AN IMPROVED UNION MODEL FOR CONTINUOUS SPEECH RECOGNITION WITH PARTIAL DURATION CORRUPTION

Ji Ming

School of Computer Science
The Queen's University of Belfast
Belfast BT7 1NN, UK

ABSTRACT

The probabilistic union model is improved for continuous speech recognition involving partial duration corruption, assuming no knowledge about the corrupting noise. The new developments include: an n-best rescoring strategy for union based continuous speech recognition, a dynamic segmentation algorithm for reducing the number of corrupted segments in the union model, and a combination of the union model with conventional noise-reduction techniques to accommodate the mixtures of stationary noise (e.g. car) and random, abrupt noise (e.g. a car horn). The proposed system has been tested for connected-digit recognition, subjected to various types of noise with unknown, time-varying characteristics. The results have shown significant robustness for the new model.

1. INTRODUCTION

This paper studies noisy speech recognition assuming that there is no knowledge about the noise, except that the noise is *shorter* than the speech utterance. We term this a partial temporal (or partial duration) corruption. Partial temporal corruption may be caused by time-limited or time-selective noise, for example, a car horn, a shut door, random channel impulses, click sounds from a keyboard or any type of burst noise occurring during the utterance and affecting only certain parts of the speech signal. There may be two different ways to deal with this type of noise for speech recognition. Firstly, we may use the conventional noise-reduction techniques to remove the noise from the signal, or to adapt the model to the noisy observations. However, this may prove difficult because these techniques usually require certain knowledge such as the spectral or cepstral characteristics of the noise, and these can be difficult to estimate given the variety, unpredictability and nonstationary nature of the abrupt noise as mentioned above. Alternatively, we may base the recognition mainly on information from the clean parts of the signal, by ignoring the noisy parts, or by making these parts play a less significant role. This recognition is possible due to the redundancy of the temporal characteristics of speech. This method is of interest because no knowledge is required for the noise, except its location. A better system may be a combination of these two methods, i.e., using the noise reduction technique to remove the noise with a known or trainable characteristic, and exploiting the redundancy in the speech signal to get around the noise with an unknown or time-varying nature. This paper is focused on the second method, but we use a simple example to demonstrate the advantage of combining the two methods.

Speech recognition, given that only partial temporal/spectral features are reliable, has been discussed previously in the context of missing feature theory (see, e.g. [1]-[5]). Instead of requiring a detailed knowledge of the noise for clearing the corrupted features, the missing feature method requires only a labelling of every feature as reliable or corrupt, for removing the unreliable features from recognition. Unfortunately, locating the corrupted data itself can be a difficult task. Recent studies have suggested that the unreliable data may be identified by explicitly measuring the local signal-to-noise ratio (SNR), based on a running estimate of the local noise spectrum via spectral subtraction [3][4]. This method performs well when the corrupting noise is stationary. For unknown or nonstationary noise, Seltzer *et al.* [5] have suggested that some characteristics of the speech signal itself, such as the harmonic nature of voiced speech, may be exploited for identifying the corrupted time-frequency regions.

For dealing with unknown, nonstationary noise, we have recently studied a new approach, i.e. the probabilistic union model [6-8]. Unlike the missing feature method, the union model does not require the identity of the noisy data, instead, it combines the local information based on the union of random events, to reduce the dependence of the model on information about the noise. The union model has been previously applied to the combination of sub-band information for speech recognition, assuming that the unknown corruption is localized in certain areas of the frequency band [6]. The present research considers the corruption localized in the time duration, which is not necessarily band-limited. A preliminary study of this, for isolated-word recognition, has been presented in [7]. The present paper deepens this study. In particular, we describe several new advances in this model. The first advance is an n-best rescoring strategy for incorporating the union model into continuous speech recognition. The second advance is a dynamic segmentation algorithm for reducing the number of corrupted segments in the union model. A further advance is a combination of the union model and conventional noise compensation methods, for dealing with a mixture of stationary noise and unknown burst noise. In the following we begin with an overview of the union model, and then describe the improvements, followed by an experimental evaluation.

2. THE UNION MODEL

Assume that in speech recognition a speech utterance can be represented by a sequence of short-term spectral vectors (i.e. frames) $X = (x_1, x_2, \ldots, x_T)$, where each frame x_t characterizes the temporal spectrum of speech at time t. The presence of a

time-limited or time-selective noise can cause some of the x_t to be corrupted. Thus, we face the problem of how to calculate the probability for X, given that some of the frames may be noisy. The idea of the missing feature method is that the acoustic mismatch due to the noise can be effectively reduced by simply ignoring the strongly affected features. However, because of the uncertainty of the noise, the identities of the corrupted frames are unknown. The probabilistic union model is a method that can be used to select usable features from a given feature set, without requiring the identity of the corrupted features.

The union model deals with the uncertainty of the corrupted frames by combining the subsets of the frames using the inclusive "or" (i.e. disjunction) operator. Let $P(X)$ be the probability of the observation sequence X. With the union model, this probability can be expressed in a general form as

$$P(X) = P(\bigvee_{t_1 t_2 \cdots t_{T-M}} x_{t_1} x_{t_2} \cdots x_{t_{T-M}}) \tag{1}$$

where the symbol $\vee$ represents the inclusive "or" operator, which is applied to combine all possible subsets $x_{t_1} x_{t_2} \cdots x_{t_{T-M}}$ of $(T-M)$ frames within $(x_1, x_2, ..., x_T)$, and M is called order of the model, with a value $0 \le M \le T-1$. For example, in the case with four frames (x_1, x_2, x_3, x_4), the union model probability $P(X)$ can take four possible forms, corresponding to order $M = 0, 1, 2$ and 3, respectively:

M=0: $P(X) = P(x_1 x_2 x_3 x_4)$ (2)

M=1: $P(X) = P(x_1 x_2 x_3 \vee x_1 x_2 x_4 \vee x_1 x_3 x_4 \vee x_2 x_3 x_4)$ (3)

M=2: $P(X) = P(x_1 x_2 \vee x_1 x_3 \vee x_1 x_4 \vee x_2 x_3 \vee x_2 x_4 \vee x_3 x_4)$ (4)

M=3: $P(X) = P(x_1 \vee x_2 \vee x_3 \vee x_4)$ (5)

A union model of order M is suited for accommodating a maximum of M noisy frames, in terms of leaving at least one subset of $(T-M)$ frames in the model not affected by the noise. To illustrate this, use the above example with order 2, assuming two corrupted frames with unknown identity. The union probability $P(X)$ for order $M = 2$ can be approximated as

$$\begin{aligned} P(X) \approx\ & P(x_1 x_2) + P(x_1 x_3) + P(x_1 x_4) \\ & + P(x_2 x_3) + P(x_2 x_4) + P(x_3 x_4) \end{aligned} \tag{6}$$

where we have omitted the terms corresponding to the joint probabilities between the $x_i x_j$'s, assuming that these are small and can be neglected in comparison to the other terms [6]. As indicated in (6), the union model includes the probabilities of all possible combinations between two frames, and thus it includes the probability for the remaining two "clean" frames, providing correct information about the probability of X. The probability containing only the clean frames should usually dominate the probability $P(X)$ for the correct model, because of small mismatch between the model and data. As such, recognition can be based on the union probability $P(X)$, and hence no information is needed for the identity of the two noisy frames.

The above union model can be implemented based on the HMM techniques. To retain the inter-frame correlation, as well as to reduce the number of combinations involved in computing the union probability (1), we model the segments instead of frames. For each test frame sequence $(x_1, x_2, ..., x_T)$, we first convert it into a sequence of segments $(z_1, z_2, ..., z_N)$, where each segment z_n consists of the same number of consecutive frames, and then compute the union probability for the segments. Given the state sequence $S = (s_1, s_2, \ldots, s_T)$ associated with the frame sequence, the segment union probability can be approximated as

$$P(X|S) \approx \sum_{n_1 n_2 \cdots n_{N-M}} P(z_{n_1}|S) P(z_{n_1}|S) \cdots P(z_{n_{N-M}}|S) \tag{7}$$

where the summation is over all possible combinations of N values $(1, \ldots, N)$ taken $N-M$ at a time, and $P(z_n|S)$ is the probability of the segment z_n, defined by

$$P(z_n|S) = \prod_{x_t \in z_n} b_{s_t}(x_t) \tag{8}$$

where $b_i(x)$ is the frame-based observation probability distribution in state i. Because local frame corruption within a segment affects the probability of the segment (i.e. (8)), a segment is considered to be noisy if part or all of its frames are noisy.

3. IMPROVEMENTS

The above model, (7) and (8), has been previously applied to isolated-word recognition [7]. In recognition, we assumed that the word-based state sequence, required for calculating the union probability (7), can be derived by using the standard Viterbi algorithm, even though there may be some noisy frames in the observation sequence. Our experiments have indicated that this appears to be effective as well as being simple. To apply the above model to continuous speech recognition, our first improvement is to adopt a two-pass, n-best rescoring approach. In the first pass, the HMMs are applied to generate n-best sentence (i.e. state sequence) alternatives by using the Viterbi algorithm. In the second pass, the union model is applied to the segment probabilities, associated with each hypothesized state sequence, to produce a union probability on which the final recognition decision is based. In rescoring, the capability of the union model for ignoring the strongly corrupted data is exploited to reduce the effect of the corrupted segments on recognition.

As described above, modeling segments of frames instead of individual frames is desirable to retain the discriminative information. The simplest way for this segmentation is to divide the test observation sequence uniformly into N segments, each segment corresponding to a specific z_n. A drawback of this method is that, for example, when there are some noisy frames that are shorter than a segment and lying across a border of two segments, then both the segments will be affected by the noise. Fig. 1 shows another example in which noisy frames shorter than two segments can affect three segments due to the fixed-border segmentation. Our second improvement is therefore a dynamic segmentation method, which adjusts the origin of the segments for each utterance such that only the smallest number of segments may be affected by the noise, as illustrated in Fig. 1. This is accomplished based on a maximum-probability criterion.

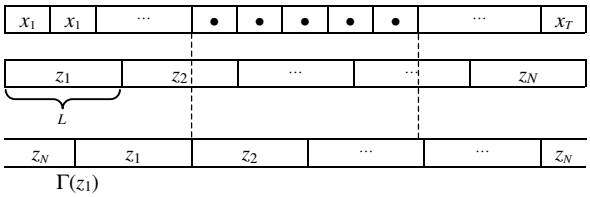

Fig. 1. Top: a frame sequence with • representing noisy frames. Middle: uniform segmentation with three segments affected by the noise. Bottom: dynamic segmentation with only two segments affected by the noise.

Denote by $\Gamma(z_1)$ the frame time that defines the origin of the first segment and hence the borders of all the segments. We then can write the union probability as a function of $\Gamma(z_1)$, i.e. $P(X \mid S, \Gamma(z_1))$, for a given segmentation. In recognition we search for the $\Gamma(z_1)$ to maximize this probability over the range of the frame time (1, *L*), where *L* represents the length of each segment. As shown in Fig. 1, as $\Gamma(z_1)$ is increased, the last segment z_N moves back to the beginning of the frame sequence, so there is no information lost. We call this method maximum-probability segmentation, in contrast to uniform segmentation. In the experiments we have tested both methods. It was found that they produced the same recognition accuracy for clean utterances. However, when noises were present, the maximum-probability segmentation method outperformed the uniform segmentation method, especially for the low SNR conditions. The maximum-probability segmentation method was used to produce the results presented in Section 4.

A further improvement is the combination of the union model with conventional noise-reduction techniques. So far we have assumed no prior knowledge about the times of occurrence and the characteristics of the noise. This is typical for random, abrupt noise. However, the real-world noise may be a mixture of stationary noise and abrupt noise. For stationary noise, with reasonably sufficient observations, it is possible to obtain an estimate of the noise statistics. Thus, we may build a system in which the union model and some conventional noise-reduction techniques are combined, to deal with this type of mixed noise. The stationary noise component may be removed, for example, by spectral subtraction or noise compensation, and the remaining unknown burst noise component can then be dealt with by the union model. An example system will be described in Section 4.

4. EXPERIMENTS

The TIDigits connected digits database was used for the experiments. This database contains a total of 6196 test utterances for speaker-independent connected digit recognition. Each test utterance may contain a string of 2, 3, 4, 5 or 7 digits, assuming no advance knowledge of the number of digits in an utterance. The speech was sampled at 8 kHz, and divided into frames of 256 samples. Each frame was featured using a 20-element vector, including 10 mel-frequency cepstral coefficients and their first-order delta parameters. Each digit was modeled with a 10-state HMM trained on clean training data, with each state containing eight mixture Gaussian densities. A silence HMM with one state was also built to account for the silences surrounding each utterance and the optional silences between digits. These HMMs were used to produce the state sequence for the union model, and also served as the baseline system for comparison.

As shown in (7), there are two parameters in the union model, i.e., the number of segments for each utterance, N, and the order of the model, M. We have tested the model with different lengths for a segment, to search for a balance between the noise localization and linguistic discrimination. We found that a segment length around ten frames (about 160ms) was suitable. Given the length of the segment, the number of segments N can be variable across utterances with different duration. As such, it would be more convenient to calculate the relative order M/N. A relative order of 0.2, for example, may accommodate up to 20% of the segments in each utterance to be corrupted. As described earlier, we used an n-best rescoring strategy for continuous speech recognition. In all the experiments, we limited the number of the rescored alternatives, n, to 50.

Firstly, we tested the models for clean utterance recognition. Table I presents the string accuracy obtained by the union model applied to rescore the top 50 string alternatives produced by the Viterbi algorithm, along with the accuracy by the baseline HMM. The n-best accuracy (i.e. the rate that the correct string is contained in the n-best alternatives) is also included in the table. As expected, the performance of the union model decreased as the order was increased, because of the disjunction between the clean segments. In practice, we need a high order to accommodate as many noisy segments as possible, but a low order to obtain an acceptable performance for clean speech recognition, i.e. a balance between robustness and clean speech performance. We have found that for connected digit recognition an order $M/N = 0.2$ provides a good balance. Therefore in the following we use this order for further experiments involving noise corruption. As shown in Table I, this order offered a string accuracy of about 94% for clean utterance recognition.

Next, we tested the union model assuming that each utterance involved a partial temporal corruption. Four different types of real-world noise, a bell, a door slam, a telephone ring, and a gunshot, were used to corrupt the utterances. The noise was additive, and the SNR was calculated relative to the part of speech where the noise was added. For each utterance, the corruption was centered at one of the five positions: beginning, middle, end, a quarter's position and three quarter's position, which was chosen randomly for each utterance. The duration of the noise was 10% and 20%, respectively, of the duration of the speech utterance. Table II presents the average string accuracy, as a function of the SNR, averaged over all the noise types. We see that the union model significantly improved upon the baseline model throughout all the noise conditions. We also see that there is still a large gap between the n-best accuracy and the union model accuracy. Note that for a 5-, 6- or 7-digit utterance, a 20% duration corruption may affect the duration of a whole word or longer. This can cause the information of a whole word to be lost, which is difficult to recover without context knowledge. Further improvement may be obtainable by combining with a language model, for recognition of a text sentence.

Table I. String accuracy (%) for clean utterances

Union model Relative order (M/N)					Baseline HMM	n-best (n=50)
0.0	0.1	0.2	0.3	0.4		
97.53	96.35	94.14	84.76	74.73	97.53	99.95

Table II. String accuracy (%) with noise corrupting the duration of each utterance by 10% and 20%, respectively

SNR (dB)	Union model		Baseline HMM		n-best (n=50)	
	Corruption					
	10%	20%	10%	20%	10%	20%
10	92.27	88.83	89.20	82.90	99.71	99.31
0	87.62	81.34	76.55	60.51	98.29	96.16
−10	80.63	62.90	61.51	39.14	93.92	83.15

We further conducted experiments by introducing multiple duration corruptions into a single utterance. In particular, we assumed that the noise occurred twice at different times within an utterance, each occurrence causing a local temporal corruption. The times at which the noise occurred were any two of the five: beginning, middle, end, a quarter's position and three quarter's position of the speech utterance, chosen randomly for each utterance. Each occurrence of the noise corrupted about 10% of the duration of the speech utterance. Table III presents the results, averaged over the four types of noise as described above.

Table III. String accuracy (%) with two noise corruptions at different times in the utterance, corrupting the duration of each utterance by a total of 20%

SNR (dB)	Union model	Baseline HMM	n-best (n=50)
10	91.20	85.17	99.56
0	83.39	59.05	95.35
−10	62.81	31.63	81.52

Finally, we tested the combination of the union model with conventional noise compensation for recognizing noisy utterances involving both stationary noise corruption and unknown burst noise corruption. The stationary noise was a car noise, and the burst noise was a car horn, which occurred at a random time within an utterance, lasting for about 10% of the duration of the utterance and simulating a further unknown unexpected corruption occurring to the utterance. The SNR of the stationary noise and burst noise were 10 dB and 0 dB, respectively, which were calculated separately relative to the clean speech data. To reduce the stationary noise, we assumed that we had the models trained in the car environment, so that the mismatch between the model and data, due to the existence of the stationary noise, could be reduced. While we assumed knowledge about the occurrence of the stationary noise, we assumed no knowledge about the occurrence of the car horn during the utterance. Table IV presents the results, showing the advantage of the combination of the union model and noise compensation technique for dealing with the mixed noise.

Table IV. String accuracy (%) with combined noise compensation and union model for mixed stationary noise (car, SNR=10 dB) and unknown burst noise (a car horn, SNR=0dB)

	Union model	Baseline HMM
No compensation	36.27	29.34
With compensation	79.52	60.23

5. SUMMARY

This paper introduced our recent efforts in enhancing the capability of the probabilistic union model for continuous speech recognition involving partial duration corruption. The new developments include an n-best rescoring strategy for union based continuous speech recognition, a dynamic segmentation algorithm for reducing the number of noisy segments in the union model, and a combination of the union model with conventional noise-reduction techniques. The improved model has been tested for connected digit recognition subjected to various types of abrupt noise with unknown, time-varying characteristics, and has shown significant noise robustness.

This work is supported by UK EPSRC grant GR/M93734.

6. REFERENCES

[1] M, Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition", *ICASSP'97*, pp. 803-806.

[2] R. P. Lippmann and A. B. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", *Eurospeech'97*, pp. 37-40.

[3] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory", *ICASSP'98*, pp. 121-124.

[4] A. Vizinho, P. Green, M. Cooke and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study", *Eurospeech'99*, pp. 2407-2410.

[5] M. L. Seltzer, B. Raj and R. M. Stern, "Classifier-based mask estimate for missing feature method of robust speech recognition", *ICSLP'2000*.

[6] J. Ming and F. J. Smith, "Union: a new approach for combining sub-band observations for noisy speech recognition", *Speech Communication*, Vol. 34, pp. 41-55, 2001.

[7] J. Ming, D. Stewart, P. Hanna and F. J. Smith, "A probabilistic union model for partial and temporal corruption of speech", *IEEE ASRU Workshop'99*, pp. 43-46.

[8] http://www.cs.qub.ac.uk/~J.Ming/