3D Human Head Geometry Estimation from a Speech

Akinobu Maejima[*]
Waseda University

Shigeo Morishima[†]
Waseda University

1 Introduction

We can visualize acquaintances' appearance by just hearing their voice if we have met them in past few years. Thus, it would appear that some relationships exist in between voice and appearance. If 3D head geometry could be estimated from a voice, we can realize some applications (e.g, avatar generation, character modeling for video game, etc.). Previously, although many researchers have been reported about a relationship between acoustic features of a voice and its corresponding dynamical visual features including lip, tongue, and jaw movements or vocal articulation during a speech, however, there have been few reports about a relationship between acoustic features and static 3D head geometry. In this paper, we focus on estimating 3D head geometry from a voice. Acoustic features vary depending on a speech context and its intonation. Therefore we restrict a context to Japanese 5 vowels. Under this assumption, to estimate 3D head geometry, we use a Feedforward Neural Network (FNN) trained by using a correspondence between an individual acoustic features extracted from a Japanese vowel and 3D head geometry generated based on a 3D range scan. The performance of our method is shown by both closed and open tests. As a result, we found that 3D head geometry which is acoustically similar to an input voice could be estimated under the limited condition.

2 Feature Extraction

We constructed a database which contains 70 individuals' range scans (50 males/20 females) with neutral faces and 5 Japanese vowel speeches for each individual. For a geometric feature, 3D head models are semi-automatically generated from each range scan in the DB based on the Radial basis functions and the non-rigid ICP algorithm [Amberg et al. 2007] with a template head model consisted of 1621 vertices and 3174 triangle polygons. We then apply Principal Component Analysis (PCA) to 3D coordinates of all generated head models to reduce its dimension (from 1621 to 90) while holding 95% variance. Here, these principal components for a 3D head model are referred to as Geometric Feature (GF). For an acoustic feature, we extract 13 Mel-frequency Cepstral Coefficients (MFCCs), their delta and delta-delta coefficients from a vowel speech at 10 msec intervals. This is because these coefficients are widely utilized in speech recognition task [Reynolds and Rose 1995] and can represent speaker's characteristics. We then combined these coefficients into a vector and refer to the 39 dimensional vector as an Acoustic Feature (AF). To reduce the dimension of AF, PCA is also applied to all subjects AFs for a vowel in the database while holding 95% variance. Finally, we can obtain 19 dimensional acoustic features for each individual at 10 msec intervals.

3 Mapping between 3D Head Geometry and Acoustic Features using Neural Network

To represent a mapping from an AF to a GF, we use a FNN which has 1 input, 2 hidden and 1 output layers, 19, 180, 180 and 90 neurons with Sigmoid function in each layer. We set the hyper parameter α and β of the sigmoid function to 0.475 and 1.0 respectively based on the 10-hold Cross validation test. To train a FNN for each vowel speech, we use pairs of AFs and GFs for all subjects's vowel speeches in the database. MFCCs represent acoustic characteristics of vocal tract (i.e the shape of mouth cavity). Thus, this mapping means the correspondence between a shape of mouth cavity and the 3D geometry of a head.

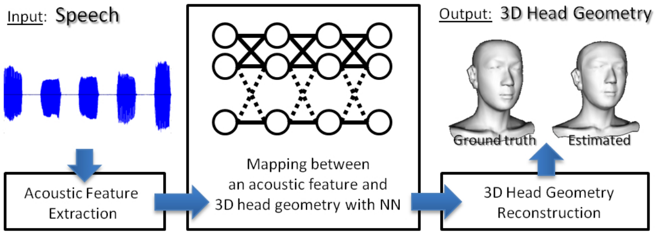
Input: Speech
Output: 3D Head Geometry
Ground truth
Estimated
Acoustic Feature Extraction
Mapping between an acoustic feature and 3D head geometry with NN
3D Head Geometry Reconstruction

Figure 1: *The overview of 3D head geometry estimation*

Table 1: *The estimation accuracy*

	Closed test	10HCV test	SSVS test
RMS (mm)	0.53	6.69	4.18

4 Results and Discussions

To verify the performance, we performed Closed and 10-Hold Cross Validation test (10HCV) using 70 individuals' GFs and AFs for Japanese vowel "a". The Root Mean Squared error between ground truth and the estimate 3D geometries for each test is shown in Table 1. As for the open test, we also evaluate for Same Subjects' Vowel Speeches recorded at the different timing from closed tests' ones (SSVS). The estimated 3D geometries are rendered in the supplemental materials. These results suggest a possibility that we can estimate plausible 3D head geometry from a speech under the limited condition when same subjects speak same vowels. Also, we found that 3D head geometry which is acoustically similar to an input voice could be estimated. Currently, the geometry estimation is sometimes unstable due to the variation of individual acoustic features. We therefore need to look for more robust and text independent acoustic features representing an individual. As a future work, we plan to enlarge the database and to perform subjective evaluations to verify the performance of our method in more detail.

References

AMBERG, B., ROMDHANI, S., AND VETTER, T. 2007. Optimal step nonrigid icp algorithms for surface registration. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1–8.

REYNOLDS, D. A., AND ROSE, R. C. 1995. Robust text-independent speaker identification using gaussian mixture speaker models. In *IEEE Trans. Acoust. Speech and Audio Processing*, vol. 3, 72–83.

[*] e-mail: akinobu@mlab.phys.waseda.ac.jp
[†] e-mail:shigeo@waseda.jp

SIGGRAPH 2012, Los Angeles, California, August 5 – 9, 2012.
ISBN 978-1-4503-1435-0/12/0008