ESTIMATION OF FUNDAMENTAL FREQUENCY FROM SURFACE ELECTROMYOGRAPHIC DATA: EMG-TO-F_0

Keigo Nakamura, Matthias Janke, Michael Wand, and Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

kei-naka@is.naist.jp, {matthias.janke, michael.wand, tanja.schultz}@kit.edu

ABSTRACT

In this paper, we present our recent studies of F_0 estimation from the surface electromyographic (EMG) data using a Gaussian mixture model (GMM)-based voice conversion (VC) technique, referred to as EMG-to-F_0. In our approach, a support vector machine recognizes individual frames as unvoiced and voiced (U/V), and voiced F_0 contours are discriminated by the trained GMM based on the manner of minimum mean-square error. EMG-to-F_0 is experimentally evaluated using three data sets of different speakers. Each data set includes almost 500 utterances. Objective experiments demonstrate that we achieve a correlation coefficient of up to 0.49 between estimated and target F_0 contours with more than 84% U/V decision accuracy, although the results have large variations.

Index Terms— Electromyography, Voice conversion, Fundamental frequency, Feature estimation.

1. INTRODUCTION

Speech is the most convenient way for communication. However, speech-driven technologies still face some major challenges like the degradation of performance in the presence of noises. In recent years, some other devices to alleviate this problem have been proposed, such as bone-conductive microphones, non-audible murmur microphones [1], and surface electromyography (EMG) [2]. EMG is a technique to record activation potentials of the articulatory muscles from the speaker's face with surface electrodes. Fig. 1 shows the positioning of electrodes as used in this paper [3]. One of the applications of EMG is its use as a silent speech interface that allows people to communicate without the necessity of emitting an audible acoustic signal. To implement a silent speech interface, input EMG signals have to be converted to output text information [4] or to synthesize speech waveforms [5] so that the listeners can understand what the speaker uttered.

The authors are grateful to Professor Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method. This research was also supported in part by Grant-in-Aid for JSPS Fellows.

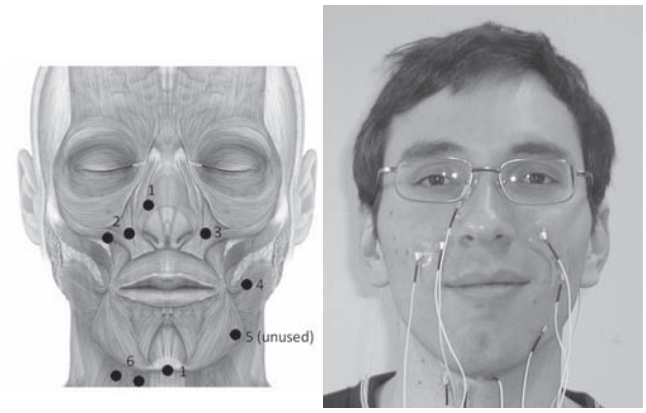

Fig. 1. Electrode positions to record EMG data [3].

Toth *et al.* [5] investigated the spectral estimation accuracy from EMG data using a voice conversion (VC) technique. They showed the possibility of EMG-based speech synthesis, although there are still some difficulties to improve the performance. They used the fundamental frequency (F_0) of simultaneously recorded audible speech for synthesizing speech waveforms. However, it would be more useful if not only spectral information but also F_0 information is estimated directly from EMG data.

This paper proposes an F_0 estimation approach from EMG data using the VC technique, referred to as EMG-to-F_0. Our approach first uses a support vector machine (SVM) [6] to discriminate unvoiced and voiced (U/V) frames. An SVM is basically used to split two patterns in order to calculate a sub-space that maximizes the distance between individual samples and the sub-space. After discriminating U/V frames, F_0 values are generated by the statistical VC method using a Gaussian mixture model (GMM) [7, 8] while keeping the unvoied frames unaltered.

This paper is organized as follows. The GMM-based VC method is briefly described in **Section 2**. In **Section 3**, EMG-to-F_0 is explained. The proposed method is experimentally evaluated in **Section 4**. Finally, we conclude this paper in **Section 5**.

2. GMM-BASED VOICE CONVERSION

VC is a feature modification technique that causes input speech (referred to as the source speech) to sound as if it is uttered by another person (referred to as the target speech). In this paper we focus on the GMM-based VC method [7, 8], in which a GMM describes the relationship between source and target acoustic features. The trained GMM can accept arbitrary input utterances.

This VC method consists of the training and conversion parts. Source and target speakers are firstly defined, and utterances of the same contents are recorded. These two speech signals are automatically aligned by dynamic time warping.

Let us define a static source and target feature vector at frame t as $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(d_x)]^\top$ and $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(d_y)]^\top$, respectively, where d_x and d_y denote the dimension of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, respectively. $\top$ denotes transposition. For training data of the source speech at frame t, a feature vector capturing the dynamic movement is used, which is denoted as $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$. For training data of the target speech, we use a similar dynamic feature vector $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$. After preparing the training data, a GMM is trained to describe the joint probability density of the source and the target feature vectors as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda) = \sum_{m=1}^{M} w_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right),$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix},$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. m denotes a mixture component index, and M denotes the total number of the mixture components. The parameter set of the GMM is denoted by λ, which consists of weights w_m, mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$ and full covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components. $\boldsymbol{\mu}_m^{(X)}$ and $\boldsymbol{\mu}_m^{(Y)}$ represent the mean vectors of the mth mixture component for the source and the target features, respectively. $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ represent the covariance matrices and $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$ represent the cross-covariance matrices of the mth mixture component for the source and the target features, respectively.

In this paper, we implement a conversion method from [7, 8] that is performed based on the manner of minimum mean-square error as follows:

$$\begin{aligned} \hat{\boldsymbol{Y}}_t &= \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \lambda\right) \boldsymbol{E}_{m,t}^{(Y)}, \\ P\left(m|\boldsymbol{x}_t, \lambda\right) &= \frac{w_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{n=1}^{M} w_n \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_n^{(X)}, \boldsymbol{\Sigma}_n^{(XX)}\right)}, \\ \boldsymbol{E}_{m,t}^{(Y)} &= \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right), \end{aligned}$$

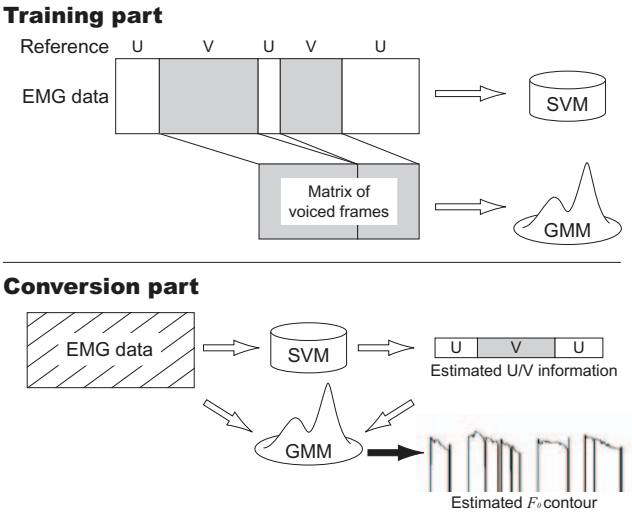

Fig. 2. Framework of EMG-to-F_0 using GMM and SVM. 'U' and 'V' denote unvoiced and voiced, respectively.

where $\hat{\boldsymbol{Y}}_t = [\hat{\boldsymbol{y}}_t^\top, \Delta\hat{\boldsymbol{y}}_t^\top]^\top$ is the target joint feature vector consisting of the estimated static feature vector $\hat{\boldsymbol{y}}_t^\top$ and dynamic feature vector $\Delta\hat{\boldsymbol{y}}_t^\top$ at frame t. We regard $\hat{\boldsymbol{y}}_t^\top$ as the finally estimated Log-scaled F_0 feature vector.

3. EMG-TO-F_0 VOICE CONVERSION

Fig. 2 shows the framework of EMG-to-F_0 using GMM-based VC. Our preliminary experiments indicated that U/V discrimination based on a GMM is insufficient. Since the task is a binary classification problem, we decided to apply an SVM for U/V discrimination.

The data for the GMM training consists of EMG feature vectors as the source data and two-dimensional vectors of the form $[F_0, \Delta F_0]^\top$ as the target data, where only voiced frames are used. We set segmental feature vectors constructed in the following method as the training data of the source EMG data; (1) concatenating current and several previous and succeeding frames, and (2) performing linear discriminant analysis (LDA) to reduce the dimension. A similar approach was successfully used to construct segmental features for source data in previous studies [9, 10]. Constructing segmental feature vectors is expected to capture more complex information than simple first or second order derivative information.

In the conversion part, segmental feature vectors of the source EMG data are constructed in the same manner as in the training part. First, the segmental EMG vectors are input to the trained SVM to decide voicedness for individual frames. Then the GMM generates F_0 values from voiced frames of segmental EMG vectors while keeping the unvoiced frames unaltered.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We used three data sets spoken by three different male speakers, referred to as *spk-1*, *spk-2*, and *spk-3*. We recorded EMG signals and normal speech simultaneously to obtain time-aligned data. EMG-to-F_0 is conducted within individual speakers. The electrode setting can be seen in Fig. 1. We used five channels and captured signals from 1) the levator angulis oris, 2) the zygomaticus major, 3) the platysma, 4) the anterior belly of the digastric and 5) the tongue. All EMG signals were sampled at 600 Hz and filtered with an analog high-pass filter.

Spk-1 recorded 50 unique sentences for 10 times in Taiwanese accented English. 380 sentences were used for GMM and SVM training, and the remaining 120 sentences were used for testing. Note that the test sentences were not included in the training. Also note that the amplifier for EMG signals was different from the setting for *spk-2* and *spk-3*. This data set of *spk-1* was identical to the previous study performed by Toth *et al.* [5].

Spk-2 and *spk-3* recorded 506 unique sentences in German accented English. Note that some sentences overlapped among these two speakers. We conducted a two-fold cross validation test in which 253 sentences were used for GMM and SVM training, and the remaining sentences were used for testing.

The applied feature extraction for source EMG signals was based on *time-domain features* [4]. For any given feature $\mathbf{f}$, $\bar{\mathbf{f}}$ was its frame-based time-domain mean, $\mathbf{P_f}$ was its frame-based power, and $\mathbf{z_f}$ was its frame-based zero-crossing rate. $S(\mathbf{f}, n)$ was the stacking of adjacent frames of the feature $\mathbf{f}$ in the size of $2n+1$ ($-n$ to n) frames.

For an EMG signal with normalized mean $x[n]$, the nine-point double-averaged signal $w[n]$ was defined as

$$w[n] = \frac{1}{9}\sum_{k=-4}^{4} v[n+k], \quad \text{where} \quad v[n] = \frac{1}{9}\sum_{k=-4}^{4} x[n+k].$$

The rectified high-frequency signal was $r[n] = |x[n] - w[n]|$. In previous EMG experiments, the best results were obtained with a feature called *TD15*, which we used in this study as well:

$$\mathbf{TD15} = S(\mathbf{f2}, 15), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_r}, \bar{\mathbf{r}}].$$

Frame size and frame shift were set to 27 ms and 10 ms, respectively. In all cases, we applied LDA on the TD15 feature to generate a final feature vector with 32 coefficients. Target F_0 features were automatically extracted using a fixed-point analysis [11].

SVM^{light} [12] was used to train the SVM in which a polynomial kernel was applied. The number of mixture components of the GMM was set to 32, since experiments were expected to give reasonable results for this setting.

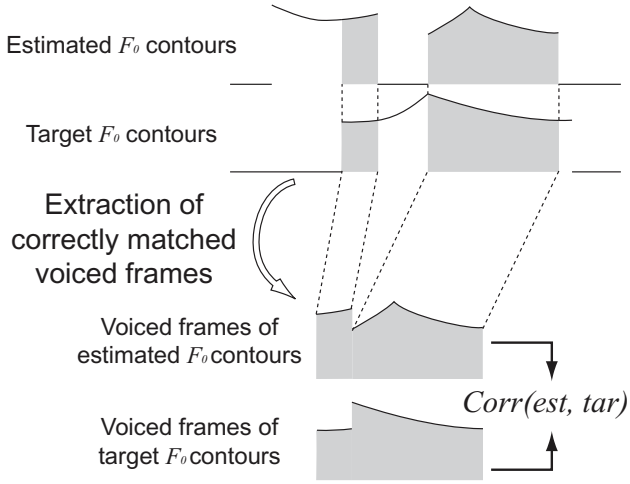

Fig. 3. Calculation of correlation coefficients between voiced frames of estimated F_0 and target F_0 contours.

The F_0 estimation accuracy was objectively evaluated by a U/V decision error rate and the correlation coefficient between frames for which the estimated and target frames both correspond to be voiced as shown in Fig. 3. Correlation coefficients were calculated utterance by utterance as follows:

$$Corr(est, tar) = \frac{Cov(est, tar)}{SD_{est}, SD_{tar}},$$

where $Cov(est, tar)$ denotes the covariance between estimated and target data, and SD_{est} and SD_{tar} denote standard deviation of estimated and target data, respectively. $Corr(est, tar)$ ranges between -1 and 1, where -1, 0, and 1 refer to negative perfect correlation, decorrelation, positive perfect correlation, respectively.

4.2. Experimental results

Table 1 shows the experimental results for all three data sets. Although the results vary, EMG-to-F_0 achieves almost 0.5 for the correlation of *spk-1* with more than 84% U/V decision accuracy. Fig. 4 shows an example of estimated and target F_0 contours of *spk-1*. The results indicate that it is possible to estimate F_0 contours from EMG data.

However, a comparison to previous studies on F_0 estimation from spectral information [9, 10] emphasizes that there is still room for improvement. While U/V discrimination gives very promising results, the correlation between estimated F_0 and target F_0 contours is rather unsatisfying and varies over speakers. F_0 estimation might reatly benefit from modifying the electrode positioning so that information of the vocal folds can be captured more accurately.

We recognize that the number of speakers is not enough to evaluate our approach. We will update our experiments by using more speakers' data. Moreover, we will also investigate and optimize experimental parameters such as (1) the

Table 1. Objective results of F_0 estimation based on correlation coefficients (Corr.) and standard deviation (S.D.). 'V → U' means rate of number of voiced frames regarded as unvoiced frames and 'U → V' means the opposite

Speaker	Corr. ± S.D.	U/V decision error rate [%]
spk-1	0.49 ± 0.19	15.8 (V → U: 6.8, U → V: 9.0)
spk-2	0.30 ± 0.23	26.8 (V → U: 10.7, U → V: 16.1)
spk-3	0.25 ± 0.18	18.0 (V → U: 9.0, U → V: 9.0)

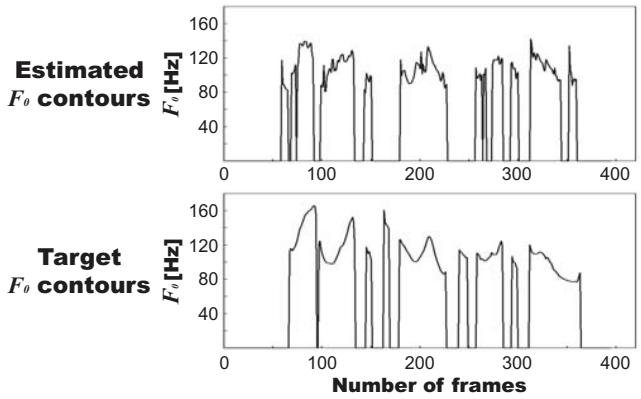

Fig. 4. Example of estimated and target F_0 of *spk-1*. Correlation of these samples is 0.67 with 87.0% U/V decision accuracy. Content is "There are lots of different ways to make a family".

relation between the F_0 estimation accuracy and the amount of training data, (2) the number of the GMM, (3) the number of compressed dimension by LDA, and (4) the kernel function used in the SVM.

5. CONCLUSION

This paper proposed EMG-to-F_0 that estimates F_0 information from EMG data only. Our approach was based on GMM-based VC method and introduced SVM to discriminate individual frames as voiced or unvoiced. As a result of experimental evaluations using data of three different speakers, one data set achieved a correlation coefficient of 0.49 between estimated and target F_0 contours with more than 84% U/V decision accuracy.

Although the achieved results are promising, F_0 estimation from EMG data only remains challenging tasks. To further investigate speaker variability, we will record EMG data of more speakers, elaborate on our algorithms, investigate different electrode positions to better capture vocal folds information, and investigate effects by changing the number of training data.

6. REFERENCES

[1] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) Recognition," IEICE Transactions on Information and Systems, vol. E89-D, no. 1, pp. 1–8, 2006.

[2] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, D.F., "Hidden Markov Model Classification of Myolectric Signals in Speech," IEEE Engineering in Medicine and Biology Society, vol. 21, no. 5, pp. 143–146, 2002.

[3] T. Schultz and M. Wand, "Modeling Coarticulation in EMG-based Continuous Speech Recognition," Speech Communication Journal, vol. 52, issue 4, pp. 341–353, 2010.

[4] S. -C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards Continuous Speech Recognition using Surface Electromyography," Proceedings of Interspeech 2006, pp. 573–576, 2006.

[5] A. R. Toth, M. Wand, and T. Schultz, "Synthesizing Speech from Electromyography using Voice Transformation Techniques," Proceedings of Interspeech 2009, pp. 652–655, 2009.

[6] V. N. Vapnik, "The Nature of Statistical Learning Theory," Springer, 1995.

[7] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," IEEE Transaction on Speech and Audio Processing (SAP), vol. 6, no. 2, pp. 131–142, 1998.

[8] A. Kain and M. W. Macon, "Spectral Voice Conversion for Text-to-speech Synthesis, " Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 285–288, 1998.

[9] T. Toda and K. Shikano, "NAM-to-Speech Conversion with Gaussian Mixture Models," Proceedings of Interspeech 2005, pp. 1957–1960, 2005.

[10] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion," Proceedings of Interspeech 2009 - Eurospeech, pp. 1431–1434, 2009.

[11] H. Kawahara, H. Katayose, A.de Cheveigné, and R.D. Patterson, "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," Proceedings of EUROSPEECH, pp. 2781–2784, 1999.

[12] SVM-Light Support Vector Machine (confirmed on 17.02.2011), http://www.cs.cornell.edu/People/tj/svm_light/