An Interactive Tool for Extracting Human Knowledge in Speech Recognition

Sayed Kamal-Aldin Ghiathi Saeed Bagheri Shouraki

Sharif University of Technology
Computer Engineering Department
Tehran, Iran

Abstract: - Conventional features for speech recognition have not been evaluated in terms of importance in human speech recognition. In this paper a method for extracting important features in an interactive process has been introduced. This method can be used as an aid for experts in an ASR expert system. It has also been shown, as an application of our method, how an expert might find out the distinguishing features between "m" and "n". As another use, it has been illustrated that how our method could be used to check the sufficiency of information in the quantized filter-bank for speech recognition.

Key-Words: - Knowledge acquisition, speech recognition, input selection, features extraction.

1 Introduction

Conventional speech recognition systems usually fail in noisy environments in which humans' intelligibility still remains to be high. This, to some extent, stems from the fact that these methods usually are based on probability theory, which is weak in representation of ignorance [1]. Still wanting to adhere to these recognizers, one can solve the problem by removing the less informative features and providing the recognizer with only important ones. Some authors have addressed this issue using an information theoretic approach [2,3,4]. In [2] the recognition phase is split into two phases:

1. Determination of relevant features with respect to each class.
2. Training recognizer based on these features (which are usually only one or two).

In this paper a new input selection method based on extraction of human knowledge is introduced. The main idea is to measure the importance of a feature with respect to its contribution to the intelligibility of speech. This is of special importance when one wants to find the discriminating feature of two similar phones (E.g. "m" and "n"). Here it has been assumed that the features to be evaluated are critical band filterbank coefficients. The problem is maintaining naturalness while manipulating speech signal in critical band filterbank space.

In section 2 our method for changing a signal to have a desired shape in filter-bank feature space is introduced. In section 3 we show a typical application of this tool to input selection for an expert system. In section 4 another application of the tool for verifying the sufficiency of features for speech recognition in a reduced feature space is discussed.

2 Proposed method

In this section we address the problem of changing speech signal to have a desired value in filter bank feature space. The result of this transformation must be a natural speech sound. It is known that to have a natural sound, the phase of original speech signal must be maintained. But amplitude can be computed at least with two ways.

2.1 Additive coefficients

One approach to finding the amplitude of each frequency component is adding up the amplitudes of triangular filters at that frequency. The problem with this approach is that it produces an artificial pitch. The reason will be clear from the following discussion. Short time spectrum of signal is computed with a fixed window size (say, with 512 samples). The following theorem shows that if window size is twice of natural frequency, odd frequencies should be zero.

Theorem 1: assume that signal x(n) is periodic with a period of N samples. If $X_N(n)$ is the DFT of x(n) with size N, we have:

$$X_{2N}(n) = \begin{cases} X_N(n/2) & if\ n\ is\ even \\ 0 & if\ n\ is\ odd \end{cases}$$

But, use of above mentioned method causes the zero value spectrums to be lost in new signal. When returning back to the time-domain, the pitch is replaced with a wrong one (say, size of window). One solution to this problem is to compute DFT over approximate fundamental frequency[1] of speaker. Another solution, which is discussed in the next section, is to maintain the relative amplitude of frequencies.

2.2 Multiplicative coefficients

Until now it has been shown that phase and relative value of frequencies must be preserved. So we must multiply each frequency with a coefficient. Assuming the center of each triangular filter to belong only to that filter, one can simply reach to the desired value of filterbank by changing centers of filters. But this will result in an artificial sound. We solved this problem by giving each filter a multiplication factor. The specific coefficient of a frequency is the weighted sum of multiplicative factors of all filters. So the problem reduces to finding appropriate multiplicative factors of filters.
Now we define some notations:

x(n): signal.
N: size of signal.
X[n]: energy of nth frequency.
fb[i]: value of ith filter.
dfb[i]: desired value of ith filter.
c[i]: dfb[i]/ fb[i].
w[i,x]: value of ith triangle at spectrum x.
α[i]: multiplicative factors of filter i. (≥ 0).

We assume that each triangular filter starts from center of its previous filter and ends at center of next filter. After finding α[i]s, the new value of each frequency is calculated using the following formula:

$$X[i]_{new} = X[i]_{old}.coef[i],$$

Where coef[i] is obtained using the following formula:

$$coef[x] = \sum_{i=1}^{\max\_filter} w[i,x].\alpha[i]$$

The problem is the determination of α[i]s such that for each filter i:

$$c[i]fb[i]_{old} = fb[i]_{new}$$

$$c[i]\log\left(1+\sum_{n=1}^{N} X[n]_{old}\, w[i,n]\right) = \log\left(1+\sum_{n=1}^{N} X[n]_{new}\, w[i,n]\right)$$

Note that we use log(1+Jx) instead of log(x) for reasons mentioned in [5]. So:

$$1+J\sum_{n=1}^{N} X[n]_{new}\, w[i,n] = \exp(c[i].fb[i]_{old})$$

$$\sum_{n=1}^{N} X[n]_{new}\, w[i,n] = \frac{\exp(c[i].fb[i]_{old})-1}{J}$$

$$\sum_{n=1}^{N} X[n]_{new}\, w[i,n] = \beta[i]\sum_{n=1}^{N} X[n]_{old}\, w[i,n]$$

Where β[i] is:

$$\beta[i] = \frac{\exp(c[i].fb[i]_{old})-1}{J\sum_{n=1}^{N} X[n]_{old}\, w[i,n]} = \frac{\exp(c[i].fb[i]_{old})-1}{\exp(fb[i]_{old})-1}$$

Now we must determine α[i] coefficients.

$$\beta[i]\sum_{n=1}^{N} X[n]_{old}\, w[i,n] = \sum_{n=1}^{N} X[n]_{new}\, w[i,n]$$

$$= \sum_{n=1}^{N} coef[i].X[n]_{old}\, w[i,n]$$

In order to compute $X[i]_{new}$, one must note that each frequency belongs to exactly two filters. Let CF[i] to be the center frequency index of filter i. So:

$$\beta[i]\sum_{n=1}^{N} X[n]_{old}\, w[i,n] = \sum_{n=1}^{N} coef[i].X[n]_{old}\, w[i,n]$$

$$= \sum_{n=1}^{CF[i]} (\alpha[i]w[i,n]+\alpha[i-1]w[i-1,n]).X[n]_{old}\, w[i,n]$$

$$+ \sum_{n=CF[i]}^{N} (\alpha[i]w[i,n]+\alpha[i+1]w[i+1,n]).X[n]_{old}\, w[i,n]$$

$$= \sum_{n=1}^{CF[i]} (\alpha[i]w[i,n]+\alpha[i-1](1-w[i,n])).X[n]_{old}\, w[i,n]$$

$$+ \sum_{n=CF[i]}^{N} (\alpha[i]w[i,n]+\alpha[i+1](1-w[i,n])).X[n]_{old}\, w[i,n]$$

[1] Pitch

We now define some new variables:

$$A^1[i] = \sum_{n=1}^{CF} X[n]_{old}\, w[i,n]$$

$$A^2[i] = \sum_{n=1}^{CF} X[n]_{old}\, w[i,n]^2$$

$$B^1[i] = \sum_{n=CF}^{N} X[n]_{old}\, w[i,n]$$

$$B^2[i] = \sum_{n=CF}^{N} X[n]_{old}\, w[i,n]^2$$

So we have:

$$\begin{aligned}\beta[i]\left(A^1[i]+B^1[i]\right) &= \alpha[i]\left(A^2[i]+B^2[i]\right)\\ &+\alpha[i-1]\left(A^1[i]-A^2[i]\right)+\alpha[i+1]\left(B^1[i]-B^2[i]\right)\end{aligned} \quad (1)$$

Note that α[0] and α[last_filter+1] are zero by definition. Since α[i] must be positive, it is difficult to find a simple formula that gives the exact values of α[i]s. So we initialize α[i]s to 1, Then we use equation (1) to re-estimate the values of α[i]s.

3 Application 1: input selection

Although the majority of current ASR systems are based on HMM and Neural Network models, some ASR systems have been developed as expert systems [6]. In these systems, it is the duty of expert to find the right feature. Current systems provide user with the ability of hearing the speech and viewing its spectrogram [6]. But, the expert cannot verify his/her guess about importance of features. This section shows how an expert could use our tool to verify his/her guess.

We, simulating an expert using our tool, try to find the discriminating features between "m" and "n" in words mad and nab. The spectrogram of these words is depicted in Figure 1.

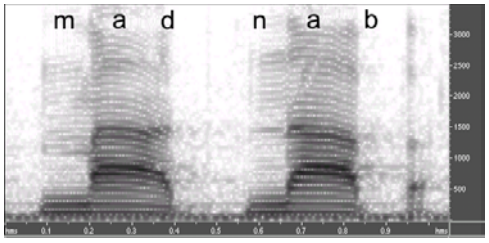

Figure 1: spectrogram of words "mad" and "nab".

As can easily be seen, there are subtle differences between spectrograms of "m" and "n". The interesting fact is that these two phonemes are distinguishable by humans even when a white noise drowns any difference between these phonemes. The spectrogram after adding white noise is depicted in Figure 2.

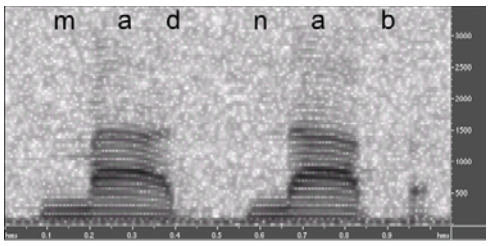

Figure 2: spectrogram of signal of Figure 1 added with a white noise.

We (still simulating the expert) guessed that the difference must be somewhere between the transition point. So we exchanged the "m" and "n" phonemes while keeping the transition point. The resulting signal was again heard as mad and nab. Our guess was true. There is some robust feature in the transition point. To find the true feature, we used our tool. Before continuing, it is necessary to see this signal in filter-bank space. Figure 3 shows the signal of Figure 1 in the filter-bank space.

We guessed three differences between "m" and "n" (highlighted with ellipses in Figure 3). Note that other differences could also be suggested (for instance, one could highlight the differences in 17^{th} filter which has a center frequency of 2000Hz and a bandwidth of 600Hz). To check our guess, we used the above algorithm to transform "nab" to "mab". The transformed signal is depicted in Figure 4. Now the signal is heard as "nad" and "mab". Another use of this approach is to put features in order based on their importance. For example, here the importance of circular area is slightly more than vertical ellipse, which in turn is more important than the horizontal ellipse.

To verify the robustness of our features, we checked them on noisy speech. As Figure 5 illustrates, our features have enough energy to be survived when noise tries to drown them.

4 Application 2: feature reduction

As another application we used our method to demonstrate the insensibility of humans to exact values of filter-banks. We quantized each filter-bank to 5 levels and re-synthesized the speech signal. Although (as a result of errors introduced when constructing the whole signal from small frames) the result was not very natural, it was completely intelligible. Figure 6 illustrates the result of quantization applied to signal of Figure 1.

5 Conclusion

In this paper a new method for manipulating speech signal in critical band filterbank was introduced. The goal is to maintain the naturalness of speech. We proposed an iterative algorithm to this problem. Our experiments shows that this method can preserve the naturalness of speech as far as requested changes are not very abrupt.

6 Future work

Although our algorithm works well, it suffers from slowness (usually when used on a large fragment of speech). One of our future works is to find a fast implementation for it. Another possible improvement to this method is to compensate for the effect of add-overlapping which is performed to reconstruct original signal from small frames.

7 Acknowledgement

This research has been supported by grants from Iran Telecommunication Research Center (ITRC).

References:

[1] G. Shafer, A Mathematical Theory Of Evidence (Princeton University Press, Princeton, NJ, 1976).

[2] H. Yang, S. van Vuuren, S. Sharma and H. Hermansky, Relevance of Time-Frequence Features for Phonetic and Speaker-Channel Classification. Speech Communication, Vol. 31(1) pp 35-50, 2001.

[3] H. Yang, S. van Vuuren, H. Hermansky, Relevancy Of Time Frequency Features for Phonetic Classification Measured by Mutual Information. Proc. ICASSP 99, Phoenix, Arizona, USA, March 1999.

[4] J. C. Segura, M. C. Benitez, A. Torre, A. J. Rubio. Feature Extraction from Time-Frequency matrices for Robust Speech Recognition. Euorospeech, Scandinavia 2001.

[5] H. Hermansky, and N. Morgan. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, 2, 578-589. 1994.

[6] HA-JIN YU, YUNG HWAN OH. Fuzzy Expert System for Continuous Speech Recognition. Expert Systems With Applications, Vol. 9. No. 1, pp. 81-89, 1995.

Figure 3: Filter-bank features of signal of Figure 1. Thickness of the line is an indicator of energy of feature.

Figure 4: The transformed signal. The second word now is heard as "mab".

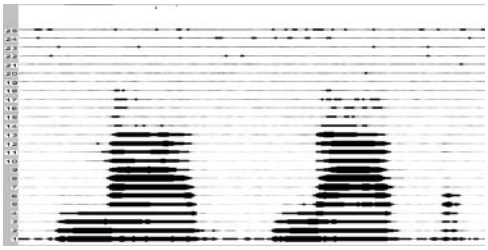

Figure 5: Filter-bank features of signal of Figure 2.

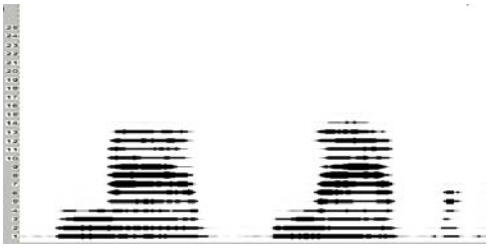

Figure 6: the result of quantization to 5 levels applied to signal of Figure 1.