figure 3c gives the waveform resulting from applying the estimated mapper to the same electret output but in a "test" region of the utterance outside of the 1.5s training interval, while Figure 3e compares the carbon-button all-pole spectrum to that of mapping the electret to carbon-button output, illustrating a close spectral match.

The characteristics of the mapper estimate are shown in Figure 4. Convergence is achieved after about 500 iterations, as seen in the training error. The postfilter takes on a bandpass characteristic, while the prefilter (not shown) is nearly flat. The nonlinearity is convex which is consistent with the observation (compare Figures 3a and 3b) that the carbon-button handset tends to "squelch" low-level values relative to high-level values. An important parameter in determining the mapper characteristic is the energy of the input signal. The mapper is designed using a particular input energy level; with a nonlinear operator, changing this level would significantly alter the character of the output. Therefore, test signals are normalized to the energy level of the training data; this single normalization, however, may over- or under-distort the data and thus further consideration of input level is needed.

We can determine the performance of the mapper by computing a smooth spectral distance, of the form in (4), between an original measurement (carbon) and its corresponding reference (electret) before and after being mapped. Figure 4d shows this measure for the mapper of Figures 4a and 4b applied to 20 seconds of a male utterance, the first 1.5s being the original training data. We see that the spectral

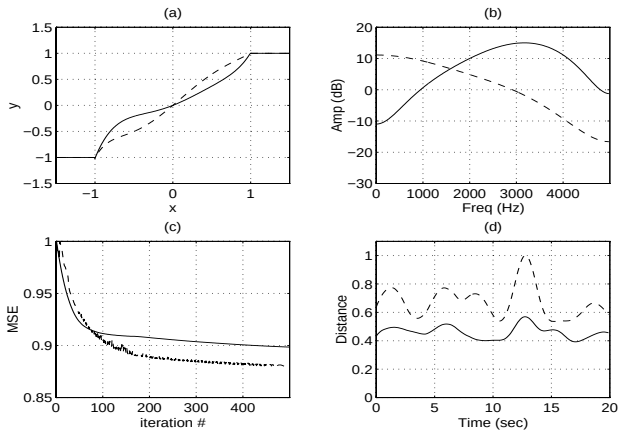

Figure 4: Characteristics of forward and inverse handset mappings: (a) nonlinearities, (b) postfilters, and (c) training error: forward mapper (solid)/inverse mapper (dashed); (d) testing error for forward mapper: no mapping (dashed)/mapping (solid).

distance reduction is preserved across the entire utterance. Similar reduction in spectral distance is seen when applying the mapper to other carbon-button test utterances within HTIMIT.

Heretofore, we have described the forward EL1-to-CB3 mapper. We can also design an inverse CB3-to-EL1 mapper simply by interchanging the reference and measurement waveforms. In this design, all specifications are the same except that the nonlinearity is of order 9 to account for a longer polynomial expansion observed in the inverse to the measured forward mapper. The inverse design is superimposed on the forward design in Figures 4a-4c. One notable observation is that the inverse nonlinearity is twisting in the opposite (concave) direction to that of the (convex) forward mapper, consistent with undoing the squelching imparted by the carbon-button handset. Although the inverse is sometimes able to remove the spectral distortion of phantom formants, bandwidth widening, and spectral flattening, we have found on the average the match to be inferior to that achieved by the forward design.

5 SPEAKER RECOGNITION EXPERIMENTS

One goal of handset (mapper) estimation is to eliminate handset mismatch between training and testing data to improve speaker recognition over telephone channels. The strategy is to assume two handset classes: high-quality (electret EL1) and low-quality (carbon-button CB3) from HTIMIT. We then design a forward EL1-to-CB3 mapper and an inverse CB3-to-EL1 mapper and apply the mappers according to handset detection [5] on training and testing data from another database (e.g., switchboard). Currently, we map only test data under a mismatch condition. Results are shown in Figure (5) using the 1996 NIST Speaker Recognition switchboard corpus [3] for different (NT) and same (TR) phone number training and testing cases. Results of using only the forward mapper are shown, illustrating significant improvement under the different phone number condition, and almost no change under the same phone number condition. With the additional use of the inverse mapper, further improvement was obtained for the different phone number condition, while some degradation in performance was found for the same condition. Similar performance trends were seen also on the 1997 NIST corpus [3] for the subset on which the mapper was applied. Unlike on the 1996 NIST corpus, however, this performance gain was averaged out on the entire database due to the far smaller number of mismatch cases.

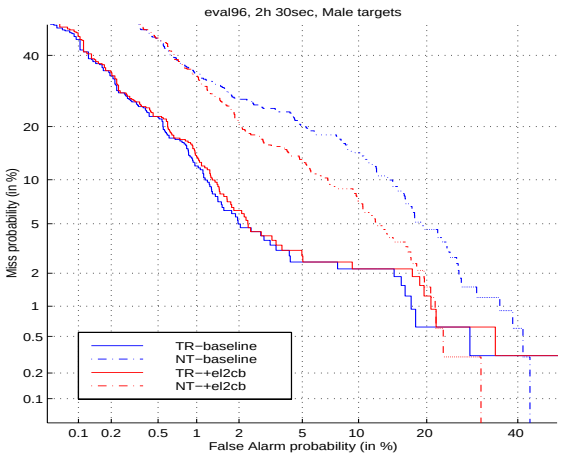

Figure 5: DET curves for 1996 NIST evaluation.

6 FUTURE

One current goal is to improve spectral matching with the forward and inverse handset mappers. The model is being refined by generalizing to nonpolynomial nonlinearities with more physically realizable constraints, while estimation is being investigated with different initial conditions and improved gradient estimates. Handset classes are being expanded to include good and bad handsets for each electret and carbon-button class. Finally, in using the mappers, we are improving input energy normalization, applying mappers to training (as well as testing) data, and exploring alternative strategies in merging with handset detection and H-normalization [7].

REFERENCES

[1] R. Fletcher, "Generalized inverses for nonlinear equations and optimization", in *Numerical methods for nonlinear algebraic equations*, Edited by P. Rabinowitz, Gordon and Breach Science Publishers, New York, NY, 1970.

[2] C.R. Jankowski, T.F. Quatieri, and D.A. Reynolds, "Measuring fine structure in speech: application to speaker identification", *ICASSP*, Detroit MI, May 1995.

[3] NIST, "NIST speaker recognition workshop notebook", NIST administered speaker recognition evaluation on the Switchboard corpus, March 1996 and June 1997.

[4] D.A. Reynolds, M.A. Zissman, T.F. Quatieri, G.C. O'Leary, and B.A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance", *ICASSP*, Detroit, MI, May, 1995.

[5] D.A. Reynolds, "HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects", *ICASSP*, Munich, Germany, April, 1997.

[6] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models" *Speech Communication*, Vol. 17, pp. 91-108, August 1995.

[7] D.A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification", *Proc. ESCA Eurospeech '97*, Rhodes, Greece, Sept., 1997.

[8] M. Schetzen, *The Volterra and Wiener theories of nonlinear systems*, John Wiley, New York, 1980.