ON VARIABLE SAMPLING FREQUENCIES IN SPEECH RECOGNITION

Fu-Hua Liu and Michael Picheny

Human Language Technologies Group
IBM T.J. Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598
Email: fhl@watson.ibm.com, picheny@watson.ibm.com

ABSTRACT

In this paper we describe a novel approach to address the issue of different sampling frequencies in speech recognition. In general, when a recognition task needs a different sampling frequency from that of the reference system, it is customary to re-train the system for the new sampling rate. To circumvent the tedious training process, we propose a new approach termed Sampling Rate Transformation (SRT) to perform the transformation directly on speech recognition system. By re-scaling the mel-filter design and filtering the system in spectrum domain, SRT converts the existing system to the target spectral range. New systems are obtained without using any data from the test environment. Preliminary experiments show that SRT reduces the word error rate from 29.89% to 18.17% given 11KHz test data and a 16KHz SI system. The matched system for 11KHz has an error rate of 16.17%. We also examine MLLR and MAP. The best result from MLLR is 17.92% with 4.5 hours of speech. In the speaker adaptation mode, SRT reduces the error rate from 15.48% to 9.71% given 11KHz test data and a 16KHz SA system while the matched 11KHz SA system has an error rate of 9.33%.

1. INTRODUCTION

It is well-known that speech recognition can achieve the best performance when test conditions match training conditions. In general, these conditions include acoustic environments ([e.g.[1,2]), speakers (e.g. [3,4]), application corpora (e.g. [5]), etc. In this paper we investigate an issue of sampling frequency mismatch. The frequency mismatch inevitably leads to severe performance degradation in speech recognition. In one of our experiments described below, the word error rate of a 16KHz speaker-independent (SI) system can increase from 14.74% to 29.89% when the sampling frequency of the test data switches from 16KHz to 11KHz.

Practically, when a speech recognition system is deployed, it is designed for a specific data sampling frequency. When another sampling rate is considered, it is customary to re-train the system for the new specific sampling rate. While it is straightforward to transform signals and re-train systems, this presents two major problems in many real-time applications. First, extra efforts are needed to supply training data at the new sampling frequency by either collecting new data or transforming existing training data. Second, the training process must be repeated to generate new system parameters. For systems that have undergone calibration processes such as speaker adaptation or acoustic adaptation, it is even more tedious to repeat them, let alone the complication of maintaining multiple prototypes. Therefore, an efficient methodology that can accomplish sampling frequency change without any burden of re-training becomes very desirable in many field applications.

In this paper, a novel algorithm, Sampling Rate Transformation (SRT), is proposed as a model-based solution to the issue of frequency mismatch. The SRT algorithm is to be evaluated at three different frequencies: 16KHz, 11KHz, and 8KHz. In Section 2, Sampling Rate Transformation will be thoroughly described, followed by the description of experiment systems and evaluation database along with comparative results in Section 3. Then, conclusion and summary will be presented.

2. SAMPLING RATE TRANSFORMATION

Interpolation and decimation have been used to change the sampling rate for data. However, the frequency conversion process is predominantly applied directly to waveform signals. In contrast, sampling rate transformation (SRT) performs “downsampling” or “upsampling” directly on a recognition system while leaving the sampling frequency of test data unchanged. In other words, the fundamental idea of sampling rate transformation is to convert a cepstral-based system that is designed for one particular sampling rate to one system that can be used for another sampling frequency.

Theoretically, the SRT algorithm can be applied to both downsampling and upsampling cases, like its counterpart techniques in time-domain. However, due to the fact that downsampling serves much more practical and useful purposes than upsampling for speech recognition, we mainly focus on the case of downsampling in the following discussions.

2.1. Frequency Transformation On Cepstral-Based Signal Data

Let $\{\bar{x}[t],\ t = 1, T\}$be a sequence of vectors of mel-frequency cepstral coefficient (MFCC) [6] for an utterance of length T with a sampling frequency, f_{ref}. The log-spectral representation of the signal can be written as

$$\bar{X}[t] = IDCT\{\bar{x}[t]\}, \quad t = 1, T \qquad (1)$$

where *IDCT* is the inverse discrete cosine transform (IDCT). In this case, each component in $\bar{X}[t]$ is actually the band energy from each individual mel-filter.

The downsampled version of the same signal for the new frequency, f_{new}, can be obtained by discarding all filters above the new Nyquist frequency, $f_{new}/2$, as

$$\bar{Y}[t] = \bar{W} \bullet \bar{X}[t], \quad t = 1, T \tag{2}$$

where $\bar{\boldsymbol{W}}$ is a rectangular-window filter with a cutoff frequency at $f_{new}/2$.

Note that Equation (2) is equivalent to filtering the signal in spectral domain with a rectangular window and masking it with an energy floor in the log-spectral domain for filters beyond the cutoff frequency.

Finally, the MFCC vectors for the downsampled version of signal with a new frequency, f_{new}, can be computed by discrete cosine transform (DCT) as

$$\bar{y}[t] = DCT\{\bar{Y}[t]\}, \quad t = 1, T \tag{3}$$

Furthermore, based on Equation (1), (2), (3), we can re-write the overall transformation as

$$\begin{aligned} \bar{y}[t] &= DCT\{\bar{Y}[t]\} \\ &= DCT\{\bar{W} \bullet IDCT\{\bar{x}[t]\}\} \\ &= \bar{\boldsymbol{A}} \bullet \bar{W} \bullet \bar{C} \bullet \bar{x}[t] \\ &= \bar{S} \bullet \bar{x}[t] \end{aligned} \tag{4}$$

where $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{C}}$ are matrices for DCT and IDCT, respectively. In other words, the frequency transformation can be characterized by matrix operation as shown in Equation (4).

It is noted that the downsampled cepstral vectors $\{\bar{y}[t], \ t = 1, T\}$ share the same filters as the original cepstral vectors, $\{\bar{x}[t], \ t = 1, T\}$. What this implies is that the design of mel-filters will remain the same regardless of the target sampling rate. To this end, a reference sampling frequency, usually the one of training data, is used to design the cutoff frequencies for all mel-filters. When test data sampled at another sampling frequency is to be processed, data points from FFT can get aligned to their corresponding filter designed based on the reference frequency with a linear warping.

2.2. Frequency Transformation On Cepstral-Based Gaussian Models

Equation (4) describes frequency transformation for individual static cepstral vectors and also serves as a fundamental block for frequency transformation. However, when it comes to the speech recognition models, more issues need to be addressed. Many state-of-the-art recognition systems utilize static cepstral vector and its first-order and second-order derivatives as the features. In addition, many systems seek to reduce speaker and environment variability by utilizing cepstral mean normalization (CMN) plus energy normalization. Therefore, in the following section, we derive the sampling rate transformation on a system that also employs dynamic cepstral features from time derivatives and CMN along with energy normalization of automatic gain control with respect to maximum value (AGC-max).

Let $\{\bar{\mu}_x[i], \ i = 1, M\}$ and $\{\bar{\Sigma}_x[i], \ i = 1, M\}$ represent the mean vectors and co-variance matrices of a set of *M* Gaussian distributions in a recognition system with a sampling frequency, f_{ref}.

Let $\bar{x}'[t] = [\bar{x}'_c[t], \bar{x}'_d[t], \bar{x}'_{dd}[t]]^T$ be the extended vector normalized with CMN and AGC-max. Furthermore, let AGC-max be written as $x'_0 = g \cdot x_0 + b(max(x_0))$. The static part of the extended observation $\bar{x}'[t]$ can be expressed as

$$\bar{x}'_c[t] = \bar{\mathbf{G}} \bullet \left(\bar{x}_c[t] - \frac{1}{T} \cdot \sum_t \bar{x}_c[t] + \bar{b}_{ref}\right) \tag{5}$$

where $\bar{\mathbf{G}} = \begin{bmatrix} g & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ .. & . & . & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix}$ $\bar{b}_{ref} = \begin{bmatrix} f(max(x_0), mean(x_0)) \\ . \\ 0 \end{bmatrix}$,

From Equation (4) and (5), the corresponding normalized vector for new sampling rate can now be re-written as

$$\bar{y}'_c[t] = \bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet (\bar{\mathbf{G}}^{-1} \bullet \bar{x}'_c t - \bar{b}_{ref}) + \bar{\mathbf{G}} \bullet \bar{b}_{new} \tag{6}$$

Note that $\bar{b}_{ref}$ and $\bar{b}_{new}$ are sentence-based shift vectors from AGC-max for the original and new sampling rates, respectively. Similarly, the corresponding dynamic features for new frequency can be written as

$$\begin{aligned} \bar{y}'_d[t] &= \bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet \bar{\mathbf{G}}^{-1} \bullet \bar{x}'_d[t] \\ \bar{y}'_{dd}[t] &= \bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet \bar{\mathbf{G}}^{-1} \bullet \bar{x}'_{dd}[t] \end{aligned} \tag{7}$$

For further simplicity, let sentence-based shifts, $\bar{b}_{ref}$ and $\bar{b}_{new}$, be replaced by global shifts, $\underset{\sim}{\bar{b}}_{ref}$ and $\underset{\sim}{\bar{b}}_{new}$. The static and dynamic features in the mean vectors for new frequency can be expressed as

$$
\begin{aligned}
\bar{\mu}'_{y_c}[i] &= \bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet (\bar{\mathbf{G}}^{-1} \bullet \bar{\mu}'_{x_c} i - \underset{\sim}{\bar{b}}_{ref}) + \bar{\mathbf{G}} \bullet \underset{\sim}{\bar{b}}_{new} \\
\bar{\mu}'_{y_d}[i] &= \bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet \bar{\mathbf{G}}^{-1} \bullet \bar{\mu}'_{x_d}[i] \qquad (8) \\
\bar{\mu}'_{y_{dd}}[i] &= \bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet \bar{\mathbf{G}}^{-1} \bullet \bar{\mu}'_{x_{dd}}[i]
\end{aligned}
$$

It can be easily shown that the co-variance matrix for new frequency can be described as

$$
\bar{\Sigma}'_{y}[i] = \bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet \bar{\mathbf{G}}^{-1} \bullet \bar{\Sigma}'_{x}[i] \bullet (\bar{\mathbf{G}} \bullet \bar{\mathbf{S}} \bullet \bar{\mathbf{G}}^{-1})^{T} \qquad (9)
$$

3. EXPERIMENTAL RESULTS

Experiment Setup. A state-of-the-art IBM large-vocabulary continuous speech recognition [7] is used in following experiments. Both training and test data are originally collected at 22KHz. Experiments are carried out for 3 different processing frequencies, including 16KHz, 11KHz, and 8KHz. The test data consists of 4 male and 6 female speakers. Each speaker records 61 utterances from a specific business office task. Each speaker also records 270 utterances for experiments of speaker adaptation.

3.1. Baseline SI Systems

Two speaker-independent (SI) baseline systems are established, one for 16KHz sampling frequency and one for 11KHz. Test data is also processed at the corresponding frequency. The word error rates (WER) for 16KHz and 11KHz baseline systems are 14.74% and 16.17%, respectively, as illustrated in the first two results in Table 1.

The baseline results reveal extra benefit from using wider-band information as observed in the 16KHz system. The 16KHz system is, thereafter, used as the reference system while the 16KHz frequency is referred to as the reference frequency in this paper. It is noted that the 16KHz and 11KHz systems share the same mel filters design based on 16KHz.

3.2. SRT

When the test data are processed at another frequency for other applications, a frequency mismatch occurs. While we maintain the same mel-band design used in the 16KHz system for the 11KHz data, the lack of high frequency component can still cause severe performance degradation.

Table 1 shows that this mismatch in frequency leads to a WER of 29.89%, twice the WER in reference system. When the SRT algorithm is applied to downsample the reference recognition system for 11KHz data, performance improves to 18.17% with SRT transforming only the mean vectors of Gaussian distribution. It is interesting to note that SRT does not get extra improvement by transforming both mean vectors and co-variance matrices. This is simply due to the fact that our reference system uses diagonal co-variance matrices.

System	Signal Processing for Test Data	WER (%)
16KHz - SI	**16KHz**	**14.74**
11KHz - SI	**11KHz**	**16.17**
16KHz - SI	**11KHz**	**29.89**
SRT - SI (Mean Only)	**11KHz**	**18.17**
SRT - SI (Mean & Var)	**11KHz**	**19.01**

Table 1: Frequency mismatch and SRT in SI systems

3.3. Adaptation Using MLLR and MAP

We would like to apply other model-based approaches such as MLLR [4] and MAP [3,8] for the issue of frequency mismatch. An adaptation data corpus with 1800 utterances (4.5 hours) from 84 speakers is processed at 11KHz. For the best performance, we also assume the 16KHz-processed data is available so that the much better alignment can be computed from the 16KHz reference system.

Table 2 lists the comparison of MLLR, MAP and MLLP+MAP. With the use of 4.5 hours adaptation data, MLLR generates the best result, 17.92%, which is comparable to SRT. The observation that the use of MAP does not offer extra improvement indicates the mismatched 16KHz system is not a good initial model with such relatively small amount of training data.

System	Signal Processing for Test Data	WER (%)
16KHz - SI	**16KHz**	**14.74**
11KHz - SI	**11KHz**	**16.17**
16KHz - SI	**11KHz**	**29.89**
SRT - SI (Mean Only)	**11KHz**	**18.17**
SRT - SI (Mean & Var)	**11KHz**	**19.01**
MLLR - SI	**11KHz**	**17.92**
MAP - SI	**11KHz**	**19.06**
MLLR+MAP -SI	**11KHz**	**18.59**

Table 2: Comparison of MLLR, MAP, SRT in SI system

3.4. Speaker Adaptation And Narrow Band

To study the performance of SRT in conjunction with speaker adaptation, we establish a speaker-adapted (SA) system using MLLR+MAP with 270 adaptation utterances for each speaker. For each sampling rate, we compute a SA system. Table 3 compares the performance of SRT in conjunction with speaker adaptation. It shows that SRT is also very effective in the speaker adaptation mode by reducing the WER from 15.48% to 9.71%, comparable to those from the matched systems. It is also interesting to note that the difference between 16KHz and 11KHz system is virtually flattened after speaker adaptation.

System	Signal Processing for Test Data	WER (%)
16KHz - SA	16KHz	9.28
11KHz - SA	11KHz	9.33
16KHz - SA	11KHz	15.48
SRT - SA	11KHz	9.71

Table 3: Comparison of SRT in SA systems

We also would like to examine SRT in narrow-band applications where the sampling frequency is set to 8KHz. Table 4 shows that the frequency mismatch between 8KHz and 16KHz degrades the performance to 28.93%, much worse than its 11KHz-16KHz counterpart which is 15.48%. SRT is shown to be able to remove the adverse impact from frequency mismatch with an impressive performance of 10.60%.

System	Signal Processing for Test Data	WER (%)
16KHz - SA	16KHz	9.28
8KHz - SA*	8KHz	10.83
16KHz - SA	8KHz	28.93
SRT - SA	8KHz	10.60

Table 4: Performance of SRT in 8KHz with SA.
*Note: the 8KHz-SA system was obtained using slightly different mel-filters but it is a useful benchmark.

3.5. Gender-Dependence And SRT

It is interesting to examine the correlation between frequency mismatch and speaker's gender. Table 5 shows the breakdown results based on speaker's gender in SA systems. Not surprisingly, it reveals that female speakers are likely to be more susceptible to the frequency mismatch than male speakers. We also observe similar comparisons in the SI systems.

System	Test Data	Total WER	Male WER	Female WER
16KHz	16KHz	9.28	8.48	9.82
11KHz	11KHz	9.33	8.35	9.98
16KHz	11KHz	15.48	9.92	19.19
SRT	11KHz	9.71	8.36	10.67
8KHz*	8KHz	10.83	9.02	12.04
16KHz	8KHz	28.93	22.59	33.15
SRT	8KHz	10.60	9.29	11.47

Table 5: Breakdown results of SRT based on speaker's gender in SA system, with 4 male and 6 female speakers

4. SUMMARY

In this paper, we study the issue of different sampling frequencies in speech recognition. Severe performance degradation is observed when a sampling frequency mismatch occurs. We propose a novel approach, SRT, to reduce adverse impact of mismatch. A significant advantage of SRT is that new systems are obtained without using any calibration data processed at the target frequency, Though we derive SRT from a system using dynamic cepstral features with CMN and AGC-max, this transformation can be easily extended to other cepstral-based systems.

In our study, SRT achieves a WER of 18.17% for 11KHz SI system and 9.7% for 11KHz SA system, comparable to 16.17% for 11KHz SI system and 9.33% for 11KHz SA system. In comparison, MLLR achieves a WER of 17.92% with the use of 4.5 hours of adaptation data. In narrow band applications, SRT can reduce the WER to 10.60% from 28.93% given a 16KHz SA system to be used with 8KHz test data. A benchmark 8KHz system achieves 10.83% for the same data.

REFERENCES

1. A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.

2. F.H. Liu, et al, "Environment Normalization for Robust Speech Recognition Using Direct Cepstral Comparison", *ICASSP-94*, pp. II-61 - II-64, April 1994.

3. C.H. Lee, C.H. Lin, B.H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", *IEEE Trans. on Signal Processing*, vol.39, no. 4, April 1991.

4. C.J. Leggetter and P.C Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech And Language*, vol. 9, pp. 171-185, 1995.

5. F.H. Liu, M.D. Monkowski, M. Novak, M. Padmanabhan, M.A. Picheny, and P.S. Rao, "Performance of the IBM LVCSR System on the Switchboard Corpus", *Proceedings of Speech Research Symposium*, P.P. 189, June, 1995.

6. S.B. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. on ASSP*, vol.28, pages 357-366, 1980.

7. L.R. Bahl, et al, "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task", *ICASSP-95*, 1995.

8. J.L Gauvain and C.H. Lee, "Maximum-A-Posteriori Estimation for Multivariate Gaussian Observations and Markov Chains", *IEEE Trans. on Speech And Audio Processing*, vol.2, no. 2, April 1994.