A DATABASE OF VOCAL TRACT RESONANCE TRAJECTORIES FOR RESEARCH IN SPEECH PROCESSING

Li Deng[1], *Xiaodong Cui*[2], *Robert Pruvenok*[3], *Jonathan Huang*[4], *Safiyy Momen*[5],
Yanyi Chen[6] *and Abeer Alwan*[2]

Microsoft Research[1], Redmond, WA, 98052
University of California, Los Angeles[2], Georgia Institute of Technology[3],
Carnegie Mellon University[4], Princeton University[5] and Cornell University[6]

ABSTRACT

While vocal tract resonances (VTRs, or formants that are defined as such resonances) are known to play a critical role in human speech perception and in computer speech processing, there has been a lack of standard databases needed for the quantitative evaluation of automatic VTR extraction techniques. We report in this paper on our recent effort to create a publicly available database of the first three VTR frequency trajectories. The database contains a representative subset of the TIMIT corpus with respect to speaker, gender, dialect and phonetic context, with a total of 538 sentences. A Matlab-based labeling tool is developed, with high-resolution wideband spectrograms displayed to assist in visual identification of VTR frequency values which are then recorded via mouse clicks and local spline interpolation. Special attention is paid to VTR values during consonant-to-vowel (CV) and vowel-to-consonant (VC) transitions, and to speech segments with vocal tract anti-resonances. Using this database, we quantitatively assess two common automatic VTR tracking techniques in terms of their average tracking errors analyzed within each of the six major broad phonetic classes as well as during CV and VC transitions. The potential use of the VTR database for research in several areas of speech processing is discussed.

1. INTRODUCTION

Acoustic resonances in the human vocal tract during speech production are perhaps the best known and the most commonly used parameters in characterizing the perception of speech sounds, especially for vowels. In this case, the resonances are also called "formant frequencies", which are spectral prominences that can be computed directly from speech waveforms. However, when anti-resonances are present, as in most consonantal sounds, the underlying resonance frequencies are often obscured. However, it is the resonance frequencies, instead of the frequencies at which spectral prominences occur, that determine the temporal trends of CV and VC formant transitions (a classic concept known as formant "loci" [1] and have perceptual relevance.)

Due to the importance of the vocal tract resonance (VTR), numerous automatic VTR or formant trajectory estimation methods have been developed (e.g., [5, 8, 6, 3, 9, 10]) over past few decades. Results of many of these methods have been applied to speech processing applications (e.g., [6, 2, 9]). Despite this, there has been a conspicuous lack of standard databases that are needed for quantitative evaluation of automatic VTR estimation methods. As with automatic speech recognition, standard databases of VTRs as a "ground truth" reference are critical for objectively evaluating different estimation algorithms and for improving the algorithms' qualities. However, VTR databases are difficult to construct due to extensive human expertise required for accurately identifying resonances based primarily on spectrographic analysis. The difficulty is especially true for many consonantal sounds where the VTRs are often not directly visible from the spectrograms, as well as for some vowels and semivowels (e.g., /y/, /iy/, /u/, /w/, /r/) where formants can be very close to one another.[1]

In this work, we attempt to overcome these difficulties by carefully applying human expertise, in combination with spectral analysis with extrapolation, to visually identify and record the VTR frequencies (F1, F2, and F3) for all classes of speech sounds in a subset of the TIMIT database. During the preparation of our "manually" labeled VTR database, we especially focus on VTR values during consonant-to-vowel (CV) and vowel-to-consonant (VC) transitions, and on speech segments involving vocal tract anti-resonances. Where the spectral prominences (i.e., "dark" or high-energy bands in the spectrogram displays) do not coincide with predicted resonances for many consonantal segments, we exploit prior knowledge of nominal VTR values [2], and use the visible VTR transitions into and out of adjacent vocalic segments to perform judicious extrapolation or interpolation. We also examine the overall spectral properties across the entire sentence, correlating the VTR values for the same phonetic units in the sentence and adjusting them when appropriate after taking into account contextual influences. Further, we exploit known effects of anti-resonances (or zero frequencies) in splitting the VTRs of nasalized vowels for example.

The paper is organized as follows. In Section 2, we describe the selection of speech data for the VTR database preparation. The VTR trajectory labeling process is presented in Section 3, including the labeling tool development and human knowledge sources used. Cross-labeler variation results are shown in Section 4. Results of a preliminary objective assessment of two automatic VTR tracking algorithms using the labeled VTR database are provided in Section 5.

2. DATA SELECTION

The VTR database is composed of 538 utterances selected as a representative subset of the well-known and widely-used TIMIT cor-

The work is part of a project carried out during the 2005 NSF-RIPS summer program at the IPAM at UCLA in collaboration with Microsoft Research, Redmond, WA.

[1] Not surprisingly, many existing formant or VTR tracking algorithms tend to make errors for these speech sounds, including transitions into and out of these sounds.

pus. TIMIT contains a total of 6300 utterances contributed by 630 speakers from 8 major dialect regions of the United States. Each speaker speaks 10 utterances. The prompts for the 6300 utterances consist of 2 dialect "shibboleth" sentences (SA), 450 phonetically-compact sentences (SX), and 1890 phonetically-diverse sentences (SI). TIMIT is divided into a training set and a test set. The former contains 4620 utterances, and the latter contains 1344 utterances where 192 out of 1344 utterances are specially chosen to form a core test set.

In preparing our VTR database, we selected a total of 538 utterances (SX and SI sentences only) and labeled F1, F2, and F3 trajectories for each 10-msec frame. The selected utterances cover all 192 utterances from the core test set and 346 utterances from the training set. There are 24 speakers in the 192-utterance test subset with 5 SX and 3 SI sentences for each speaker, and 173 speakers in the 346-utterance training subset with each speaker contributing 1 SX and 1 SI sentences. In this way, the selected 538 utterances well represent a balanced selection of speaker, dialect, gender and phoneme. These utterances also contain rich phonetic contexts and thus are a good collection of acoustic-phonetic phenomena that exhibit interesting VTR variations.

3. VTR TRAJECTORY LABELING

To facilitate the process of VTR trajectory labeling in the database preparation, we first obtained initial trajectory estimates based on an automatic VTR tracking algorithm as described in [3].[2] Based on these initial estimates, extensive manual correction is performed to provide accurate VTR labeling.

A Matlab GUI tool has been developed to enable VTR correction and manual labeling. A screen shot of this tool is shown in Fig.1. The tool shows the waveform and wideband spectrogram of a speech signal together with its word- and phone-level transcriptions. The phone boundaries are marked also in the spectrogram to facilitate labeling. The spectrogram can be zoomed in and out for detailed or coarse spectrographic information. A contrast bar is also available to tune the intensity contrast of the spectrogram to help make decisions on the VTR values. To correct trajectories, a labeler simply needs to click on the desired points and a local spine interpolation is implemented in the tool that automatically smoothes out the modified trajectory to the visual satisfaction of the labeler.

Figs. 2 and 3 are two example sentences to illustrate the correction and labeling process and results. Blue dashed lines (F1, F2, and F3, respectively) are the initial, automatic VTR trajectory estimates described in [3] marked as "MSR" in legend. Manually corrected versions of them are shown as red solid lines (with 20 to 30 point corrections per sentence by manual mouse clicks followed by automatic local spline smoothing). For comparison purposes, we also show the automatic F1/F2/F3 tracking results, as green solid lines, from the popular open source tool WaveSurfer, which uses the same algorithm for VTR/formant tracking as ESPS/xwaves introduced in [8].[3] For most of the vocalic portions in the utterances where the "dark/high energy" bands in the wideband spectrograms are clearly identifiable, both automatic trackers give rather accurate results. Exceptions are occasional errors in 1) F3 for /r/ or /er/ when F3 is low and close to F2; 2) F2 for /uw/ or /w/ when F2 is low and close to F1; and 3) F2 and/or F3 for /y/ and /iy/ when they are close to each other. These errors occur mostly during relatively rapid formant transitions.

More difficult situations arise for the frames where there is a lack of spectral prominences or when spectral prominences do not coincide with predicted resonances for consonantal segments. In this case, nominal consonant-specific values of VTRs (Chapter 10 in [2]) are provided as crude references for male speakers as a guideline. These values are increased by approximately 20% for female speakers. To determine the final labels, we also make use of the visible VTR transitions into and out of adjacent vocalic segments to extrapolate to the consonantal VTR "loci". "Velar pinch" patterns (F2 and F3 coming together) are identified and used for labeling F2 and F3 values related to velar consonants (/k/, /g/, /ng/) in the front-vowel context. The overall spectral properties across the entire sentence with a fixed speaker are exploited to equalize the VTR values for the same phonetic units in the sentence; adjustments may be made based on analysis of contextual effects. For nasal consonants or nasalized vowels, the "pole-zero-pole" pattern in their spectra is analyzed if such a pattern exists. In this case, the final VTR values may be determined by the frequencies where spectral valleys instead of spectral peaks occur, especially if such values are consistent with the nominal values based on prior knowledge.[4] Examples of some of the above analyses and decisions during the VTR correction/labeling process have been included in Figs. 2 and 3.

4. CROSS-LABELER VARIATIONS

The VTR labeling effort of the database is distributed among several labelers. A problem that naturally arises is the possible inconsistency among them. In difficult cases where the VTRs in a spectrogram are not obvious or are ambiguous, the manually labeled VTR values from different labelers often differ from each other.

In order to assess the degree of such inconsistency, independent pair-wise cross-labeling is performed for a subset (80 in total) of the 538 manually labeled utterances. Five pairs of labelers are made, with each pair performing independent cross-labeling of common 16 utterances (including two utterances from separate four male and four female speakers). The absolute difference between the two different labelers separated into six broad phonetic classes and into F1, F2, and F3, averaged over all frames for each class, is listed in Table 1. The magnitude of the variation is somewhat higher than was expected, warranting additional checking for possible labeling errors.

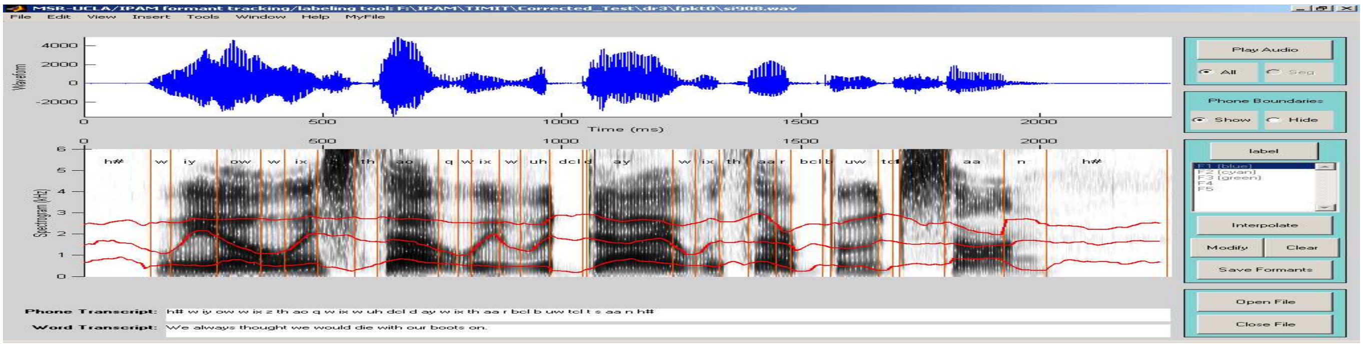
Classes	Absolute Diff per frame (Hz)		
	F1	F2	F3
vowels	55	69	84
semivowels	68	80	103
nasal	75	112	106
fricatives	91	113	125
affricatives	89	118	135
stops	91	110	116

Table 1. Averaged pair-wise cross-labeling absolute difference per frame for F1, F2, and F3 and for each of the six broad phonetic classes (in Hz).

[2] This algorithm is based on a version of the structured speech model consisting of continuous-valued hidden dynamics and a piecewise-linearized prediction function from resonance frequencies and bandwidths to LPC cepstra.

[3] Software download site: http://www.speech.kth.se/wavesurfer.

[4] This is an example where the underlying resonances in the vocal tract may not correlate with the spectral peaks in the speech signal.



MSR-UCLA/IPAM formant tracking/labeling tool: F:\IPAM\TIMIT\Corrected_Test\dr3\fpkt0\si908.wav

Phone Transcript: h# w iy ow w ix z th ao q w ix w uh dcl d ay w ix th aa r bcl b uw tcl t s aa n h#

Word Transcript: We always thought we would die with our boots on.

Fig. 1. Matlab-based GUI labeling tool for manual VTR trajectory correction and labeling.

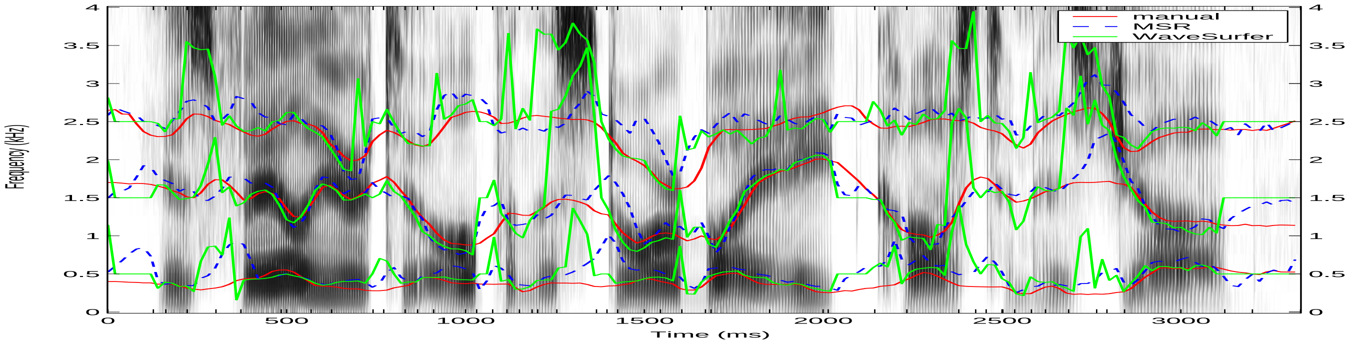


Fig. 2. Spectrogram of a TIMIT utterance, *His failure to open the store by eight cost him his job*, superimposed with the F1/F2/F3 trajectories corrected/labeled by hand (red solid lines; "Manual" in legend) and from two automatic VTR trackers. One tracker is described in [3] (blue dashed lines; "MSR" in legend), and the other is described in [8] (green solid lines; "Wavesurfer" in legend).

5. COMPUTING ERRORS IN AUTOMATIC TRACKING ALGORITHMS

The 538 manually labeled utterances with F1, F2, and F3 trajectories in the VTR database described above can serve as a basis for quantitatively computing the errors produced by existing VTR tracking algorithms. The labeled trajectories are used as the references for determining the errors. We selected the "MSR" algorithm in [3] and the "WaveSurfer" algorithm (footnote 3) for analysis; the resulting errors are listed in Table 2. The errors are defined to be the absolute VTR difference between the reference and the estimated values averaged over all frames for a particular broad phonetic class. We have used six such broad classes defined in the TIMIT database, with all segment boundaries made available for frame averaging. Compared with the inter-labeler variation shown in Table 1, the difference in the errors made by the two algorithms for the sonorant speech classes (vowels, semivowels, and nasals) appears to be relatively minor. Greater differences occur for the obstruent speech classes (fricatives, affricatives, and stops).

We also examined the errors of the two algorithms when limiting the error-counting regions to only the CV and VC transitions. The "transition regions" are fixed to be 6 frames, with 3 frames to the left and 3 frames to the right of CV or VC boundaries defined in the TIMIT database. The detailed results are listed in Table 3.

As a control experiment, we have corrected and labeled a small subset (about 10%) of VTR trajectories which are initialized from the Wavesurfer's outputs instead of MSR algorithm's outputs. The differences shown above are reduced, mainly for vowels and for semivowels. Definite conclusions can not be drawn until after a larger set of the data is labeled.

Classes	MSR			WaveSurfer		
	F1	F2	F3	F1	F2	F3
vowels	64	105	125	70	94	154
semivowels	83	122	154	89	126	222
nasal	67	120	112	96	229	239
fricatives	129	108	131	209	263	439
affricatives	141	129	149	292	407	390
stops	130	113	119	168	210	286

Table 2. VTR tracking errors (in Hz) measured by averaging absolute VTR differences between the reference and estimated values over all frames for a particular broad phonetic class. Both the algorithms in [3] (MSR) and for the WaveSurfer algorithm are used. Results are listed for F1, F2, and F3, and for each of the six phonetic classes separately.

6. SUMMARY AND DISCUSSION

In this paper, we report on the development of a soon-to-be publicly available database in which the VTR, or formant, trajectories are labeled. We plan to have the database, after further careful verification and correction (which is currently on-going at MSR), be posted online on our webpages by the time of the ICASSP 2006 conference. F1, F2, and F3 trajectories, at every 10-msec frame, of a represen-

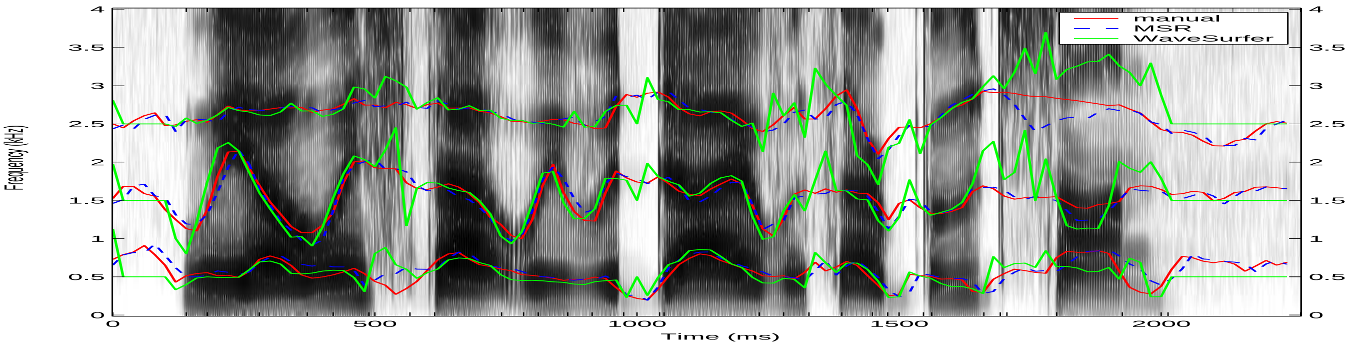


Fig. 3. Same as Figure 2, except with a different TIMIT utterance: “We always thought we would die with our boots on”.

Classes	MSR			WaveSurfer		
	F1	F2	F3	F1	F2	F3
CV transitions	106	101	119	156	192	273
VC transitions	48	92	120	59	88	157

Table 3. Same as for Table 2 except for the focus on temporal regions of consonant to vowel (CV) transitions and vowel to consonant (VC) transitions instead of on broad phone classes.

tative subset (538 sentences) of the TIMIT corpus with respect to speaker, gender, dialect and phonetic context are labeled. The key part of the database creation process is the development and use of a software labeling tool based on high-resolution spectral and temporal displays. This enables visual identification of, for example, VTR or formant frequency values in the spectrograms, typically in vocalic speech segments. When ambiguity arises or when no clear “dark/high-energy bands” are visible in the spectrograms, the underlying resonance frequencies have to be inferred using human expertise to balance a number of knowledge sources. In this paper, we discussed major knowledge sources that guided our labeling process while creating the VTR database, including the use of consonantal VTR targets (or “loci”) and their gender-based adjustment, the use of visible CV and VC transitions for extrapolation, consistency of the within-utterance VTR targets and VTR values for the same phone, the adjustment of targets and VTR values based on contextual influences and on possible target undershooting, the use of the distinct “velar pinch” pattern, and the effects of nasalization on pole splitting.

We report in this paper an exploratory use of the database on quantitative assessment of two common automatic VTR tracking techniques measured by their averaged tracking errors analyzed within each of the six broad phonetic classes as well as those during CV and VC transitions. We also report a small-scale experiment where interlabeler variation (as an indicator of the quality of the database) is examined.

A clear benefit of the established VTR database is in quantitative and rigorous evaluation of existing and new VTR or formant tracking algorithms and in fostering the future high-quality algorithm development. In addition to this use in speech analysis research, other potential uses of the VTR database can be foreseen in different areas of speech processing as well. In particular, the importance of the resonances or formants in speech perception makes it desirable to incorporate them into speech recognition systems either as new features in the discriminative framework [7], or as part of model structure in the generative framework [4]. With a standard database made available, the feature extraction techniques and recognition models can be better designed or beneficially initialized in automatic training. In formant-based speech synthesis, the standard database for VTR trajectories in speech will also help improve the formant-generation component of the system, especially for generating natural-style utterances with varying degrees of phonetic reduction.

Acknowledgment: We thank Prof. Mark Green of IPAM at UCLA for the organization of the IPAM program which makes this collaborative work with MSR possible. The initial work of Leo Lee in designing the interface component of the labeling tool at MSR and useful discussions with David Talkin related to the work in [8] are gratefully acknowledged. We also thank Alex Acero for suggesting cross-labeling experiments and for encouragement and support of this work.

7. REFERENCES

[1] P. C. Delattre, A. M. Liberman, and F. S. Cooper. “Acoustic loci and transitional cues for consonants,” *J. Acoust. Soc. Am.* Vol. 27, 1955, pp. 769-773.

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

[3] L. Deng, L. J. Lee, H. Attias, and A. Acero. “A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances,” *Proc. ICASSP*, Vol. 1, 2004, pp. 557-560.

[4] L. Deng, D. Yu, and A. Acero. “A bi-directional target-filtering model of speech coarticulation and reduction: Two-stage implementation for phonetic recognition,” *IEEE Trans. Speech & Audio Proc.*, Vol. 14, 2006, pp. 256-265.

[5] S. McCandless. “An algorithm for automatic formant extraction using linear prediction spectra,” *IEEE Trans. Acoust. Speech & Sig. Proc.*, Vol. 22, 1974, pp. 135-141.

[6] L. Welling and H. Ney. “Formant tracking for speech recognition,” *IEEE Trans. Speech & Audio Proc.*, Vol. 6, 1998, pp. 36-48.

[7] F. Pereira. “Linear models for structure prediction”, *Proc. Interspeech*, Lisbon, Sept 2005, pp. 717-720.

[8] D. Talkin. “Speech formant trajectory estimation using dynamic programming with modulated transition costs” *J. Acoust. Soc. Am.*, S1, 1987, pp. S55.

[9] B. Strope and A. Alwan, “Robust word recognition using threaded spectral peaks,” *Proc. ICASSP*, 2004, Vol.II, pp. 625-629.

[10] Y. Zheng and M. Hasegawa-Johnson. “Formant tracking by mixture state particle filter,” *Proc. ICASSP*, 2004, Vol.1, pp. 565-568.