date of current version 17 February 2025. The review of this article was arranged by Associate Editor Nobutaka Ito.

Digital Object Identifier 10.1109/OJSP.2025.3530876

Formant Tracking by Combining Deep Neural Network and Linear Prediction

SUDARSANA REDDY KADIRI [1] (Member, IEEE), KEVIN HUANG [1] (Graduate Student Member, IEEE), CHRISTINA HAGEDORN [2], DANI BYRD [1], PAAVO ALKU [3] (Fellow, IEEE), AND SHRIKANTH NARAYANAN [1] (Fellow, IEEE)

[1]University of Southern California, Los Angeles, CA 90007 USA
[2]College of Staten Island, The Graduate Center, City University of New York, New York, NY 10017 USA
[3]Aalto University, 02150 Espoo, Finland

CORRESPONDING AUTHORS: SUDARSANA REDDY KADIRI; PAAVO ALKU (e-mails: skadiri@usc.edu; paavo.alku@aalto.fi).

This work was supported in part by National Science Foundation under Grant 2311676 and in part by the Academy of Finland under Grant 330139.

ABSTRACT Formant tracking is an area of speech science that has recently undergone a technology shift from classical model-driven signal processing methods to modern data-driven deep learning methods. In this study, these two domains are combined in formant tracking by refining the formants estimated by a data-driven deep neural network (DNN) with formant estimates given by a model-driven linear prediction (LP) method. In the refinement process, the three lowest formants, initially estimated by the DNN-based method, are frame-wise replaced with local spectral peaks identified by the LP method. The LP-based refinement stage can be seamlessly integrated into the DNN without any training. As an LP method, the study advocates the use of quasiclosed phase forward-backward (QCP-FB) analysis. Three spectral representations are compared as DNN inputs: mel-frequency cepstral coefficients (MFCCs), the spectrogram, and the complex spectrogram. Formant tracking performance was evaluated by comparing the proposed refined DNN tracker with seven reference trackers, which included both signal processing and deep learning based methods. As evaluation data, ground truth formants of the Vocal Tract Resonance (VTR) corpus were used. The results demonstrate that the refined DNN trackers outperformed all conventional trackers. The best results were obtained by using the MFCC input for the DNN. The proposed MFCC refinement (MFCC-$DNN_{QCP\text{-}FB}$) reduced estimation errors by 0.8 Hz, 12.9 Hz, and 11.7 Hz for the first (F1), second (F2), and third (F3) formants, respectively, compared to the Deep Formants refinement ($DeepF_{QCP\text{-}FB}$). When compared to the model-driven KARMA tracking method, the proposed refinement reduced estimation errors by 2.3 Hz, 55.5 Hz, and 143.4 Hz for F1, F2, and F3, respectively. A detailed evaluation across various phonetic categories and gender groups showed that the proposed hybrid refinement approach improves formanttracking performance across most test conditions.

INDEX TERMS Formant tracking, linear prediction, MFCCs, spectrogram, machine learning, deep learning.