A Deep Neural Network Based Approach to Mandarin Consonant/Vowel Separation

Yen-Teh Liu, Yu Tsao, and Ronald Y. Chang
Research Center for Information Technology Innovation, Academia Sinica, Taiwan

***Abstract*—In this paper, we study the problem of Mandarin consonant/vowel separation which is an integral part of many Mandarin speech applications. We propose a deep neural network (DNN) based approach and compare its performance with the support vector machine (SVM) method. Our results demonstrate an improved separation performance yielded by the proposed method, especially on consonant identification.**

I. Introduction

Voice classification, the process of classifying human speech signals into few classes with different features, finds many applications in speech processing. For example, in hearing aid signal processing, classifying the speech signals into consonant and vowel classes and applying the frequency transposition algorithms to the consonant class can allow for improved listening for people with high-frequency hearing loss [1]. In fact, consonant/vowel separation is a fundamental and common application of voice classification due to the distinct features of the two types of speech and consequently the opportunity of customizing speech processing algorithms according to these distinct features. Existing methods for English consonant/vowel separation include threshold-based methods [2] and methods incorporating a high-pass prefilter [3] and a detector of the vowel onset point [4]. These methods have the limitations of high signal distortion and unsuitability for on-line operations, and do not translate directly to Mandarin consonant/vowel separation due to the different linguistic characteristics of the English and Mandarin Chinese languages.

In this paper, we propose a deep neural network (DNN) [5] based approach to Mandarin consonant/vowel separation. We build a classification model exploiting simple features, which proves efficiency for on-line operations and robustness against signal distortion. We compare the proposed method with the conventional support vector machine (SVM) [6] based method and report an improvement by the proposed method.

II. Methods

A. Procedures

The training speech includes 18 sentences produced by a female speaker and a male speaker both speaking in normal speed. The speech signals are sampled at 16 kHz and quantized in 16 bits. To simulate an on-line operation environment, we cut the speech signal into frames of 8 ms (128 sample points). We remove the silent parts of the speech first and then label the remaining parts as either -1 or 1 (corresponding to consonant or vowel, respectively). Mandarin syllables /ㄇ(m)/, /ㄋ(n)/, /ㄌ(l)/, /ㄖ(r)/, which are sonorants and voiced consonants, are counted as vowels for their spectrum with formants similar to those of the vowels. The testing speech is also recorded at 16 kHz sampling rate and 16 bit quantization. Then, speech features are extracted from the training and testing speech signals and fed into the SVM and DNN frameworks.

B. Feature Extraction

We adopt three features that characterize Mandarin consonant/vowel speech:

1. Energy ratio: The energy ratio between high-frequency and low-frequency components is computed, i.e.,

$$\text{ER} = |X|^2_{f>2000}/|X|^2_{f<500}, \quad f = (f_s/N) \times l,\ l = 1, 2, \ldots, N/2 \tag{1}$$

where X is the frequency-domain signal corresponding to the time-domain signal x, f is the discrete frequency bin, f_s is the sampling frequency, and N is the number of points in the discrete Fourier transform.

2. Normalized autocorrelation: The one-point autocorrelation is calculated, i.e.,

$$\text{NC} = \frac{\sum_{n=1}^{N} x(n)x(n-1)}{\sqrt{\left(\sum_{n=1}^{N} x^2(n)\right)\left(\sum_{n=1}^{N} x^2(n-1)\right)}}. \tag{2}$$

3. Zero-crossing rate: The zero-crossing rate, which measures the frequency at which a discrete signal passes through the zero, is calculated, i.e.,

$$\text{ZCR} = \sum_{n=2}^{N} f\big(x(n), x(n-1)\big),\ f(x,y) = \begin{cases} 1, & \text{if } x \cdot y < 0 \\ 0, & \text{if } x \cdot y > 0 \end{cases}. \tag{3}$$

We normalize the values of the three features to the range of $[0, 1]$ before feeding the features into SVM and DNN.

C. Support Vector Machine

SVM is a classifier that creates a hyperplane that separates the data points with as large margin to the hyperplane as possible. The data points are mapped to a higher dimension through a Kernel function and thus the SVM can find a linear separating hyperplane with maximum margin in the higher dimensional space. We implement a common Kernel function, i.e., the Gaussian radial basis function $K(x, y) = \exp(-\gamma \left\|x - y\right\|^2)$, where γ is chosen to maximize the performance in training data. The decision function is

$$d(x) = \text{sgn}\left(\sum_i w_i K(x_i, x) + b\right) \tag{4}$$

where $\text{sgn}(\cdot)$ is the sign function, x is the test vector, and x_i, b, and w are the support vector, bias, and weight obtained after the training process, respectively.

TABLE I
MANDARIN CONSONANT/VOWEL SEPARATION RESULTS (IN PERCENT CORRECT)

Method\Accuracy	Overall	Consonants	Vowels
SVM	92.44%	73.95%	99.64%
DNN (1 layer)	93.14%	76.90%	99.78%
DNN (2 layers)	93.19%	77.43%	99.64%
DNN (3 layers)	93.80%	79.51%	99.64%

D. Deep Neural Network

A neural network based approach was inspired by the human's nervous system. In our application, we adopt a DNN based approach with 1–3 hidden layers (10 nodes each layer), as additional layers beyond three offer negligible performance gain. In the training stage, we first pre-train the system with deep belief network (DBN) to initialize the weights and then update the weights during training to obtain the best model. The decision is made based on calculating the hidden units and applying the softmax function on the final hidden layer to calculate the probability of each class [7]. The jth ($j = 1, 2, \ldots, 10$) hidden unit of the ith ($i = 1, 2, 3$) layer is given by

$$h_j^i(x) = \text{sigm}\left(b_j^i + \sum_k W_{jk}^i h_k^{i-1}(x)\right) \quad (5)$$

where $\text{sigm}(a) = \frac{1}{1+e^{-a}}$ is the sigmoid function, b and W are the bias and weight of hidden layers, respectively, and $h^0(x) = x$.

III. RESULTS AND DISCUSSION

Table I summarizes the Mandarin consonant/vowel separation results for SVM and DNN. As can be seen, DNN achieves an overall 1% improvement over SVM, with predominant contribution from a better consonant identification performance (over 5%). SVM misclassifies a larger number of consonants as vowels. For DNN, the accuracy increases consistently as the number of layers increases.

The suffered performance of consonant identification in both models may be explained as follows. First, vowels generally have more conspicuous characteristics in the spectrogram as compared to consonants, and thus the considered simple features represent vowels better than consonants. Second, talker variability (different speakers speaking in different rates and in different conditions) tends to produce higher variation in consonants than vowels. Third, the fact that a consonant is followed immediately by a vowel in Mandarin also affects consonant identification.

An example of Mandarin consonant/vowel separation by the DNN method is given in Fig. 1 for the Mandarin syllable /chu/ (Tone 2). As can be seen, the affricate consonant and the vowel exhibit remarkable differences in the spectrogram; in particular, the energy of the vowel is more centered (at lower frequencies) than that of the consonant. Also, a close examination of Fig. 1 reveals that the consonant part has a higher zero-crossing rate and less periodicity. These observations justify the fitness of the adopted features.

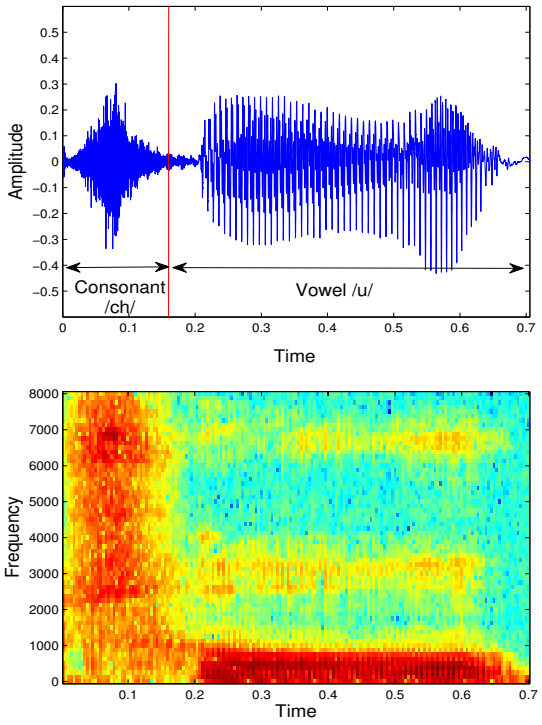

Fig. 1. An example of consonant/vowel separation by the proposed DNN method for the Mandarin syllable /chu/ (Tone 2) and its spectrogram.

IV. CONCLUSION

We have proposed a DNN-based model for Mandarin consonant/vowel separation. The proposed method achieves higher accuracy than SVM-based method and is suitable for real-time processing. The proposed method carries immense potential for other Mandarin speech processing applications.

REFERENCES

[1] J. D. Robinson, T. Baer, and B. C. Moore, "Using transposition to improve consonant discrimination and detection for listeners with severe high-frequency hearing loss," *International Journal of Audiology*, vol. 46, no. 6, pp. 293–308, 2007.

[2] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *Proc. American Society for Engineering Education (ASEE) Zone Conference*, 2008, pp. 1–7.

[3] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, May 1999.

[4] A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Improved vowel onset point detection using epoch intervals," *AEU - International Journal of Electronics and Communications*, vol. 66, no. 8, pp. 697–700, 2012.

[5] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, pp. 1–127, 2009.

[6] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, Apr. 2011.

[7] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *The Journal of Machine Learning Research*, vol. 10, pp. 1–40, 2009.