SPEAKER NORMALIZATION ON CONVERSATIONAL TELEPHONE SPEECH

Steven Wegmann, Don McAllaster, Jeremy Orloff and Barbara Peskin

Dragon Systems, Inc.
320 Nevada Street, Newton, MA 02160, USA

ABSTRACT

This paper reports on a simplified system for determining vocal tract normalization. Such normalization has led to significant gains in recognition accuracy by reducing variability among speakers and allowing the pooling of training data and the construction of sharper models. But standard methods for determining the warp scale have been extremely cumbersome, generally requiring multiple recognition passes. We present a new system for warp scale selection which uses a simple generic voiced speech model to rapidly select appropriate frequency scales. The selection is sufficiently streamlined that it can moved completely into the front-end processing. Using this system on a standard test of the Switchboard Corpus, we have achieved relative reductions in word error rates of 12% over unnormalized gender-independent models and 6% over our best unnormalized gender-dependent models.

1. INTRODUCTION

Since the 1994 CAIP workshop, a number of groups ([1], [2], and [3]) have reported on methods of vocal tract length normalization. This is an old idea ([4]) which was revived by [3] in the following form. If we assume a uniform tube with length L for the model of the vocal tract, then each formant frequency will be proportional to $1/L$. The idea is to rescale (or warp) the frequency axis during the signal processing step, to make speech from all speakers appear as if it was produced by a vocal tract of a single standard length. Dragon Systems, Inc., used a method similar to [3] in its 1994 Wall Street Journal (WSJ) evaluation system [1]. We found, in that work, that speaker normalization allowed us to pool all of the training data to build a single set of gender independent models, which outperformed our gender dependent models by an estimated 12%. Unfortunately, the method used to choose the warp factor was somewhat cumbersome, requiring a coarse recognition pass and then rescoring the transcription at each of the warp scales. In this paper a new method for selecting the scaling factors will be described, which was developed on the SWITCHBOARD (SWB) corpus. It gives the performance benefits described above but at much lower computational cost. Furthermore, since full recognition passes are no longer required, the warp selection process can be moved completely into the front end, resulting in a great simplification to the overall recognition system.

2. THE FRONT END

Dragon's signal processing front end ([1]) computes 44 channel normalized features every 10 milliseconds: 1 overall energy term, 7 log spectral magnitudes, 12 mel cepstral terms, 12 mel cepstral differences, and 12 mel cepstral second differences. An IMELDA transform, derived from linear discriminant analysis (see [5]), is used to reduce these to 24 features.

The frequency warping is done using a piecewise linear transformation of the frequency axis that has fixed points at 0 kHz and at the Nyquist frequency (4 kHz when processing 8 kHz SWB data). We choose a point A below the Nyquist frequency (A is chosen to be 3.5 kHz when working with SWB data). The map from 0 kHz to A is a line through the fixed point at the origin with a slope chosen from the range 0.88 to 1.2. The map from A to the Nyquist frequency is a line that intersects the previous line at A and ends at the fixed point at the Nyquist frequency. We use 10 such maps, which we will refer to as warp scales. Four of the warp scales have slopes less than one, which compress the frequency axis and are commonly selected by female speakers, five of the warp scales have slopes bigger than one, which expand the frequency axis and are commonly selected by males, and there is a warp scale with slope one. These transformations are applied in the following manner after the FFT is computed: Let X and Y denote the original and transformed frequency axes, and let $f : X \rightarrow Y$ be a warp scale. Given y in Y, there is a unique x in X with $y = f(x)$. Since, in general, x will not be one of the frequencies where we computed the FFT, we estimate the value of the FFT at x, $FFT(x)$, by linearly interpolating the values of the FFT at the two frequencies nearest x where the FFT was computed. We then set $FFT(y) = FFT(x)$, and use these new values of the FFT when we extract

0-7803-3192-3/96 $5.00©1996 IEEE

the spectral and cepstral features.

3. THE NEW WARP SCALE SELECTION METHOD

A model of generic voiced speech is used to select the warp scale for each speaker. This model is a single probability distribution, consisting of a mixture of 256 multivariate Gaussians with diagonal covariances. To select the best warp scale, we signal process a speaker's data using each of the 10 warp scales, then score the resulting feature files against this generic voiced speech model. Although we still do the signal processing at each warp scale, the scoring is now very fast. Eventually the scoring will be integrated into our signal processing front-end, eliminating the need to process the complete message at each warp scale.

We use an iterative process to construct the generic voiced speech model. A small set of acoustic data, which is gender balanced, is used for training the warp model. (In the experiments conducted to date this set has consisted of about three hours of data taken from 7 (WSJ) to 40 (SWB) speakers.) To seed the iteration, a generic voiced speech model is trained from the voiced frames of the unwarped data. We use a "harmonicity" feature to decide which frames are voiced. This harmonicity feature is derived by methods similar to traditional cepstrum peak analysis for making voicing decisions, but it is more robust to noise. To compute harmonicity, a limited order autocorrelation analysis is applied to the lowest 1.5 kHz of the power spectrum before applying the cosine transform and determining the peak value. For more details, see [6].

The EM and LBG algorithms are used to construct the mixture model from the selected voiced speech frames. The iteration step consists of using the previous generic voiced speech model to determine the best warp scale for each of the training speakers. This warped data is then used to train a new model. The best warp scale for a speaker's training data is determined by signal processing the speech at each of the 10 warp scales (this only needs to be done once during the training process), and then scoring only the voiced frames with the generic voiced speech model: the warp scale that scores best is selected. The process is iterated until the average score per speaker against the generic speech model is minimized. This usually occurs at about the fourth iteration.

For corpora such as SWB, which is 4-wire recordings, it may be necessary to chop up the speech stream before warp selection in order to remove crosstalk, nonspeech events, and long intervals of silence. This reduces the single speech stream to a sequence of speech-rich "chopped" segments.

The warp scales that are selected generally follow a bimodal distribution, roughly corresponding to the genders, although there is a fair amount of spread around the two modes. This warp selection pattern demonstrates the power of speaker normalization: an appropriate choice of warp scales reduces differences not only between the genders but within each gender class, allowing data from all speakers to be merged into a single set of well-focused models.

4. EXPERIMENTAL RESULTS

To evaluate this speaker normalization scheme on the SWB corpus, we built gender independent (GI) and gender dependent (GD) models from speaker normalized (SN) and unnormalized (UN) acoustic training data. The data used to train the generic voiced speech model was a three-hour subset of the SWB training set, consisting of two message halves from each of 40 speakers (20 males and 20 females). Using this generic speech model, appropriate warp scales were selected for each speaker in the full training set.

The full training set consisted of 6-10 message halves from each of 80 men and 80 women in the SWB corpus, which encompassed around 65 hours of speech after chopping out silence, crosstalk, and other nonspeech events. All of the acoustic models are decision tree based mixtures of up to 16 Gaussian components with diagonal covariances (see [7]). The GD models had about 9000 output distributions for each gender's models, and the GI models had about 14,000 output distributions. For the SN models, one warp scale was used for all of a given speaker's message halves. This warp scale was determined by picking one of the speaker's message halves, signal processing it at each of the 10 warp scales, and using the generic voiced speech model to pick the best scoring warp scale (scoring only the voiced frames).

Not only is the scoring that determines the warp scales fast, it requires very little speech to make a decision: for the SWB message halves, this method selected the warp scale after seeing an average of only 52 seconds of the chopped speech, much less than the over 2 minutes of speech contained in an average processed message half.

The models were tested on the so-called "CAIP set" of 20 SWB message halves. A trigram language model and a roughly 10k vocabulary were used for recognition. When the speaker normalized acoustic models were used, the test data was also speaker normalized. The table below displays the word error rates (averaged over the number of words) for the recognition runs.

	unnormalized	normalized
gender dept	46.9%	45.0%
gender indt	49.8%	43.9%

Table 1. Word error rates with and without speaker normalization.

The gender independent, speaker normalized models give a 6% improvement relative to our previous best models, the gender dependent unnormalized models, and a full 12% relative gain over gender independent unnormalized models. As noted above, while the distribution of scales that the speakers selected is bimodal, the speakers spread themselves out over the scales. This spread of warp scales within the genders reduced the variabilty within the gender-specific training data, which explains why the SN GD models were better than the UN GD models. The SN GI models were better still, because they made effective use of the entire pool of training data, twice as much data as each GD model was trained from.

5. FUTURE WORK

Even though this new algorithm works well on SWB and WSJ data, where there is only one speaker per speech stream, significant work is still required before it can be used effectively on speech that contains multiple speakers in a speech stream, such as that in the CALLHOME and Marketplace corpora. In order for this speaker normalization algorithm to be effective in such environments, some form of speaker change detection will probably be necessary. Ideally, speaker clustering would also be done, so that the warp scale selection would use the biggest possible chunks.

In preliminary experiments on CALLHOME messages containing only one speaker in a speech stream, SN GI models outperformed UN GI models by less than 5% (relative), much less than one would expect given our results on WSJ and SWB. The reasons for this disparity still need to be understood and further refinements may need to be added to the normalization system.

In our earliest experiments with SWB data, before we restricted the warp selection to only voiced speech frames, SN GI models gave only modest improvements over UN GI models and UN GD models outperformed SN GI models, a symptom that the warp scale selection process was not working properly. We believe that the uneven quality of "silence" in SWB was part of the problem — a primary motivation for our move to scoring only voiced frames when determining the warp scale. Similarly, we may need to do something more sophisticated when attempting to select warp scales for corpora such as Marketplace, since some amount of music and noise may be slipping through our voiced speech detection.

REFERENCES

[1] Roth, R., Gillick, L., Orloff, J., Scattone, F., Gao, G., Wegmann, S., and Baker, J., "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer", *Proc. of the Spoken Language Systems Technology Workshop*, Austin, TX, January 1995, pp 116-120.

[2] Eide, E., and Gish, H., "A Parametric Approach to Vocal-Tract-Normalization", *Proc. of the 15th Annual Speech Research Symposium*, CLSP, Johns Hopkins University, Baltimore, MD, June 1995, pp 161-167.

[3] Kamm, T, Andreou, G., and Cohen, J., "Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability", *Proc. of the 15th Annual Speech Research Symposium*, CLSP, Johns Hopkins University, Baltimore, MD, June 1995, pp 175-178.

[4] Bamberg, P., "Vocal Tract Normalization", *Verbex Internal Technical Report*, 1981.

[5] Hunt, M. J., *et al.*, "An Investigation of PLP and IMELDA Acoustic Representations and of their potential for Combination", *ICASSP-91*, Toronto, Canada, May 1991, pp 881-884.

[6] Hunt, M. J., "A Robust Method of Detecting the Presence of Voiced Speech", *Proc. 15th International Congress on Acoustics*, Trondheim, Norway, June 1995.

[7] Peskin, B., Connolly, S., Gillick, L., Lowe, S., McAllaster, D., van Mulbregt, P., Nagesha, V., and Wegmann, S., "Improvements in Switchboard Recognition and Topic Identification", *ICASSP-96*, Atlanta, GA, May 1996.