Classification of formant estimation methods in transformed auditory feedback experiments using convolutional neural networks

Fumiaki Taguchi^{1,*}, Sadao Hiroya², Yasufumi Uezu² and Takemi Mochida²

¹Graduate School of Design, Kyushu University, Shiobaru 4-9-1, Minami-ku, Fukuoka, 815-8540 Japan

²NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, 3-1 Morinosato Wakamiya Atsugi, 243-0198 Japan

(Received 6 April 2020, Accepted for publication 27 April 2020)

Keywords: Auditory feedback, Speech production, PEAR, LPC, CNN

PACS number: 43.70.-h, 43.70.Jt, 43.70.Mn, 43.72.Ar [doi:10.1250/ast.41.800]

1. Introduction

Speech communication is essential in daily life. The auditory feedback mechanism, by which we monitor our own speech in real time and correct errors in it, plays an important role in our ability to stably control speaking. However, the mechanism is not well understood [1]. In attempts to clarify its contribution to speech motor control, formant transformed auditory feedback (TAF) experiment have been conducted. In the experiment, when a speaker utters while listening to a speech in which the formant frequency of a vowel of spontaneous speech is perturbed in real time, the speaker utters so as to compensate for the perturbation [2].

Although linear predictive coding (LPC) [3] has been used for formant estimation in many previous studies [4,5], the naturalness of transformed speech deteriorates due to estimation error [6]. Therefore, we have proposed a phase equalization-based autoregressive exogenous model (PEAR) [7] to improve formant estimation accuracy, and have shown that the compensatory response to perturbations in PEAR is greater than that in LPC [8] and that brain activity patterns are different [9]. These results suggest that the naturalness of transformed speech affects auditory feedback experiments.

A difference between PEAR and LPC must exist in the transformed speech, but if we listen to and compare transformed speech after an experiment, we can hardly tell the difference. Thus, we hypothesized that the difference between PEAR and LPC becomes noticeable in the context of listening while speaking, i.e., during formant TAF experiments. Such a difference between passive and active listening is known in brain research [10], but subjects have difficulty rating the naturalness of transformed speech during a formant TAF experiment because doing so imposes a burden on the subject and may affect the formant TAF experiment itself (e.g., [11]).

In this study, we tested the above hypothesis by classifying existing speech data of the formant TAF experiment comparing PEAR and LPC [8] by using a convolutional neural network (CNN) [12]. If the hypothesis is correct, then PEAR and LPC should not be classified by transformed speech alone, but by both uttered speech and transformed speech.

2. Formant estimation methods

Formant TAF experiments require a real-time formant estimation method. LPC is widely used in such experiments because of its theoretical relationship with the speech production model and its low computational complexity. However, since LPC assumes that the source signal is Gaussian noise, an error occurs in the estimated formant frequency [6]. An autoregressive exogenous model (ARX) [13] has been proposed that uses a pulse train corresponding to the fundamental period as a sound source signal, but it assumes a minimum phase difference from the phase characteristic of the input speech. PEAR [7] solves this problem by combining ARX with phase equalization [14], which converts the phase characteristics of input speech to the minimum phase.

Let s be the input signal, e be the LPC residual signal, R be the autocorrelation function of the input speech, and w be the window function. PEAR can determine LPC coefficients a that are robust to the fundamental frequency by solving the following equation:

$$\begin{pmatrix} R(0) & \dots & R(P-1) \\ \vdots & \ddots & \vdots \\ R(P-1) & \dots & R(0) \end{pmatrix} \begin{pmatrix} a(1) \\ \vdots \\ a(P) \end{pmatrix} = \begin{pmatrix} R(1) - V(1) \\ \vdots \\ R(P) - V(P) \end{pmatrix}, \quad (1)$$

with

$$V(p) = \sum_{i=0}^I w(t_i - p)w(t_i) \sum_{\tau=-M}^M e(t_i - \tau)s(t_i - p - \tau),$$

where t_i is a pitch mark (glottal closing instant or peak time of residual signal), P is the dimension of the LPC coefficients, $I+1$ is the number of pitch marks in a frame, and $2M+1$ is the number of taps of the phase equalization filter. The convolution operation $E * S$ on the right side represents the phase equalization process, and the subtraction from R represents the ARX. The formant frequencies are obtained by finding the roots of the LPC polynomial. Assuming Gaussian noise as source signals, $V(p) = 0$ is obtained when there is no t_i . In that case, Eq. (1) is equivalent to the

*e-mail: taguchi.f.664@s.kyushu-u.ac.jp

Fig. 1 Formant TAF experiment.

autocorrelation method for LPC. Thus, PEAR can be said to be a generalization of LPC.

The formant estimation error with LPC prevents pulsing of the residual signal. As a result, in formant TAF experiments using LPC, the naturalness of the transformed speech is reduced due to the interaction between the nonpulsed residual signal and the converted formant frequency. PEAR, on the other hand, is less prone to this problem [8].

3. Experiments

3.1. Formant TAF experiment data

In the formant TAF experiment [8], fifteen native Japanese speakers (eight females) spoke Japanese /he/ as soon as text appeared on the display. The speech signals from the microphone (Utter) were transmitted to a formant TAF system to generate formant-shifted speech signals. The transformed speech signals (Trans) were presented via headphones with a delay of 12 ms (Fig. 1). The sampling frequency was 8 kHz. Pink noise was added to the sound from headphones to mask bone-conducted signal. Trans compensated for the delay to Utter and did not include pink noise. Note that two speakers were excluded from the analysis because we failed to record their transformed speech: The data for thirteen speakers (seven females) were used in the analysis.

Figure 2 shows perturbation patterns of the first (F1) and second (F2) formant frequency in the vowel interval in blocks comprising 140 uttering trials. Either PEAR or LPC was selected in each block. A block consisted of four phases: Baseline (Trial 1–20), Ramp (Trial 21–70), Hold (Trial 71–90), Return (Trial 91–140). In the Baseline phase, speakers produced utterances with no transformed feedback. In the Ramp phase, speakers produced utterances while receiving transformed speech feedback. The absolute value of the magnitude of perturbation in the formant frequency was increased by a certain amount per trial. In the Hold phase, speakers uttered while receiving feedback speech with formant frequencies transformed at the maximum level of perturbation. In the Return phase, speakers produced utterances with normal feedback, which was the same as in the Baseline phase, to wash out aftereffects. The maximum levels of perturbation for formant frequencies were $(F1, F2) = (+200, -200)$ Hz. In order to increase the reliability of the results, speakers performed the test conditions twice.

Fig. 2 F1 and F2 perturbation pattern.

In [8], the amounts of compensatory response in formant frequencies were determined by subtracting the value in the Baseline phase from that in the Hold phase, and it was shown that the compensatory response of F1 for the /he/ with PEAR was significantly larger than that with LPC. Note that data for /hi/ were excluded from this analysis because there was no significant difference in the compensatory response between PEAR and LPC [8].

3.2. Classification method

A model for classifying PEAR and LPC by leave-one-out cross-validation for a speaker was created from the speech in the Hold phase, and classification from the speech for the unknown speaker in the Hold phase was carried out using the obtained model. Since only the evaluation in the Hold phase was carried out in the past study [8], this study also carried out the classification in the Hold phase. That is, the task was to classify the formant estimation method of PEAR or LPC from the speech at the maximum perturbation. The input feature was the 40-th order mel-spectrum (Fig. 3). There were three types of inputs: (1) Utter alone, (2) Trans alone, and (3) two-channel input of Utter and Trans (2 ch). The output is the probability that the input feature is PEAR. Models were created for each input type: There were 39 models in total.

The classification model is a CNN (Fig. 4). Here, the number n in the figure indicates that the filter size is n in the frequency direction. Since the stationary part of the vowel is used as the input feature, convolution in the time direction was not performed in this study. It is known that the accuracy

Fig. 3 Example of mel-spectrogram.

Fig. 4 Structure of CNN and domain adversarial losses for speakers and trials for learning. Conv1D: One-dimensional convolutional layer. BN: Batch normalization layer. FC: Fully connected layer. MLP: Multi-layer perceptron.

of automatic speech recognition can be improved with a CNN by using correlation in the frequency direction of speech [15]. CNN is also effective for 2-ch input. The number of input frames in one trial was 20 (totally 200ms) extracted from the stationary part (40–80%) of the vowel interval. If the utterance section was larger than 20 frames, 20 frames were randomly extracted from the utterance section for each iteration, and if the utterance section was smaller than 20 frames, padding was performed by repeating the utterance section. In the classification test, all the stationary parts of the vowel interval were used. We use the Adam optimizer [16] with dropout = 0.5, weight decay = 10^{-5} , and gradient clipping = 5. We set batch size to 512 and convolution filter number to 64.

To promote generalization from twelve speakers to unknown speakers, we introduced domain adversarial losses [17] for speakers and trials. A network for estimating a speaker and a network for estimating a trial from an input to a fully connected layer of CNN were constructed, and CNN was trained so that the loss of speaker and trial classification was maximized by using a gradient reversal operation. This makes it difficult for the CNN to classify speakers and trials, so the CNN is expected to be robust against unknown speakers. The multi-layer perceptron consists of two layers, each having 64 units (Leaky ReLU function). Only the CNN parts in Fig. 4 were used for inference. All error functions used cross entropy.

3.3. Classification results

Figure 5 shows the average identification rate for each input type. If the output of the CNN is greater than 0.5, it is judged as PEAR; otherwise, it is judged as LPC, and the identification rate is obtained from the correct number of answers. Therefore, the higher the identification rate, the better the discrimination between PEAR and LPC. A one-sample *t*-test revealed that the average identification rate was not significantly different from the chance level of 50% for both Utter alone ($t(12) = -0.14$, $p = 0.44$) and Trans alone

Fig. 5 Classification results. Error bars indicate standard error of the mean.

($t(12) = -0.15$, $p = 0.44$). On the other hand, the average identification rate was significantly higher than 50% for the 2-ch inputs of Utter and Trans ($t(12) = 2.01$, $p < 0.05$). The effect size between Trans alone and 2 ch was 0.51.

4. Discussion

It is obvious that PEAR and LPC cannot be classified in Utter alone, because no speech transformation is performed. The result that Trans alone is also unclassifiable is consistent with the fact that we can hardly tell the difference between them after an experiment, as mentioned in the introduction. This means that the speech transformed using LPC does not have a feature common to the speaker, which is the basis of unnaturalness, because the degree of naturalness of the transformed speech differs depending on the speaker due to various factors such as the residual signal and the amount of perturbation. On the other hand, modeling the relationship between Utter and Trans by 2 ch represents a kind of speech perception-production interaction mechanism. Therefore, the result that can be classified in 2 ch supports the hypothesis that the difference between PEAR and LPC becomes noticeable in the context of listening while speaking. Furthermore, the magnitude of effect size indicates the effectiveness of 2-ch input compared to Trans alone. These suggest that even if the auditory system alone cannot perceptually detect the difference in the quality of the feedback speech, it can be detected by feedback-based speech motor control based on sensory prediction during speaking.

To further test this hypothesis, using the same 2-ch model as that constructed in Sect. 3.3, we calculated the identification rate for 2-ch inputs that were mismatched by randomizing the Utter and Trans trials in the same speaker. The rate dropped to 53.4% (2.7 S.E.), which was not significantly different from 50% ($t(12) = 1.23$, $p = 0.12$). This suggests that the causality between uttered and feedback speech may be important in the prediction of auditory feedback (e.g., [18]).

It is known that the perception of the same speech differs between passive listening and listening while speaking. Curio *et al.* [10] have reported a phenomenon of speaking-induced auditory suppression in which the magnitude of brain activity in the auditory cortex when listening to one's own voice during speaking is reduced compared to when listening to the same voice recorded and played back. Daliri *et al.* [19] used

the formant TAF system to measure formant frequency discrimination thresholds during active and passive listening and showed that the discriminations were more ambiguous for active listening. Therefore, the results of our study suggest that the presence or absence of auditory feedback prediction may explain these phenomena.

In such a CNN study, it is important to construct a speaker-independent model. In this study, domain adversarial loss was used. If it had not been used, the learning of the model would not have converged, and the variance of the identification rate among speakers would have increased. Thus, domain adversarial loss is considered to be effective when dealing with data obtained with a relatively small number of speakers such as experimental data.

However, since this study only showed that PEAR and LPC can be classified by 2-ch input, the relationship between the model and the human auditory-speech motor mechanism remains unclear. Moreover, from this study using a CNN, it is difficult to understand which frequencies contribute to the naturalness of the transformed speech. Exploring these issues remains as future work.

5. Conclusions

In this study, formant TAF experimental data directly comparing PEAR and LPC were classified by using a CNN. As the result, it was proved that the average identification rate was significantly high in the case of two-channel input of Utter and Trans. This suggests that speech prediction is important in determining the naturalness of feedback speech.

References

- [1] S. Hiroya, “Brain science of “speaking and listening to speech”: The relationship between speech production and perception in brain,” *J. Acoust. Soc. Jpn. (J)*, **73**, 509–516 (2017) (in Japanese).
- [2] J. F. Houde and M. I. Jordan, “Sensorimotor adaptation in speech production,” *Science*, **279**(5354), 1213–1216 (1998).
- [3] F. Itakura and S. Saito, “A statistical method for estimation of speech spectral density and formant frequencies,” *Electron. Commun. Jpn., A*, **53**, 36–43 (1970).
- [4] D. W. Purcell and K. G. Munhall, “Compensation following real-time manipulation of formants in isolated vowels,” *J. Acoust. Soc. Am.*, **119**, 2288–2297 (2006).
- [5] V. M. Villacorta, J. S. Perkell and F. H. Guenther, “Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception,” *J. Acoust. Soc. Am.*, **122**, 2306–2319 (2007).
- [6] S. Hiroya, “Formant analysis of vowels: Process and hypotheses,” *J. Acoust. Soc. Jpn. (J)*, **70**, 538–544 (2014) (in Japanese).
- [7] H. Oohashi, S. Hiroya and T. Mochida, “Real-time robust formant estimation system using a phase equalization-based autoregressive exogenous model,” *Acoust. Sci. & Tech.*, **36**, 478–488 (2015).
- [8] Y. Uezu, S. Hiroya and T. Mochida, “Vocal-tract spectrum estimation method affects the articulatory compensation in formant transformed auditory feedback,” *Acoust. Sci. & Tech.*, **41**, 720–728 (2020).
- [9] S. Hiroya and T. Mochida, “Neural mechanisms underlying the impact of speech sound naturalness during transformed auditory feedback,” in *Society for Neurobiology of Language* (2018).
- [10] G. Curio, G. Neuloh, J. Numminen, V. Jousmäki and R. Hari, “Speaking modifies voice-evoked activity in the human auditory cortex,” *Hum. Brain Mapp.*, **9**, 183–191 (2000).
- [11] Y. Uezu, S. Hiroya and T. Mochida, “A study on the relationship between compensatory responses and sense of self-agency under formant transform auditory feedback,” *Proc. Autumn. Meet. Acoust. Soc. Jpn.*, pp. 593–594 (2019) (in Japanese).
- [12] J. Schmidhuber, “Deep learning in neural networks: An overview,” *CoRR*, abs/1404.7828 (2014).
- [13] T. Ohtsuka and H. Kasuya, “Robust ARX speech analysis method taking voicing source pulse train into account,” *J. Acoust. Soc. Jpn.*, **58**, 386–397 (2002) (in Japanese).
- [14] M. Honda, “Speech coding using waveform matching based on LPC residual phase equalization,” *Proc. ICASSP 90*, pp. 213–216 (1990).
- [15] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, **22**, 1533–1545 (2014).
- [16] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *Proc. 3rd Int. Conf. Learning Represent.*, 15 pages (2015).
- [17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, **17**, 2096–2030 (2016).
- [18] A. Toyomura and T. Omori, “Auditory feedback control during a sentence-reading task: Effect of other’s voice,” *Acoust. Sci. & Tech.*, **26**, 358–361 (2005).
- [19] A. Daliri and L. Max, “Formant frequency discrimination in adults who stutter,” in *Society for Neuroscience* (2018).