한국음향학회지 제42권 제1호 pp. 67~74 (2023)
The Journal of the Acoustical Society of Korea Vol.42, No.1 (2023)
https://doi.org/10.7776/ASK.2023.42.1.067

pISSN : 1225-4428
eISSN : 2287-3775

영상 기반 음성합성에서 심도 영상의 유용성

The usefulness of the depth images in image-based speech synthesis

이기승[1†]

(Ki-Seung Lee[1†])

[1]건국대학교 전기전자공학부

(Received December 30, 2022; revised January 19, 2023; accepted January 25, 2023)

초 록: 발성하고 있는 입 주변에서 취득한 영상은 발성 음에 따라 특이적인 패턴을 나타낸다. 이를 이용하여 화자의 얼굴 하단에서 취득한 영상으로부터 발성 음을 인식하거나 합성하는 방법이 제안되고 있다. 본 연구에서는 심도 영상을 협력적으로 이용하는 영상 기반 음성합성 기법을 제안하였다. 심도 영상은 광학 영상에서는 관찰되지 않는 깊이 정보의 취득이 가능하기 때문에 평면적인 광학 영상을 보완하는 목적으로 사용이 가능하다. 본 논문에서는 음성 합성 관점에서 심도 영상의 유용성을 평가하고자 한다. 60개의 한국어 고립어 음성에 대해 검증 실험을 수행하였으며, 실험 결과 객관적, 주관적 평가에서 광학적 영상과 근접한 성능을 얻는 것을 확인할 수 있었으며 두 영상을 조합하여 사용하는 경우 각 영상을 단독으로 사용하는 경우보다 향상된 성능을 나타내었다.
핵심용어: 심도영상, 음성 합성, 음성인식, 다중 퍼셉트론

ABSTRACT: The images acquired from the speaker's mouth region revealed the unique patterns according to the corresponding voices. By using this principle, the several methods were proposed in which speech signals were recognized or synthesized from the images acquired at the speaker's lower face. In this study, an image-based speech synthesis method was proposed in which the depth images were cooperatively used. Since depth images yielded depth information that cannot be acquired from optical image, it can be used for the purpose of supplementing flat optical images. In this paper, the usefulness of depth images from the perspective of speech synthesis was evaluated. The validation experiment was carried out on 60 Korean isolated words, it was confirmed that the performance in terms of both subjective and objective evaluation was comparable to the optical image-based method. When the two images were used in combination, performance improvements were observed compared with when each image was used alone.
Keywords: Speech synthesis, Speech recognition, Depth image, Multi-layer perceptron

PACS numbers: 43.72.Ja, 43.72.Kb

I. 서 론

음성은 의사와 정보 전달을 위한 가장 기본적인 수단이다. 음성을 이용한 정보 전달이 불가능한 환경(예: 극심한 잡음 환경, 음성 보안이 요구되는 환경 등)에서는 발성 음 외에 다른 수단을 사용하여 의사 전달이 이루어져야 한다. 이러한 목적으로 음성을 발성하지 않고 발성 행위만으로 의사 전달이 이루어지는 무 음성 대화 기술(Silent speech interface) 이 제안되었다.[1] 무 음성 대화기술에는 음성과 유의한 상관관계를 갖는 신호들이 사용되는 데, 입 주변에서 취득한 근전도 신호,[2] 얼굴 하단의 영상 신호,[3] 초단파[4]/초음파 도플러 신호[5,6] 등이 대표적이다. 이 중 영상 신호는 비접촉, 비 침습 적 취득 방법으로 사용

†**Corresponding author:** Ki-Seung Lee (kseung@konkuk.ac.kr)
Department of Electronic Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea
(Tel: 82-2-450-3489, Fax: 82-2-450-3437)

자에게 불편감이 덜 하며 휴대폰에 장착된 카메라 등을 사용하여 비교적 쉽게 구현할 수 있다는 장점이 있다. 그러나 영상을 이용한 방법은 주변광의 영향을 받을 수 있으며 광원이 반드시 존재해야 한다는 문제가 있다. 이러한 단점은 광 노출이 허용되지 않는 매복 환경이나 조명이 없는 야간에서 사용이 제한될 수 있다.

심도 영상은 2차원 좌표계에서 각 지점의 깊이 분포를 나타내며, 대상 체의 표면 정보를 얻을 수 있다. 입 주변에서 취득한 심도 영상은 발성하는 얼굴 하단의 3차원 정보를 제공하며, 광학 영상과 마찬가지로 발성 음과 유의한 상관관계를 갖는 다고 가정할 수 있다. 본 연구에서는 광학영상의 대안으로서 심도영상을 이용하는 음성 합성 기법을 제안하였다.

심도 영상은 2개 이상의 광학 카메라로 취득된 영상을 조합하여 합성하거나[7] 심도 카메라를 사용한 방법이 있다. 이 중 심도 카메라를 사용한 방법은 근적외선대의 독립된 광원을 사용하기 때문에 광학 영상이 갖는 단점인 주변광 및 불 균일 광원의 영향, 낮은 조도에서 사용 불가 문제를 부분적으로 해결할 수 있다. 그러나 광학 영상에 비해 일반적으로 화질이 떨어지고, 고해상도 심도 영상을 얻기 위해서는 고가의 전문 장비가 필요하다는 단점이 있다.

본 연구의 주 목적은 음성 합성 관점에서 심도 영상의 유용성을 광학 영상과 비교하여 검증하는 것이다. 이를 위하여 심도 카메라를 이용하여 화자 입 주변에서 광학 영상과 심도 영상을 취득하고 동시에 발성 음을 녹음한다. 이들 데이터로부터 각 영상과 발성음간의 대응관계를 추정하였다. 추정된 대응 관계를 이용하여 음성을 합성하고 합성의 품질을 객관적, 주관적인 척도로 평가하였다.

II. 데이터 취득

본 연구에서는 광학영상과 심도영상의 취득을 위해 Intel사의 RealSense LiDAR Camera(Model: L515)를 사용하였다. 이 카메라의 측정 범위는 0.4 m ~ 9 m로서 데이터 취득 시 화자와 카메라간의 거리도 측정 범위 내(0.5 m) 로 유지하도록 하였다. 데이터 취득은 비교적 조용한 환경에서 백색 LED광원이 사용된 조명이 얼굴로 향하게 설치하고 카메라는 삼각대를 이용하여 고정하였다.

음성 신호의 동시 취득을 위해 심도카메라 하단에 별도의 마이크로폰을 장착하였으며 Desktop PC를 통해 음성 신호를 녹음하였다. 1명의 피시험자(남성, 26세)로부터 60개의 한국어 고립어[5]를 50번 반복하여 총 3000개의 발화를 취득하였다. 매 반복 녹음 시 다른 운율과 스타일로 발성하도록 요구되었다. 녹음은 매 500 발화에 대한 녹음 후 일정 시간 휴식을 갖도록 하였다. 영상 촬영 시 피시험자의 머리가 움직이는 것을 억제하기 위해 목 고정대가 장착된 의자에 편한 자세로 앉도록 하였다.

Fig. 1에 취득된 광학영상과 깊이영상의 예를 나타내었다. 모음 “아”를 발성할 때 취득된 영상으로서 깊이 영상은 입술의 안쪽 윤곽 형태를 보이고 있다. 사용된 카메라의 깊이 해상도는 3 mm로서 “아” 발성 시에는 입술이 얼굴 표면과 비교하여 충분히 돌출되지 않았기 때문에 입술의 형태는 희미하게 나타난다. 반면 모음 “우”를 발성할 때는 입술을 오므리면서 전면으로 돌출되어 오른쪽 그림과 같이 입술의 형태가 비교적 뚜렷하게 나타남을 알 수 있다. 이러한 예는 깊이 영상을 통해 입술의 3차원적인 정보를 취득할 수 있음을 나타낸다.

광학 영상과 대비되는 깊이 영상의 또 다른 차이는 피시험자의 피부면의 반점 등이 발견되지 않는다는 점이다. 피부면의 반점, 화장 등은 얼굴면의 균질성을 떨어트리고 자체로 얼굴 내 패턴을 형성하여

Fig. 1. Examples of the acquired optical (top) and corresponding depth (bottom) images when vocalizing vowel “ah” (left) “uh” (right).

음성 추정 시 영향을 끼칠 수 있다. 깊이 영상에서는 이러한 피부 표면의 비 균일성으로 인한 성능 저하를 피할 수 있다. 그러나 깊이 영상은 위 예에서 보듯이 광학 영상에 비해 입술의 형태, 특히 바깥 윤곽 정보를 정확하게 파악할 수 없으며, 이는 발성 음에 따른 입술 모양의 미세한 변화를 관찰하기 어렵다는 것을 의미한다. 이러한 문제는 1 mm 이하의 깊이 해상도를 갖는 깊이 카메라를 사용함으로써 해결이 가능할 것으로 기대된다.

III. 데이터 전처리

취득된 영상은 피시험자의 머리를 고정시켰음에도 불구하고 미소한 움직임에 따라 프레임 마다 입술의 위치가 일정하지 않게 나타났다. 또한 사용된 심도 카메라의 최소 취득 거리가 0.4 m로서, 시야각에 따른 취득 범위가 발성과는 무관한 코와 턱 부분도 포함된다. 따라서 입술부분 만을 추출하는 전처리 과정이 필요하다.

주어진 얼굴 영상에서 입술 부분만을 추출하기 위해, 입술의 형태학적인 특징이 뚜렷하게 나타나는 영역을 영상의 전 영역에서 탐색하는 방법으로 구현할 수 있다.[8] 이러한 기법은 화자마다 각기 다른 입술의 모양을 반영할 수 없으며, 심도 영상에 대해서는 적용이 불가능하다는 문제가 있다. 본 연구에서는 상관 계수를 이용한 반자동 추출 기법을 사용하였다. 먼저 영상열의 첫 번째 영상에 대해 수작업을 통해 입술 영역을 분할한다. 다음 영상에 대해서는 이전 영상에서 분할된 입술 영상과 상관 값이 가장 큰 위치를 탐색하고, 이 좌표 값을 기준점으로 현재 영상의 입술 영역을 얻는다. 즉, n-번째 영상에 대한 입술 영역 S_n은 다음과 같이 나타낼 수 있다.

$$S_n = f_n(x^* \le x \le x^* + \Delta x \ . \\ y^* \le y \le y^* + \Delta y) \tag{1}$$

$$x^*, y^* = \mathrm{argmax}_{x,y} \left[\frac{\sum_{j=0}^{\Delta y-1}\sum_{i=0}^{\Delta x-1} f_n(x-i, y-j) S_{n-1}(i,j)}{\sqrt{\| f_n(x,y) \|^2 \| S_{n-1} \|^2}} \right]. \tag{2}$$

$$\| f(x,y) \|^2 = \frac{1}{\Delta x \Delta y} \sum_{i=0}^{\Delta x-1}\sum_{j=0}^{\Delta y-1} f_n^2(x-i, y-j). \tag{3}$$

$$\| S_{n-1} \|^2 = \frac{1}{\Delta x \Delta y} \sum_{i=0}^{\Delta x-1}\sum_{j=0}^{\Delta y-1} S_{n-1}^2(i,j), \tag{4}$$

여기서 $f_n(x,y)$는 $n-$번째 영상의 좌표 (x,y)에서의 밝기 값을, Δx와 Δy는 각각 입술 영역의 가로, 세로 크기를 나타낸다. 심도 영상에 대한 입술 영역은 광학 영상의 입술 영역 시작 지점 (x^*, y^*)을 두 영상의 크기 차이(광학영상 640×480, 심도영상 320×240)를 반영하여 $(x^*/2, y^*/2)$인 지점을 시작 지점으로 간주하여 추출하였다.

IV. 음성 신호 추정

영상 신호로부터 음성 신호를 추정하는 것은 2차원 데이터와 1차원 데이터 간 대응관계를 추정하는 문제로 간주할 수 있다. 이러한 문제 해결 방법으로 2차원 convolution이 사용된 합성 곱 신경망(Convolutional Neural Networks, CNN) 이 고려될 수 있다. 그러나 본 연구의 실험 결과에 따르면 취득된 2차원 광학/심도 영상에 대해 CNN을 적용하는 경우 매우 낮은 성능이 얻어지는 것으로 관찰되었다. 따라서 본 연구에서는 영상 신호를 직접 사용하지 않고 특징 변수로 변환하여 사용하는 방법을 사용하였다.

일반적으로 영상 신호는 음성 신호에 비해 데이터 량이 많고 이 중 상당수는 잉여 정보로 존재한다고 알려져 있다. 따라서 데이터 감축을 수행하는 것이 바람직하다. 본 연구에서는 입 주변 영상 신호에 대해 이산여현변환(Discrete Cosine Transform, DCT) 또는 주요성분분석(Principle Component Analysis, PCA)을 적용하여 데이터 감축을 수행하고 이를 1차원으로 재배열한 후, 음성 추정을 위한 파라메터로 사용하였다.

기존 무 음성 대화 기술에서는 선형 예측 계수,[3] 멜스펙트럴 계수,[2] 단 구간 푸리어 크기 스펙트럼[5] 등이 음성 파라메터로 사용되었는데 본 연구에서는

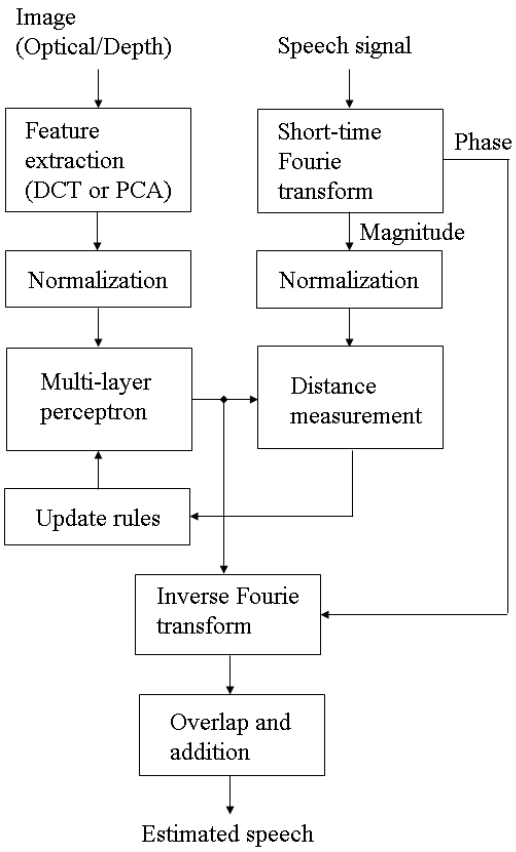

Fig. 2. Block diagram of the proposed image-based speech estimation scheme.

단 구간 푸리어 크기 스펙트럼을 사용 하였다. 48 msec의 길이를 갖는 hamming window를 사용하였으며, 33 msec 만큼 이동시켜 가면서 푸리어 변환을 수행하였다. 33 msec는 광학 영상과 깊이 영상의 프레임 레이트인 30 fps와 일치하는 값이다.

Fig. 2에 본 연구에서 제안한 음성 추정 기법의 블록도를 나타내었다. 영상 특징 변수로부터 음성 파라메터를 추정하기 위해 다중 퍼셉트론(Multi-Layer Perceptron, MLP) 이 사용되었다. MLP의 입력은 영상 신호에 대한 DCT/PCA 값으로서, 현재 프레임에 대한 DCT/PCA값과 함께 인접한 프레임에 대한 값을 함께 사용하였다. 이 경우 MLP의 입력 값은 아래와 같이 나타낼 수 있다.

$$X_T(t) = [\vec{x}(t-T), \ldots \vec{x}(t), \ldots, \vec{x}(t+T)], \tag{5}$$

여기서 $\vec{x}(t)$는 t-번째 프레임에 대한 영상 특징변수를 나타낸다. 인접 프레임의 개수 T는 추정 성능이 최대가 되도록 실험적으로 결정하였다.

완전한 무 음성 대화 기법이 구현되기 위해서는 크기 스펙트럼과 함께 위상 스펙트럼이 함께 추정되어야 하나, 영상 신호와 위상 스펙트럼 간에는 의미 있는 상관관계가 존재하지 않는 것으로 알려져 있다.[1] 기존 무 음성 대화 기법에서는 Griffin과 Lim이 제안한 최소 자승 오차 법,[9] 랜덤 위상 기법[6] 등이 적용되었는데 이러한 방법은 크기 스펙트럼이 본래의 음성과 매우 근접한 경우에만 만족할 만한 성능을 나타내었다. 본 연구에서와 같이 추정된 크기 스펙트럼을 사용하는 경우엔 위상 스펙트럼이 자체의 왜곡이 중첩되면서 합성음의 음질 저하가 크게 나타나는 것으로 관찰되었다. 본 연구에서는 심도 영상의 크기 스펙트럼에 대한 추정 효용성을 검증하는 것이 주요 목적으로서, 위상 스펙트럼의 추정은 고려하지 않고 본래 음성 신호에서 추출된 값을 사용하였다.

MLP의 구조는 학습 데이터의 20 %에 해당하는 데이터로부터 별도의 검증 데이터를 사용하여 결정하였다. 결론적으로 3개의 은닉 계층과 각 은닉 계층의 노드 수는 입력 노드 수의 1.5배로 설정하였다. Sigmoid activation 함수가 사용되었으며 출력 노드에서는 linear함수가 사용되었다. Learning rate는 0.001로, batch size는 100으로 설정하였다. MLP의 학습에 사용된 손실함수는 다음과 같다.

$$L = \frac{1}{N_B}\sum_n [(1-w_P) D_{MSE,n} + w_P(w_s D_n^{(s)} + w_a D_n^{(a)})]. \tag{6}$$

N_B는 batch size를 나타내며 w_P, w_s, w_a는 각각 Perceptual Disturbance(PD), symmetrical, asymmetrical disturbances[10]에 대한 가중치를 나타낸다. $D_{MSE,n}$은 $n-$번째 프레임에 대한 MSE를 나타낸다.

$$D_{MSE,n} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{1}{\sigma_m^2}\left(\log\frac{|P_{m,n}|^2}{|\hat{P}_{m,n}|^2}\right)^2. \tag{7}$$

M은 크기 스펙트럼의 차수를 나타내며, $P_{m,n}$과

$\hat{P}_{m,n}$은 각각 본래 음성과 복원된 음성의 n-번째 프레임의 $m-$번째 파워 스펙트럼을 나타낸다. Disturbance $D^{(s)}$와 $D^{(a)}$는 각각 인간의 청각 특성을 반영한 perceptual domain transformation과 loudness convertsion을 통해 계산되는데 자세한 과정은 Reference [10]에 제시되었다. Eq. (6)에서 $w_P=0$인 MLP는 MSE만 최소화하는 방향으로 학습이 되며, $w_P>0$인 경우는 인간의 청각 특성에 기반 한 왜곡이 고려되어 MLP가 학습된다. 다음 장에서 실험 결과를 이용하여 w_P에 따른 성능 변화를 제시한다.

V. 실험 결과

5.1 객관적 성능 평가

심도 영상의 음성 추정 측면에서 유용성을 평가하기 위해 취득된 데이터를 이용해 검증 실험을 수행하였다. 전체 데이터의 75 %를 음성 추정을 위한 대응 규칙의 학습에 사용하였으며 나머지 데이터를 검증에 사용하였다. 성능 검증을 위한 척도로는 Perceptual Evaluation of Speech Quality(PESQ)[11]와 Root Mean Squared Error(RMSE)를 사용하였다.

먼저 영상 데이터의 feature수와 인접된 프레임 수를 결정하기 위해 두 값을 변경하면서 광학/심도 각 영상에 대한 평균 PESQ값을 관찰하였다. Fig. 3에 이에 대한 결과를 제시하였다. 두 영상 모두 feature수 100인 경우에 높은 PESQ값을 나타내었으며, 인접 프레임수가 증가함에 따라 PESQ값이 증가하는 경향을 보였다. 이는 feature수가 200, 300인 경우 Pearson 상관 값이 모두 0.85이상의 값을 나타낸 것으로 입증되었다. 다만 feature수가 100인 경우, 프레임수가 3인 경우에 가장 낮은 PESQ를 보였으나 대체적으로 프레임수와 PESQ간에 양의 상관관계를 나타내었다. ($R^2=0.765, 0.713$) 유의도 검증[analysis of variance (ANOVA) test]에서 광학/심도 영상 모두 인접 프레임 개수 가 평균 PESQ에 유의한 영향을 끼치는 것으로 나타났으며(p=0.02) feature수는 상대적으로 낮은 영향을 나타내었다(p=0.3). 두 광학/심도 영상 간 큰 차이는 관찰되지 않았으며 이를 통해 심도 영상이 광학 영상과 음성 추정 면에서 유사한 성능을 갖는 것

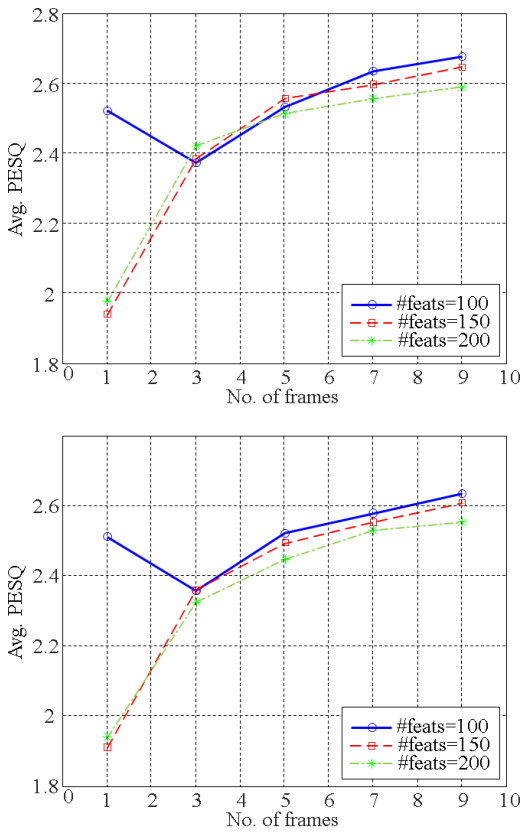

Fig. 3. (Color available online) Average PESQs for optical image (top) and depth image (bottom) for the different number of image features and frames.

으로 확인되었다. Fig. 3의 결과는 영상 특징 변수로 DCT를 사용한 경우이며, PCA를 사용한 경우도 이와 유사한 결과를 얻었다. 실험 결과로부터 이후 실험은 영상 feature수=100, 인접 프레임 수=9로 설정하여 수행되었다.

Fig. 4에 Eq. (6)에 제시한 PD의 가중치 w_P를 변경시키면서 얻어진 평균 PESQ와 RMSE값을 영상/특징 변수별로 도시하였다. PD는 인간의 청각 특성을 반영한 왜곡 척도이며, PESQ와 높은 관련성을 갖는 것으로 알려져 있다.[10] 따라서 w_P가 증가되면 PESQ도 증가되는 것으로 예상할 수 있으나 실험 결과에서는 MSE와 PD가 동일하게 강조되는 경우($w_P=0.5$)에 높은 PESQ 값이 얻어짐을 알 수 있다. 이는 PD값 만을 최소화하는 경우 추정 스펙트럼이 본래 스펙트럼과 형태적으로 큰 차이를 보이는 것에 기인한다. Fig. 4의 하단에 제시한 RMSE에서 $w_P=1$인 경우 값이

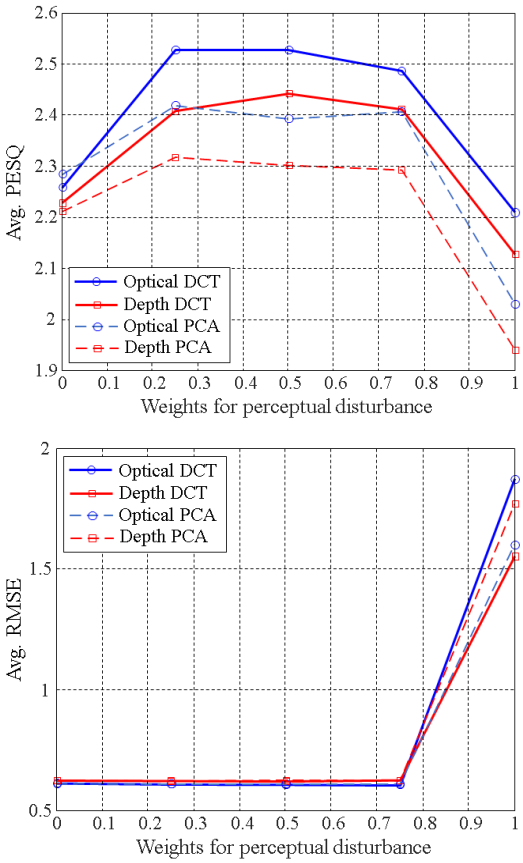

Fig. 4. (Color available online) Average PESQ (top) and RMSE (bottom) according to the weights of perceptul disturbance, for optical/depth images when DCT and PCA were adopted as image feature.

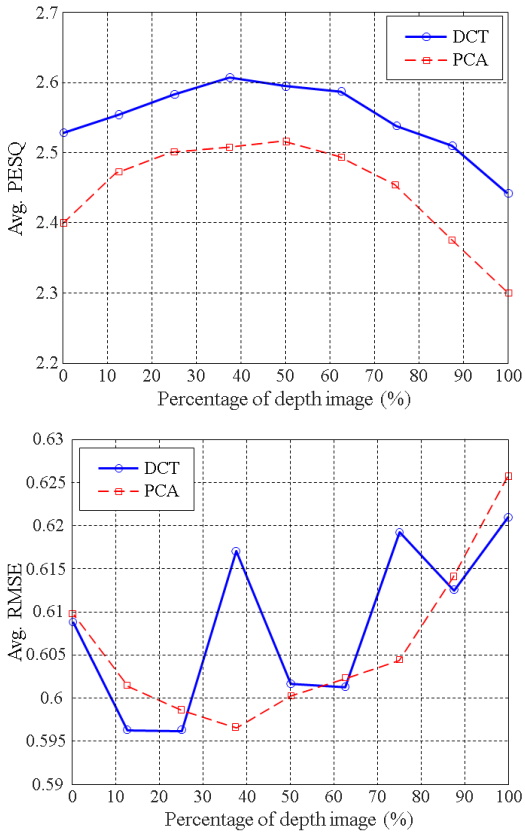

Fig. 5. (Color available online) Average PESQ (top) and RMSE (bottom) according to the percentage of depth image, when DCT and PCA were adopted as image feature.

크게 증가하는 것이 이를 입증한다. Fig. 4의 결과를 보면 $w_P < 1$에서 광학/심도 영상 간, DCT/PCA특징 변수 간 차이가 PESQ에는 비교적 뚜렷하게 나타나지만 RMSE는 거의 없음을 알 수 있다. 이는 $w_P < 1$에서는 MSE값을 일정하게 유지하면서 PD값만을 최소화하는 방향으로 MLP의 학습이 이루어짐을 나타낸다.

DCT를 특징 변수로 사용하는 경우가 PCA와 비교하여 높은 PESQ값을 나타내었으며, 광학 영상이 심도 영상보다 약간 높은 PESQ를 나타내었다. 이는 Fig. 3에 제시한 결과와 일치하는 것으로, 입모양을 보다 세밀하게 표현할 수 있는 광학 영상이 음성 합성 측면에서 상대적으로 높은 유용성을 갖는 것으로 해석할 수 있다.

Fig. 5에 광학영상과 깊이 영상을 조합하여 사용한 경우 PESQ와 RMSE값을 제시하였다. 이 결과는 $w_P = 0.5$로 설정한 경우에 얻어진 것이다. 두 영상의 조합은 한 프레임에 대한 DCT(또는 PCA) 계수의 전체 개수를 200으로 설정하여 광학/심도 영상의 비율에 따라 각 계수를 연결함으로써 구현된다. 예로서 50 %비율로 조합하는 경우 광학 영상에 서 얻은 100개의 DCT(또는 PCA) 계수와 심도 영상에서 얻은 100개의 DCT(또는 PCA) 계수를 서로 조합하여 200차원의 크기를 갖는 특징벡터를 구성한다. Fig. 5의 가로축은 심도 영상의 비율을 나타낸 것으로 0 %와 100 %는 각각 광학 영상, 심도 영상만으로 음성을 추정하는 경우에 해당한다. 평균 PESQ를 살펴보면 이전 결과와 동일하게 광학 영상만을 사용한 경우(가로축 0 %) 가 심도 영상만 사용한 경우와 비교하여 약간 높게 나타난 것을 알 수 있다. 마찬가지로, RMSE도 광학 영상만 사용한 경우가 낮게 나타났다. DCT를 특

징 변수로 사용한 경우 심도 영상의 비율에 따른 RMSE 값의 변동이 크게 나타났다. 이는 $w_P = 0.5$에서 MLP의 초기 값에 따라 Eq. (6)의 MSE항 또는 PD항이 우선적으로 최소화되는 방향으로 학습되면서 나타난 현상으로 $w_P \neq 0.5$에서는 PESQ와 마찬가지로 RMSE도 포물선 형태를 나타내었다.

최대 평균 PESQ값은 DCT계수를 사용한 경우 광학 영상의 37.5 %, PCA를 사용한 경우 광학 영상의 50 %가 반영되었을 때였다. 앞선 결과와 동일하게 DCT 계수를 사용한 경우가 PCA 보다 대략 1.0 정도 높은 PESQ를 나타내었다. 이러한 결과는 광학 영상을 단독으로 사용하는 것 보다는 심도 영상을 일부 조합하여 사용하는 것이 본래 음성과 더 가까운 합성음이 얻는데 유리함을 나타낸다. 다만 최대 PESQ는 광학 영상의 비중이 약간 높은 경우에 얻어졌는데 이는 광학 영상이 심도 영상에 비해 음성 추정 면에서 상대적으로 더 유용함을 의미한다. 본 연구에서 사용된 카메라는 얼굴의 3차원적 구조를 영상화하는 목적 보다는 자율 운행 체와 같은 중/장거리 방해물의 탐지가 주된 용도라 할 수 있다. 만약 발성에 따라 특이적으로 나타나는 얼굴의 3차원 정보를 세밀하게 영상화할 수 있는 고해상도 카메라를 사용한다면 음성 합성과 관련된 유용성이 더욱 증가할 것이라고 판단된다.

5.2 주관적 성능 평가

합성음의 품질을 청취자 관점에서 평가하기 위해 informal listening test를 수행하였다. 청취 테스트에 참여한 시험자는 총 10명으로 연령분포는 23세 ~ 55세(중앙값 28) 남성 8명, 여성 2명이었고 알려진 청각 장애 병력은 모두 없었다. 시험자는 비교적 조용한 환경에서 헤드폰을 착용하고 음성을 청취하였다, 총 75개의 합성음 중 50개를 랜덤하게 선택하여 들려주었으며 합성음의 품질을 인지도와 자연성 관점에서 평가하도록 하였다. 시험자는 각 발화에 대해 총 3개의 음성(원 음성, 광학 영상으로 합성된 음성, 심도 영상으로 합성된 음성)을 청취하여 두 개의 합성음 중 선호하는 음성을 선택하도록 요청되었으며 선호하는 음성이 없는 경우 "No preference"를 선택하도록 하였다. 시험 자는 판단 전까지 횟수의 제한 없이 반복

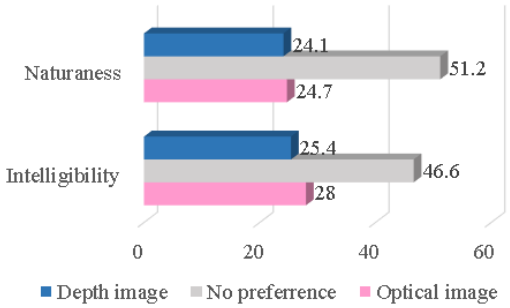

Fig. 6. (Color available online) Subjective listening test result.

청취가 가능하였다.

청취 테스트의 결과를 Fig. 6에 제시하였다. 인지도와 자연성 모두 광학 영상으로부터 합성된 음성이 약간 높게 나타났으나 두 음성 간의 차이는 유의하게 나타나지 않았다($p = 0.1$). 이는 객관적 척도의 결과와 일치하는 것이다. 선택 항목 중 "No preference"가 가장 높은 빈도를 보였으며 이는 두 음성간의 차이가 청취자에게 크게 느껴지지 않았음을 나타낸다. 청취자가 합성음성을 올바르게 인식하는 비율은 광학 영상, 심도 영상 각각에 대해 73.3 %, 67.8 %로 나타났다. 상당수의 청취자는(8명) 두 합성음의 품질이 원래 음성과 비교하여 명료성이 다소 떨어진다는 의견을 제시하였는데 이는 추정된 음성의 스펙트럼을 관찰하였을 때 미세 구조가 대부분 손실된 것에 기인 된다. 결론적으로, 주관적 청취 테스트에서도 심도 영상을 이용하여 합성된 음성은 기존 광학 영상으로 합성된 음성과 비교하여 큰 차이가 나타나지 않음을 확인할 수 있었다.

VI. 결 론

가시 광 조명을 사용하는 기존 영상 기반 음성합성 기법의 대안으로 본 논문에서는 심도 영상을 사용한 음성합성 기법을 제안하였다. 신경 회로망을 이용하여 영상과 음성 간의 대응 관계를 추정하였으며, 한국어 60개 고립어에 대한 검증 실험을 수행하였다. 실험 결과 기존의 광학 영상 기법이 약간 우수한 성능을 보였으나 심도 영상을 사용하여 합성된 음성과 큰 차이를 나타내지 않았다. 조명 광량의 변동, 저조도 환경에서 광학 영상 기반 음성합성 기법

은 성능 변동이 크게 나타날 것으로 예측되며 심도 영상은 이러한 환경에서도 안정된 성능을 나타낼 것으로 기대된다. 추후 연구는 이러한 상황에서의 심도 영상의 유용성을 검증하고자 한다.

감사의 글

본 논문은 한국연구재단 연구과제인 "심도영상을 이용한 무 음성 대화 기술 개발"(과제번호: 2022R1F1A1068979112068207325010 1) 의 연구 결과 중 일부입니다.

References

1. B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," Speech Comm. **52**, 270-287 (2010).
2. K.-S. Lee, "EMG-based speech recognition using hidden markov models with global control variables," IEEE Trans. Biomed. Eng. **55**, 930-940 (2008).
3. I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," IEEE Trans. Audio, Speech, Language Proc. **19**, 1642-1651 (2011).
4. S. Li, Y. Tian, G. Lu, Y. Zhang, H. Lv, X. Yu, H. Xue, H. Zhang, J. Wang, and X. Jing, "A 94-GHz milimeter-wave sensor for speech signal acquisition," Sensors, **13**, 14248-14260 (2013).
5. K.-S. Lee, "Speech synthesis using Doppler signal" (in Korean), J. Acoust. Soc. Kr. **35**, 134-142 (2016).
6. K.-S. Lee, "Ultrasonic doppler based silent speech interface using perceptual distance," Appl. Sci. **12**, 827 (2022).
7. M. A. Subhi, S. H. M. Ali, A. G. Ismail, and M. Othman, "Food volume estimation based on stereo image analysis," IEEE IMM, **6**, 36-43 (2018).
8. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proc. IEEE CSPV, 511-518 (2001).
9. D. W. Griffin and J. S. Lim, "Signal estimation from the modified short-time fourier transform," IEEE Trans. on Acoustic, Speech Signal Proc. **32**, 236-243 (1984).
10. J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," IEEE Signal Process. Lett. **25**, 1680-1684 (2018).
11. ITU-T, Rec. P. 862, *Perceptual Evaluation of Speech Quality(PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow Band Telephone Networks and Speech Codecs,* Int. Telecomm. Union-Telecomm. Stand. Sector, 2001.

저자 약력

▸ 이 기 승 (Ki-Seung Lee)

1991년 2월 : 연세대학교 전자공학과 학사
1993년 2월 : 연세대학교 전자공학과 석사
1997년 2월 : 연세대학교 전자공학과 박사
1997년 10월 ~ 2000년 9월 : AT&T Labs-Research, Senior Technical Staff.
2000년 11월 ~ 2001년 8월 : 삼성전자(주) 종합기술원 전문 연구원
2001년 9월 ~ 현재 : 건국대학교 전기전자공학부 교수.