VOICING FEATURE INTEGRATION IN SRI'S DECIPHER LVCSR SYSTEM

Martin Graciarena, Horacio Franco, Jing Zheng, Dimitra Vergyri, Andreas Stolcke

Speech Technology and Research Laboratory,
SRI International, USA.
{martin,hef,zj,dverg,stolcke}@speech.sri.com

ABSTRACT

We augment the Mel cepstral (MFCC) feature representation with voicing features from an independent front end. The voicing feature front end parameters are optimized for recognition accuracy. The voicing features computed are the normalized autocorrelation peak and a newly proposed entropy of the high-order cepstrum. We explored several alternatives to integrate the voicing features into SRI's DECIPHER system. Promising early results were obtained in a simple system concatenating the voicing features with MFCC features and optimizing the voicing feature window duration. Best results overall came from a more complex system combining a multi-frame voicing feature window with the MFCC plus third differential features using linear discriminant analysis and optimizing the number of voicing feature frames. The best integration approach from the single-pass system experiments was implemented in a multi-pass system for large vocabulary testing on the Switchboard database. An average WER reduction of 2% relative was obtained on the NIST Hub-5 dev2001 and eval2002 databases.

1. INTRODUCTION

Most state-of-the-art speech recognition systems use Mel cepstral (MFCC) features as input features. The typical parameters of a MFCC feature front end (e.g., 25 ms window, 10 ms frame rate) are a compromise solution chosen to represent all phones. Because of the constraint of a fixed single set of parameters, the MFCC features may fail to capture many discriminative cues that have longer or shorter time spans.

We propose to augment the Mel cepstral feature representation with phonetic features computed with independent front ends. The parameters from each front end specific to a phonetic feature can be optimized to improve accuracy of a recognition system using the augmented feature. The front end parameters that can be optimized include the window duration and type, the fast Fourier transform (FFT) size, and others. A single frame rate must be set for all the front ends to be able to concatenate the features. The idea is that robust broad-class phonetic features could provide "anchor points" in acoustic phonetic modeling. While this is a general framework for multiple phonetic features, our initial approach explores the use of voicing features.

Previous work in incorporating voicing features into speech recognition systems includes the following. In [1] experiments with an autocorrelation-based voicing measure are reported. Standard MFCC plus delta and delta-delta features were augmented by the voicing measure as well as its second and third derivatives. In [2] the fundamental frequency and a voicing measure were combined with MFCC features using linear discriminant analysis (LDA). In [3] three alternative voicing features were tested in combination with MFCC features using LDA.

Our approach differs from previous work in a number of ways. First, we use a novel voicing feature, the entropy of the high-order cepstrum, in combination with a more standard autocorrelation-based voicing estimator. Second, we perform extensive optimization of the voicing front end for the recognition task. Third, the improvement we obtain exceed the one found in a recent comparable study, which had obtained about 1% relative reduction in word error rate (WER) on English conversational speech [2]. Finally we show that the improvement is preserved across the different stages of a multi-pass evaluation-type system.

We explored several alternatives for integrating the voicing features into SRI's DECIPHER large vocabulary recognition system [4]. In an initial system, we first concatenated the voicing features to the standard Mel cepstral features and optimized the window duration for the voicing feature front end. We extended the window duration to explore whether the voicing activity was captured more reliably with a longer time span. Then we explored integrating the voicing feature in a more complex recognition system with Mel cepstral features augmented with third differential features, using heteroscedastic linear discriminant analysis (HLDA) [7]. Different integration approaches were evaluated in this system, revealing the usefulness of a multi-frame window. Finally, a multi-pass system which includes adaptation, N-best generation, rescoring with more complex language models and confusion-network decoding, was built to evaluate the voicing features. All the training, testing, and optimization were done using conversational telephone speech (CTS) databases.

The paper is organized as follows. Section 2 presents the voicing feature extraction algorithms. Section 3 presents voicing feature integration into SRI's DECIPHER recognition system. Section 4 presents a multi-pass evaluation-type system. Section 5 presents our conclusions.

2. VOICING FEATURE EXTRACTION ALGORITHMS

The first voicing feature used in this paper is the normalized peak autocorrelation, similar to the one described in [3]. The

second voicing feature used is a newly defined entropy of the high-order cepstrum. For the time windowed signal $x(t)$ of duration T the high order cepstrum is defined as $C = IDFT(\log(|DFT(w(t)\cdot x(t))|^2))$, $w(t)$ is the Hamming window of duration T, zero padding is used prior to the computation of the DFT. The entropy of the high-order cepstrum is computed as follows.

$$H(C) = -\sum_r P(C(r))\log(P(C(r)))$$

$$P(C(r)) = \frac{C(r)}{\sum_{r'} C(r')}$$

The entropy is computed over the indexes r and r' corresponding to a pitch region from 80 Hz to 450 Hz. The entropy of the high-order cepstrum is a measure that depends on the complete cepstral vector, rather than on a single value. Therefore it should be a more robust measure.

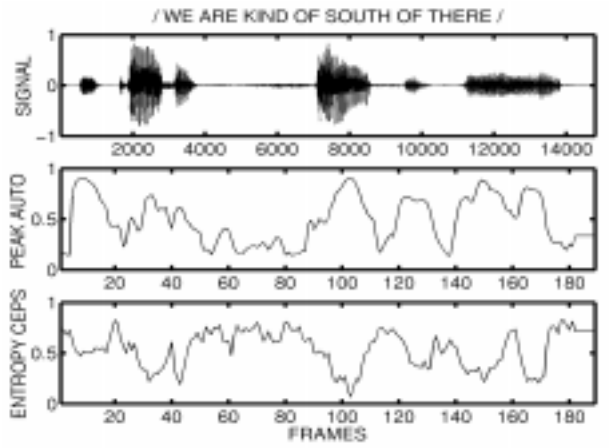

FIGURE 1: Voicing Features Graph

We chose to use the two voicing features together because of the complementary behavior of the time-domain-based features and the frequency-domain-based features for high and low pitch values. For low pitch values, the pitch periods are well separated, enabling well-defined correlation peaks and therefore providing reliable voicing estimation. For high pitch values, the harmonics in the spectrum are well separated and the high-order cepstrum peak is well defined, therefore providing more reliable voicing estimation.

Figure 1 shows the two voicing features for a waveform extracted from the Switchboard database. On top the waveform is presented. The "PEAK AUTO" graph corresponds to the normalized peak autocorrelation and the "ENTROPY CEPS" corresponds to the entropy of the high-order cepstrum.

3. VOICING FEATURE INTEGRATION

Here we describe the integration of the voicing features into SRI's DECIPHER large vocabulary continuous speech recognition (LVCSR) system.

3.1. Concatenating Voicing Features

The initial approach was to concatenate the two voicing features described in the previous section with the traditional 13-dimensional Mel cepstral feature vector (including energy) plus delta and delta-delta features (MFCC+D+DD), and then to train acoustic models with the resulting extended feature vector of dimension 41.

TABLE 1: Temporal Window Duration Optimization of the Voicing Feature Front End.

Recognition System	% WER
Baseline (No Voicing Features)	41.4 %
Baseline + 2 Voicing (25.6 ms)	41.2 %
Baseline + 2 Voicing (75.0 ms)	40.7 %
Baseline + 2 Voicing (87.5 ms)	40.5 %
Baseline + 2 Voicing (100.0 ms)	40.4 %
Baseline + 2 Voicing (112.5 ms)	41.2 %
Baseline + 2 Voicing (125.0 ms)	41.1 %

The voicing feature front end used the same frame rate as the standard feature front end. We optimized the window duration of the voicing feature front end to explore whether more reliable voicing activity estimation was achieved by increasing the temporal scope. Table 1 presents word error rate (WER) results, for both sexes, for different window durations of the voicing feature front end.

The details of the experiment are as follows: Acoustic models were trained with a 64-hour subset of the Switchboard 1, CallHome English, and Switchboard Cellular databases. These models were of the Genone (bottom-up clustered) type [5], gender-dependent, and were trained with maximum likelihood estimation (MLE). The standard features used were MFCC+D+DD of dimensionality 39, with a 25.6 ms window every 10 ms. The two voicing features were concatenated to the standard features. Vocal tract length (VTL) and speaker mean and variance normalizations were performed on the standard features only, as the voicing features are self-normalized. The test was done on the NIST Hub-5 dev2001 database. A bigram language model was used in decoding.

From Table 1 we see that using the voicing features with the same window duration as the standard front end produces a small WER reduction. As the window duration is increased the WER is further reduced. The optimum window duration for this task is found at 100 ms. These experimental results support the use of an independent front end, as clearly the optimal voicing feature window duration is different from the standard Mel cepstral feature window duration. The WER increases beyond the optimal point. For subsequent experiments a window duration of 75 ms was used in order to avoid being too close to the region where performance degrades sharply.

The behavior of the WER as a function of voicing feature window duration may be explained as follows: a longer voicing window than the initial value results in a better spectral resolution and therefore in a better representation of the harmonic structure. This helps capture the voicing activity more reliably. For a very long window the pitch may vary inside the window and therefore the harmonic structure may be less defined, producing a less reliable voicing activity estimation.

3.2. Voiced/Unvoiced Posterior Probability Features

In this experiment, the same concatenation scheme described above was used. This time, however, a log posterior feature was derived from a two state hidden Markov model (HMM) trained for voiced/unvoiced detection. This procedure is similar to the one presented in [6]. The normalized log frame energy and the two voicing features were used as input features for this HMM.

TABLE 2: Voice/Unvoiced Posterior Feature from Two-State HMM.

Recognition System	% WER
Baseline	39.2 %
Baseline + Posterior Feature	39.7 %

The HMM had one Gaussian per state and was trained in an unsupervised way for each sentence. The parameters were initialized so that one state represented a voiced region, and the other an unvoiced region.

Once the HMM was trained, the log posterior probability corresponding to the voiced state was computed. This log posterior probability was used as an additional feature and was appended to the MFCC features. The total feature dimension was thus 40. The experimental conditions were the same as described in Section 3.1. Results are presented in Table 2 (for male speakers only).

Use of this feature actually degraded performance. One possible explanation for this result may be that the posterior probability feature may not represent the soft transitions between voiced and unvoiced segments. Analyzing the log posterior probability feature, we observed that it has sharp transitions between voiced and unvoiced states. A hard decision may not be well suited for multi-state models commonly used in acoustic modeling, and a smoothing scheme may be needed.

3.3. Voicing Multi-Frame Window plus HLDA

The next step was to integrate the voicing features into a more state-of-the-art front end, incorporating third differential features and the HLDA algorithm to reduce the feature dimensionality. In this system we extended the concatenation from single-frame voicing features to a multi-frame window of voicing features. The idea was to explore whether a linear combination of multiple frames of voicing features produced a more reliable voicing estimator than the voicing features from a single frame. The HLDA algorithm was applied to the combined feature vector, and the final feature dimension was 39.

We optimized the number of voicing feature frames to be appended to the standard Mel cepstral feature. Table 3 reports the WER for a single-frame window, five-frame window and nine-frame window, for both sexes. In all cases the HLDA algorithm reduced the feature dimensionality to 39.

In Table 3 we find that using single-frame voicing features results in a small WER reduction. Extending the number of frames to five leads to a further WER reduction. This shows that voicing activity is estimated more reliably by combining the voicing features from multiple frames. For a nine-frame window the gain was smaller.

TABLE 3: Multi-Frame Window Optimization of Voicing Features plus HLDA.

Recognition System	% WER
Baseline + HLDA	39.9 %
Baseline + 1 frame, 2 voicing + HLDA	39.6 %
Baseline + 5 frame, 2 voicing + HLDA	38.8 %
Baseline + 9 frame, 2 voicing + HLDA	39.3 %

The weighting of the window of voicing features produced by the HLDA algorithm in the five-frame window is similar to an average. This indicates that for improved recognition it is better to apply some temporal smoothing to the voicing features.

3.4. Delta of Voicing Features plus HLDA

As an alternative to the previous approach, we explored computing first and second differences of the voicing features instead of a multi-frame window of voicing features. The total number of voicing features per frame was six. The concatenation of the MFCC+D+DD features and the voicing features plus their deltas generated an extended feature, which was reduced with the HLDA algorithm to 39 dimensions. The recognition setup was the same as before. The WER results of this experiment are presented in Table 4 and include only male speakers.

TABLE 4: Deltas of Voicing Features plus HLDA.

Recognition System	% WER
Baseline + HLDA	37.5 %
Baseline + Delta Voicing + HLDA	37.6 %

We did not find any improvement using this feature representation, probably because the variability in the voicing features across frames produces deltas with high variance.

4. TWO EVALUATION-TYPE EXPERIMENTS

In Section 3 it was found that the best combination for integrating the voicing features was a five-frame window of voicing features together with the HLDA algorithm, using a window duration of 75 ms. This combination was integrated into SRI's conversational telephone speech recognition system. Described below are two experiments evaluating the WER reduction obtained using the voicing features in evaluation-type multi-pass systems similar to those used for the NIST speech recognition evaluations [9].

4.1. A Multi-Pass Recognition System

We trained within-word gender-dependent genonic triphone acoustic models with and without voicing features on approximately 418 hours of the Switchboard, CallHome English and Switchboard Cellular databases.

We then tested a multi-pass recognition system on the NIST Hub-5 eval2002 database, which contains approximately 5000 utterances from 120 speakers. This multi-pass system was tested with models with and without voicing features.

In Pass 1, the acoustic models were adapted using a phone loop, and N-best lists were generated. In Pass 2, the N-best lists were rescored with a 4-gram SuperARV almost-parsing language model [8]. In Pass 3, the N-best lists were rescored with duration models and decoded with a confusion network algorithm. The goal of Passes 2 and 3 was to improve the N-best lists in order to provide better hypotheses for unsupervised adaptation. In Pass 4,

a Maximum Likelihood Linear Regression (MLLR) adaptation of the non-cross-word models was performed using the 1-best hypotheses obtained from the rescored N-best lists. Lattices were generated from the MLLR adapted model and from these lattices new N-best lists were generated. Pass 5 was the same as Pass 2, but with the new N-best lists. Pass 6 was the same as Pass 3, but with the new N-best lists. The WER results after each pass, with and without the voicing features, are presented in Table 5 for both sexes.

TABLE 5: Multi-Pass Recognition System Test on EVAL2002 Database. Results in WER, Relative Percentage Reduction in Parentheses

	System	Without Voicing Features	With Voicing Features
1	Phone Loop Adapt. N-best Generation	38.6%	37.8% (-2.1%)
2	Rescored N-best with 4-gram SuperARV Language Models	34.7%	33.6% (-3.2%)
3	Rescored N-best with Duration and Confusion Network	33.6%	32.5% (-3.3%)
4	MLLR Adaptation on 1-best Hypotheses New N-best Generation	36.5%	35.7% (-2.2%)
5	Rescored new N-best with 4-gram SuperARV Language Models	31.5%	31.0% (-1.6%)
6	Rescored N-best with Duration and Confusion Network	30.6%	30.0% (-2.0%)

In Table 5 we see that the gain from Pass 1 is a relative WER reduction of 2.1% compared to the system without voicing features. The gain increased after Passes 2 and 3, resulting in better adaptation hypotheses for the voicing feature models. After Pass 4, the gain from the voicing features is smaller, showing that the adaptation did not take full advantage of the better N-best lists from the voicing features. From the adapted model, lattices were generated in Pass 4. The lattice error rate was computed (which is the WER of the best possible word combination in the lattice) with and without the voicing features. The lattice error rate for the voicing features models was 4.41% and the lattice error rate for the models without voicing features was 4.46%. After rescoring the new N-best lists the final gain from the voicing features is 2.0% relative. It is worth noting that the relative gain from this new knowledge source, the voicing features, is preserved in this multi-pass system.

4.2. Adaptation Experiment in Advanced Pass

In this experiment, we used SRI's complete evaluation system [9] and tested the effect of voicing features just prior to the final N-best rescoring stage. The acoustic models in this case are cross-word triphone models trained with maximum mutual information estimation (MMIE). The models both with and without voicing features have been adapted by a nine-transform full-matrix MLLR, based on prior recognition output from a separate PLP-based system that also incorporated voicing features. (We therefore expect less difference between the two contrasting systems.) The results are presented in Table 6.

TABLE 6: Voicing Features In Advanced Pass

Recognition System	% WER
Baseline EVAL	25.6 %
Baseline EVAL + Voicing Features	25.1 %

The relative WER reduction from the voicing features is again 2.0%, but now relative to a significantly better baseline than that used in the previous experiment.

5. CONCLUSIONS

We explored the integration of voicing features in a large vocabulary speech recognition system. We validated the use of an independent front end to compute the voicing features. After exploring several integration approaches, we found that the best was to combine a multi-frame window of voicing features with the MFCC plus third differential features, using linear discriminant analysis. We achieved a consistent gain across the different stages of a multi-pass system in evaluation-type experiments.

6. ACKNOWLEDGEMENTS

This work was funded by DARPA under contract No. MDA972-02-C-0038.

7. REFERENCES

[1] D. L. Thomson and R. Chengalvarayan, "Use of Voicing Features in HMM-based Speech Recognition", *Speech Communication*, vol. 37, pp. 197-211, 2002.

[2] A. Ljolje, "Speech Recognition Using Fundamental Frequency and Voicing in Acoustic Modeling", in *International Conference on Spoken Language Processing*, Denver, CO, September 2002, pp. 2177-2140.

[3] A. Zolnay, R. Schulter and H. Ney, "Extraction Methods of Voicing Feature for Robust Speech Recognition", *Proceedings of Eurospeech*, September 2003, pp. 497-500.

[4] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub. "Large-vocabulary dictation using SRI's DECIPHER speech recognition system: Progressive search techniques", in *Proceeding ICASSP*, vol. II, pp. 319-322, Minneapolis, 1993

[5] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model Based Speech Recognizers", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 281-289, 1996.

[6] M. Arcienega and A. Drygajlo, "Robust Voiced-Unvoiced Decision Associated to Continuous Pitch Tracking in Noisy Telephone Speech", *ICSLP2002*, Denver, Colorado, pp. 2433-2436

[7] M. J. F. Gales, "Semi-tied Covariance Matrices for Hidden Markov Models", *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272-281, 1999.

[8] W. Wang and M. P. Harper, "The SuperARV Language Model: Investigating the Effectiveness of Tightly Integrating Multiple Knowledge Sources", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, pp. 238-247.

[9] DARPA RT-03 Workshop, Boston, May 2003. http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/.