Performance adjustment of Speech rate in Automatic Speech Recognition

E. Chandra[1]

ABSTRACT

Rate of speech (ROS) has a greater influence in both spectral features and word pronunciations that affect automatic speech recognition (ASR) systems. To deal with these ROS effects, the research proposes to use parallel, rate-specific, acoustic models: one for fast speech, the other for slow speech. Rate switching is permitted at word boundaries, to allow modeling within sentence speech rate variation, which is common in conversational speech. The Input Signal Processing configures the front end for ROS to identify the fast and slow speech and customize the application to handle both the situation well.

Keywords : Hyper-Threading, Multithreading, Performance, Threading

1. INTRODUCTION

Rate of speech (ROS) is an important factor that affects the performance of a transcription system,[1]. Possible reasons are that some features commonly used in recognition systems are duration related and clearly influenced by speech rate, such as delta and delta features, and that some pronunciation phenomena such as co-articulation and reduction are also speech rate related [7]. Thus, using rate-dependent acoustic models seems to be a promising way to improve robustness against speech rate variation.

[1]Department of Computer Applications, D.J.Academy for Managerial Excellence, Coimbatore-32.

This paper proposes a new approach of word level rate-dependent acoustic modeling. Under this approach, each typical word is given as fast version pronunciation and a slow-version pronunciation, each consisting of rate-specific phonemes (elementary probabilistic models of basic linguistic units). The recognizer is allowed to select the fast or the slow pronunciation for each word automatically during search, based on the maximum likelihood criterion.

To train the rate specific phoneme models, the input signal processing model used as a duration-based ROS measure to partition the training data into rate-specific categories.

2. ROS MEASURE

Two methods are typically used to estimate ROS of an input utterance. One is based on phoneme durations. The Input Signal Processing model (ISP) [5] is proposed with the confidence that the input signal speed can be fine tuned for better realization by the Sphinx Speech Engine. In this research work, the ISP model configures the Result.getFrameNumber() function in the Result class of Sphinx software by multiplying with the window Shift function which is 10 milliseconds by default, to get the length of the result. A standard reference is set to identify the silence as well as the speech based on several repeated experiments. The work also lowered the speech recogniser classifier 'threshold' property in the config files to make the input signal to be loud enough for the Sphinx engine [9] to recognize. By using repeated experiment results the proposed model arrived at a set of standard for optimal speed for the input signal and configured the

module (ISP). When the utterance transcription is known, this duration based method can provide robust ROS estimation [1]; however, when the transcription is unknown, the hypothesis from a prior recognition run was referred, whose quality can be not very precise. The second method involves estimating OS directly from the waveform or acoustic features of the input utterance [3][5]. To achieve robust ROS estimation, th computation is often based on a data window witl sufficient length. Under the proposed approach, to trai the rate specific models the training data is partitione into rate-specific categories at the word level, an therefore need the ROS for each word to be estimate locally. The output of this process should give each wor in the training transcription a rate class label. As the firs step to ROS modeling, the research decided to use onl two ROS classes: fast or slow. Since only there is a need to compute ROS for the training data that have transcriptions, it is relatively straightforward to obtain the duration of each word and its component phones by computing forced Viterbi alignments, and then applying duration-based ROS estimation methods. Fig. 1 illustrates the duration distributions of 46 categories of monophones estimated from the training corpus. It illustrates the duration distribution across different phone types differ substantially. The approach, use a relative ROS measure, RW(D) , defined as percentile of word's ROS distribution:

$$R_W(D) = P_W(d > D) = 1 - \sum_{d=0}^{D} P_W(d) \qquad (1)$$

where W is a given word, D is the duration of W, and PW(d) is the probability of that type of word having duration d. RW(D) is the probability of W having a duration longer than D=2E The measure RW(D) always falls within the range [0,1], and can be compared between different word categories. However in practice, PW(d) is hard to estimate directly due to the data sparseness problem. To address this it was assumed that in a word the duration distributions of its component subword units, such as phonemes, are independent of each other[6]. Thus, a word's duration distribution equals the convolution of its component subword units distributions, which are easier to estimate from training data.

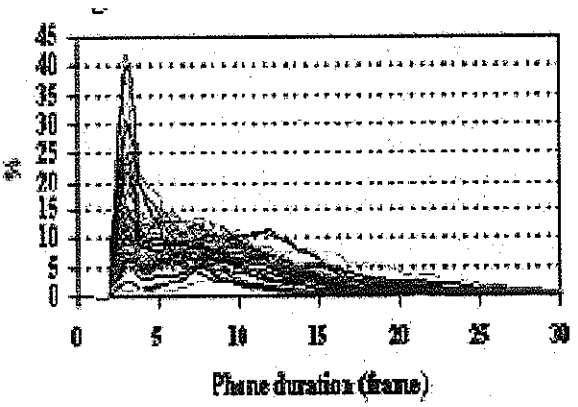

Figure 1: Duration Distribution Of Different Phone Types

The proposed model used this measure to calculate the ROS for all the words in the training data, and found that 80% of sentences with five or more words have at least one word belonging to the fastest one third and one word belonging to the slowest one third of all the words. This suggests that in conversational speech, speech rate is usually not uniform within a sentence[11]. In fact, the measure defined in Eq. (1) can also be applied to subword units, thus allowing us to calculate the ROS of phonemes. This measure studied the phoneme's ROS variation within words vs. within sentences. Fig. 2 shows a histogram of the standard deviation of the phoneme's ROS within words and within sentences for all training data, suggesting that the word is a better unit than the sentence for ROS modeling, because the average phoneme-level ROS variation within a word is significantly smaller than within a sentence.

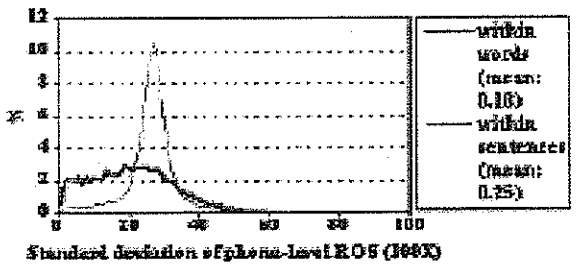

Fig 2 : Histogram Of Standard Deviate On Of Phoneme-level ROS: Within Words Vs. Within Sentences.

3. Rate-Dependent Acoustic Modeling

In this proposed method, each word is given parallel fast- and slow-version pronunciations in the recognition lexicon. Both fast- and slow version pronunciations are initialized from the original rate-independent version, with the simple replacement of rate-independent phonemes by rate-specific phonemes. For example, the original rate-independent pronunciation of "WORD" is / w er d/. Consequently the fast-version pronunciation is / wf erf df/ and the slow-version /ws ers ds/, consisting of fast and slow phonemes, respectively. The recognizer in this case the normal forced alignment modes [5] automatically find the best pronunciations that maximize the likelihood score during the search, and thus avoid the need for ROS estimation before recognition. In addition, the search algorithm, Breadth first searching [5] is allowed to select pronunciations of different rates across word boundaries and thus can cope with the problem of speech rate variation within a sentence.

A. Acoustic Training

The experiment is based on Sphinx4 system, which uses continuous-density genomic hidden Markov models (HMMs) [4] the application is configured with only the first-pass recognizer based on gender-dependent non-crossword genonic HMMs (1730 genones with 64 Gaussians each for male, 1458 genones for female) and a bigram grammar with a 33,275- word vocabulary. The recognition lexicon was derived from the CMU V0.4 lexicon with stress information stripped. The recognizer used a two-pass (forward pass and backward pass) Viterbi beam search algorithm; in the first pass a lexical tree was used in the grammar back off node to speed up search. The report results from the backward pass. The features used were 9 cepstral coefficients (C1- C8 plus C0) with their first- and second-order derivatives in 10ms time frames. The research first calculated the ROS for all the words in the training corpus based on the above-mentioned measure, sorted these words accordingly, and then split them into two categories: fast and slow. The ROS threshold for splitting was selected to achieve equal amounts of training data for the fast and the slow speech. The training transcriptions were labeled accordingly. The research then prepared a special training lexicon: words with a fast label were given the fast-version pronunciation, and words with a slow label the slow-version pronunciation[13][15].

In this way, the proposed model was able to train the fast and slow models simultaneously. The research used genonic training tools to do standard MLE (Maximum Likelihood Estimation) gender dependent training [4] and obtained rate dependent models with 3233 genones for male and 2501 genones for female. The genone clustering for rate-dependent models used the same information loss threshold as the training of rate-independent models. The proposed model compared the rate-dependent acoustic model with the rate-independent acoustic model (baseline system) on a development data set, which is a subset of the Sphinx4 data set, consisting of 1143 sentences from 20 speakers (9 male, 11 female). Table 1 shows the word error rate (WER) for both models.

Table 1 : Wer Comparison Between The Baseline System With Rate-independent Model And The System With Rate-dependent Model On The Development Data Set

	Male	Female	All
rate-independent model	55.6	64.5	58.9
rate-dependent model from training	51.7	61.7	58.9

Rate-dependent modeling brings an absolute WER reduction of 1.9%, which is statistically significant. To eliminate the possible effect of different numbers of parameters, there was an adjustment in the information loss threshold for genone clustering to obtain another rate independent model that had a number of parameters similar to that of the rate-dependent model.

B. Adaptation Vs. Standard Training

In our previous work based on the ISP [5], instead of using the training method proposed here, the proposed model trained the rate-dependent model based on Speech rate as the major influencer. However, in the current task of speech transcription the proposed work more training data, and the research use a different strategy to partition the data into two classes instead of three, yielding more training data for each rate class.. Thus, the proposed model was able to train the rate-dependent models robustly with standard training methods. For comparison the proposed model tested the Bayesian adaptation approach [2] on the current training set. Similar to [2], even though the research has used separate rate-specific models for each triphone, the research has not created separate copies of the genones, but let the fast and slow models for a given triphone share the same genone. In this way, the same number of Gaussians was used for the rate-dependent model as for the rate-independent model. Table 2 shows the results on the same development data set used in the previous section. This approach brings an advantage of 1.0% over the baseline, less than the standard training scheme[14]. This indicates that the difference between fast and slow speech in the acoustic space is significant, and that standard training might be better than the previous adaptation scheme to capture this difference. These differences might explain why the adaptation scheme did not achieve as much improvement as the standard training.

Table 2 : Wer Comparison Between The Baseline System With Rate-independent Model And The System With Rate-dependent Model From Adaptation On The Development Set.

	Male	Female	All
rate-independent model	55.3	63.4	59.8
rate-dependent model from training	54.0	62.6	58.8

Table 3 : Minimal Pair Comparison Based On An Improved Baseline System Using A Wider Front End And Vtl Normalization On The Development Set

	Male	Female	All
WER of baseline system	44.3	53.3	47.3
WER of rate-dependent system	43.6	53.0	46.8

4. Experimenal Setup

The proposed research work used Sphinx4[10] in a windows 2000 platform with the following configurations, Viterbi algorithm based HMM[16] and Flat structured viterbi search for decoding and continuous density acoustic mode and ASCII simple N-gram model and uses Breadth first search and the baseline system had been enhanced substantially. Below the research supply some minimal pair experiments based on different baseline systems during the development process. The baseline system in Table 3 used a wider-band front end (with 13 cepstral coefficients instead of 9), and vocal tract

length (VTL) normalization [5] during training. The success from introducing word-level rate dependency is still 1.9%, over a baseline that was itself improved by 5.0%

Table 4 : Minimal Pair Comparison Based On A Multiword-augmented Baseline System On The Development Set

	Male	Female	All
WER of baseline system	50.6	57.9	54.6
WER of rate-dependent system	49.2	55.6	52.7

Another major addition to the evaluation system was the introduction of multiword pronunciations. Here a multiword is a high frequency word bigram or trigram, such as "a lot of", that is handled as a single word in the vocabulary. By using handcrafted phonetic pronunciations describing various kinds of pronunciation reduction phenomena for these multiwords, the work achieved better modeling of crossword coarticulation. In Sphinx4 system 1200 multiwords were introduced. Experiments showed that the multiword pronunciation modeling brought about a 4.0% absolute win on top of the improved baseline system in Table 3, [5][8].

The possible reasons for the diminished effectiveness of ROS modeling may lie in the following aspects. First, each multiword is given multiple parallel pronunciations reflecting both full and reduced forms. This by itself models fast and slow speech variants to some extent. Words, the work fail to model the rate variation occurring within the multiwords, and thus may influence the quality of the rate-dependent acoustic models. Third, due to our current implementation, the introduction of multiwords made the search much more expensive than before; rate-dependent modeling on top of the multiword dictionary made this problem even worse, and may have produced a loss in performance due to search pruning[12].

Based on the above analysis, another scheme was tested: instead of treating multiwords as ordinary words the research trained them with multiword-specific phoneme units, that is, using separate phonetic models to describe the multiwords. Similar to the original approach, trained three classes of phoneme models simultaneously: fast models for ordinary words, slow models for ordinary words, and a separate set of phone models trained only on the multiword data. With this approach, the research improved the WER reduction to 0.7%, as shown in Table5.

Table 5 : Minimal Pair Comparison On The Development Set Between The Multiword Augmented Baseline System And The Rate Dependent System With Multiword-specific Phone Models

	Male	Female	All
WER of baseline system	44.3	53.3	49.3
WER of rate-dependent system	43.6	52.6	48.6

5. Conclusion

Thus the paper proposed a rate-dependent acoustic modeling scheme, which is able to model within-sentence speech rate variation, and does not rely on ROS estimation prior to recognition. Experiments shows that this method results in a 1.9% (absolute) word error rate reduction on sphinx4 speech transcription test set.

References

[1] N. Mirghafori, E. Fosler and N. Morgan, *"Towards Robust-ness to Fast Speech in ASR."*, Proc. ICASSP96, Vol 1, PP. 335-338, 1996.

[2] J. Zheng, H. Franco, F. Weng, A. Sankar and H. Bratt., *"Word-level Rate-of-Speech Modeling Using Rate-Specific Phones and Pronunciations."*, Proc. ICASSP00, Vol 3, PP. 1775-1778, 2000.

[3] N. Morgan and E. Fosler, *"Combining Multiple Estimators of Speaking rate."*, Proc. ICASSP98, Vol 2, PP. 729-732, 1995.

[4] V. Digalakis, P. Monaco and H. Murveit, *"Genones, Generalized Mixture Tying in Continuous Hidden Markov Model-based Speech Recognizers,"* IEEE TSAP, Vol 4. No 4. PP. 2E 281-289, 1996.

[5] Dr.E.Ramaraj,E.Chandra, *"Influence of Acoustics in Speech Recognition for Oriental Language"*, Published in IJCPOL, International Journal of Computer Processing of Oriental Language, C073, December 2005,World Scientific Publications, Singapore,Vol:18, No:4, PP 265-280.

[6] Ephraim.Y and N. Merhav, *"Hidden Markov processes"*, IEEE Trans. Inform Theory, Vol: 48, PP. 1518-1569, June 2002.

[7] Huang X, Lee K, Hon H. and Hwang. M, *"Improved Acoustic Modeling for the Sphinx Speech Recognition System"*, IEEE International Conference on Acoustics, Speech, and Signal Processing. Toronto, Ontario, CANADA, 1991, PP. 345-348.

[8] Jelinek.F, *"Continuous speech recognition by statistical methods,"* Proceedings of the IEEE, Vol. 64, PP. 532 -556, 1976.

[9] Lee. K, *"Automatic Speech Recognition: The Development of the Sphinx System"*, Kluwer Academic Publishers, Boston, 1989.

[10] Lenzo K.A, *"CMU Sphinx: Open Source Speech Recognition"*, [Online]. Available: http:// www.speech.cs.cmu.edu/speech/sphinx/.2002.

[11] Lippmann. R.P, *"Speech recognition by machines and humans"*, Speech Communication, Vol 22 No.1 PP:1-16, July 1997.

[12] Martinez. F, Tapias. D and Alvarez. J, *"Towards Speech Rate Independence in Large Vocabulary Continuous Speech Recognition"*, IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle. PP. 725-728, May 1998.

[13] O'Shaughnessy. D, *"Speech Communication-Human and Machine"*, Reading MA: Addison-Wesley, 1987.

[14] Pisoni, David. B, Remez, Robert. E. (Eds.) *"The handbook of speech perception"*, Oxford: Blackwell. ISBN 0-631-22927-2, 2004.

[15] Stevens, Kenneth, *"Acoustic Phonetics"*, The MIT Press, New Ed edition, ISBN 0-262-69250-3, 2004.

[16] Lee. K, Hon. H, Hwang. M. and Huang X, *"Speech Recognition Using Hidden Markov Models: A CMU Perspective"*, Speech Communications, Vol.9, 1990.

Authors Biography

Dr. E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University ,Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007 . At present she is working as a Head and Assistant professor at Department of Computer Applications in D. J. Academy for Managerial Excellence, Coimbatore. She has published more than 20 research papers in National, International journals and conferences. She is guiding 4 Ph.D., Scholars and guided for more than 30 M.Phil., research scholars. Her research interest lies in the area of Data Mining, Artificial intelligence, neural networks, speech recognition systems and fuzzy logics. She is an active member of CSI, Society of Statistics and Computer Applications.