ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ROBUST FEATURE SELECTION USING PROBABILISTIC UNION MODELS

Ji Ming, Peter Jancovic, Philip Hanna, Darryl Stewart, F. Jack Smith

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

ABSTRACT

This paper provides a summary of our recent work on robust speech recognition based on a new statistical approach - the probabilistic union model. In particular, we considered speech recognition involving partial corruption in frequency bands, in time duration, and further in feature components. In all these situations, we assumed no prior knowledge about the corrupting noise, e.g. its band location, occurring time and statistical distribution. The new model characterizes these partial, unknown corruptions based on the union of random events. For the evaluation, we have conducted isolated-word recognition tasks by using both a speaker-independent E-set database and the TiDigits database, each being corrupted by various types of additive noise with unknown, time-varying statistics. The results indicate that the probabilistic union model offers robustness to partial corruption in speech utterances, requiring little or no knowledge about the noise characteristics.

1. INTRODUCTION

This paper provides a summary of our recent work on robust speech recognition based on a new statistical approach - the probabilistic union model [1-5]. In particular, we consider speech recognition subjected to partial corruption in three aspects:

1) in frequency bands;
2) in time duration;
3) in feature components.

Partial corruption accounts for the effects of many real-world noises. Partial frequency-band corruption may be caused by frequency-selective noise, for example, a phone ring, a passing car, a siren or a random channel tone. Partial temporal duration corruption may be caused by abrupt noise, for example, a shut door, a channel impulse or any type of burst noise. For human beings, these temporally localized noises normally do not destroy the intelligibility of an utterance. However, for automatic speech recognition systems this typically proves problematic, especially if the noise is unpredictable and of an unknown or time-varying nature. Such noise characteristics cause particular difficulties for obtaining accurate and sufficient information for model adaptation or compensation, given that the noise may occur in the middle of a speech utterance.

Recent studies towards a solution to the partial frequency-band corruption have included the multi-band approach (e.g. [6][7]). This involves a division of the entire speech frequency-band into several sub-bands, to isolate the effect of the frequency-localized noise; those sub-bands that remain unaffected by the noise can thus be used for recognition. To achieve this, one needs to select, in a given set of bands, a sub-set of bands that carry useful information about the speech utterance. This selection can be difficult without prior information about the noisy bands.

To deal with the partial temporal corruption, we may use a multiple time-segment analysis method to extract features for the utterance [2][4]. This shares certain characteristics with the multi-band analysis method used for dealing with the partial frequency-band corruption. The multi-segment approach divides each utterance into several segments and processes each segment independently, such that the effect of any local temporal corruption could be isolated from other usable segments. Again, we face the problem of how to select, in a given set of segments, a sub-set of segments that carry useful information. This selection can be difficult without knowledge about the noisy segments.

In addition to the above partial frequency-band and partial temporal duration corruption, one may also encounter partial feature component corruption. This refers to some of the components in a given acoustic feature vector being noisy. In speech recognition, a speech utterance may be represented by multiple feature streams, typically, the static spectra and dynamic spectra, over varying time scales. In real-world applications, due to the background noise or channel effects, there may be only a subset of the given feature components that remains useful. For example, in the presence of stationary noise, the static spectral components will be corrupted; but because the dynamic spectral components (e.g. the delta components) are less affected by the stationary noise, they may still provide useful information for recognition. However, without prior knowledge about the noise conditions, it can be difficult to decide which subset of the feature components provides useful information.

The above three problems can be unified as a feature selection problem, i.e. selecting features that carry useful information from a given feature set $o = \{o_1, o_2 \ldots, o_N\}$, where each o_n may represent a feature component for a specific sub-band, or for a specific segment, or for a specific feature stream. We are given that some of the o_n's may be noisy, but without appropriate knowledge about the noise characteristics, particularly, the position (i.e. which o_n's are affected) and intensity of the noise. We tackle this feature selection problem involving partial unknown corruption by using the probabilistic union model, previously described in [1-5]. In the following we start to describe the general principle of the model, and then move to its specific applications to speech recognition involving partial unknown corruption in frequency-band, in temporal duration and in feature components, respectively.

2. PROBABILISTIC UNION MODEL

Assume an observation consisting of N feature components $o = \{o_1, o_2 \ldots, o_N\}$. The recognition is based on the likelihood of o associated with each word model. To calculate this likelihood, the traditional approach is to combine the feature components o_n's using the "and" (i.e. conjunction) operator $\wedge$ (although this is not usually explicitly stated). Thus, assuming independence between the feature components, we can obtain the likelihood $p(o)$ as the product of the individual likelihoods $p(o_n)$'s, i.e.

$$\begin{aligned} p(o) &= p(o_1 \wedge o_2 \wedge \cdots \wedge o_N) \\ &= p(o_1)p(o_2)\cdots p(o_N) \end{aligned} \quad (1)$$

For convenience, we call (1) the *product model*. This model has a drawback: when the individual probability densities $p(x_n)$'s are trained on clean speech and used for modeling an observation with some noisy components, then the corresponding $p(\tilde{o}_n)$'s for the noisy $\tilde{o}_n$'s will be highly inaccurate–when the noise is strong they can become almost zero. This destroys the model's ability to discriminate between correct and incorrect word classes. Unless the noisy components can be identified this is difficult to correct.

As an alternative, given no knowledge about the noise, we can assume that the useful features in the given observation may be *any* of the o_n's, $n = 1, \cdots, N$, or *any* of the combinations among the o_n's up to the complete feature set. This can be expressed, using the inclusive "or" (i.e. disjunction) operator $\vee$, as

$$o_\vee = o_1 \vee o_2 \vee \cdots \vee o_N \quad (2)$$

where $o_\vee$ is a combined observation based on $\vee$, representing the useful features within $\{o_1, o_2 \ldots, o_N\}$. For example, using a 3-component observation, the expression $o_\vee = o_1 \vee o_2 \vee o_3$ assumes that the useful features within the given (o_1, o_2, o_3) may be o_1, or o_2, or o_3, or $o_1 \wedge o_2$, or $o_1 \wedge o_3$, or $o_2 \wedge o_3$, or $o_1 \wedge o_2 \wedge o_3$. These combinations characterize, respectively, an observation in which there are two-component, one-component and no component corruption, therefore covering all possible partial corruptions, including the no corruption case that may be encountered in a 3-component observation. In general, if an observation consists of N components, and these components may be subjected to some partial unknown corruption, then the useful information contained in this observation may be modeled by (2). This model takes into account all possible partial corruptions, thereby requiring no knowledge about the actual noise.

Assume that the o_n's are discrete random events, then $o_\vee$ is the union of the o_n's. Thus, we can compute the probability $P(o_\vee)$ based on the rules of probability for the union of random events. This probability, for each modeled word, can then be used to decide the recognized word based on the maximum-likelihood principle. Assume independence between the o_n's, note that $\vee_{m=1}^{n} o_m = (\vee_{m=1}^{n-1} o_m) \vee o_n$, so $P(o_\vee)$ can be computed using a recursion

$$P(\vee_{m=1}^{n} o_m) = P(\vee_{m=1}^{n-1} o_m) + P(o_n) - P(\vee_{m=1}^{n-1} o_m) P(o_n) \quad (3)$$

for $n = 2, \cdots, N$. This computation requires only the probability distributions of the individual components $P(x_n)$'s, which are assumed to be trained on clean training data. We call (2)-(3) the *probabilistic-union model*, as opposed to the product model (1).

Since the $P(o_n)$'s are generally not large, (3) is effectively the sum of the individual probabilities. The advantage of (3) over (1) for noisy speech is that, for $P(x_n)$'s trained for clean speech, the value of $P(\tilde{o}_n)$ for a noisy $\tilde{o}_n$ can be very small and as such makes a small contribution to (3). Therefore the almost random variation of $P(\tilde{o}_n)$ between the correct and incorrect words will have little effect on $P(o_\vee)$. So $P(o_\vee)$ is dominated by noiseless feature components. The disadvantage of (3) is that it effectively averages the ability of each feature component to discriminate between correct and incorrect words, unlike (1) where each component reinforces the other. This problem may be overcome by combining the use of "and", "or" operators between the feature components. For an N-component observation, such a combined model can be expressed in a general format as

$$o_\vee = \bigvee_{n_1 n_2 \cdots n_{N-M}} o_{n_1} o_{n_2} \cdots o_{n_{N-M}} \quad (4)$$

where the $\wedge$ operator between the o_n's has been omitted; $0 \le M < N$; and the "or" is taken over all possible combinations of $n_1 n_2 \cdots n_{N-M}$ with each $n_i \in (1, \cdots, N)$, giving a total of ${}^N C_{N-M}$ combinations. We call (4) a *union model of order M*. This model is suited to the observation with a maximum of M noisy components but without information about where these are located. In this case (4) will include, through the inclusion of all possible conjunctions of $(N - M)$ components, at least one conjunction of $(N - M)$ clean components which will dominate the union probability. The other conjunctions including noisy components will have low probabilities (because the $P(x_n)$'s are trained on clean speech) and therefore make only a small contribution to the union probability. Model (4) is reduced to model (2) when order $M = N - 1$ and to the product model (1) when order $M = 0$. In our experiments, we choose a model order to accommodate as much noise as possible, subject to an acceptable performance for clean speech recognition [5].

3. APPLICATION TO SUB-BAND BASED SPEECH RECOGNITION

Consider speech recognition subjected to partial unknown frequency-band corruption. We tackle this problem based on the multi-band analysis method. Then the feature set $o = \{o_1, o_2 \ldots, o_N\}$ corresponds to N sub-band observations, with a possibility that some of the o_n's are corrupted, but no information about the noisy bands. The union model described above is employed to select the sub-band features from the given feature set for recognition. In particular, the above union model (4) has been built into an HMM to combine these sub-band observations on the frame level [1][3][5]. To calculate the sub-band features, a multi-channel, Mel-scaled filter bank is used to estimate the log-amplitude spectra of speech; these filter-bank channels are then grouped uniformly into sub-bands, for which the MFCC's and the corresponding ΔMFCC's are computed as the sub-band observations. Experiments are based on a speaker-independent E-set database provided by BT. This database contains three repetitions of each E-set word (b, c, d, e, g, p, t, v) by 104 speakers, 53 male and 51 female. A typical baseline HMM achieves an accuracy of ~85% over this database.

The models are trained on clean training data and tested on noisy test data, generated by adding noise to each test utterance. Various types of noise are considered, including: 1) the stationary narrow-band noise, with a bandwidth of 100 Hz and different central-frequencies, i.e. 900, 1800, 2700 and 3500 Hz, respectively; 2) time-varying narrow-band noise, with a bandwidth of 100 Hz and a time-varying central-frequency which changes from 900 to 1800 Hz and then to 2700 Hz during the utterance; and 3) real-world noises including the sounds of a ding, a telephone ring and a laser, extracted from the sound files "ding.wav", "ringin.wav" and "laser,wav", respectively, provided in the Windows NT OS. It can be found that both the ding and telephone ring included multiple narrow-band components, and the laser included one dominant narrow-band component with both the bandwidth and central frequency being time-varying [3][5]. The entire speech frequency band is divided into 5 sub-bands (i.e. $N = 5$). Table 1 presents a summary of the performance of the union model, with an order $M = 2$, over all the above test conditions. For comparison, we also included the performance of the product model given in (1), which generates the likelihood of the observation by simply taking the product of the individual sub-band likelihoods.

Table 1 indicates that the union model offers a significant improvement over the product model throughout all the noisy test conditions. The union model reduced the error rate by an average of 57.1% in comparison to the product model.

Table 1: Summary of performance of the union model for speech recognition involving partial, unknown, time-varying frequency-band corruption and comparison with the product model

Noise condition	SNR (dB)	Union model (%)	Product model (%)
Clean		84.7	87.0
Stationary narrow-band	10	81.9	58.3
	0	77.2	40.7
Time-varying narrow-band	10	81.2	44.1
	0	72.8	17.3
Ding	10	74.9	34.9
Phone ring	10	65.1	37.8
Laser	10	71.9	36.2
Average		76.2	44.5

4. APPLICATION TO SPEECH WITH PARTIAL TEMPORAL CORRUPTION

In speech recognition, a speech utterance may be represented by a time series of short-term spectral vectors (i.e. frames). By partial temporal corruption we mean that some of the frames are noisy. This can result when a shut door, a channel impulse, or any type of burst noise occurs during the utterance. As described in Section 1, we tackle this problem by using a multi-segment analysis method, in which each frame sequence is divided into N consecutive segments, forming the feature set $o = \{o_1, o_2 \ldots, o_N\}$, where each o_n corresponds to a segment. Thus, we deal with speech recognition given that some of the o_n's may be noisy, but without information about where the noise occurs. The union model is employed to select the segments from the given segment set for recognition. This has been implemented by incorporating the union model (4) into an HMM framework [2][4].

Experiments are based on the TiDigits database, containing the speech data of connected digit sequences from 225 adult speakers (111 male and 114 female) for speaker-independent recognition. From this database the isolated-digit parts were extracted for the tests. This includes eleven isolated-digit words: "one" to "nine", "zero", and "oh". The noisy data were generated by adding noise into different positions of each test utterance, in particular, the beginning, the middle and the end of each utterance. The duration of the noise is measured in percentages relative to the duration of the speech. The noise is controlled for each utterance so that all utterances are corrupted to an equal percentage. Fig.1 shows some examples of the noisy speech data dealt with in recognition, involving 20%, 30% and 40% temporal corruptions, respectively. Different types of corrupting noise are considered, which include the white noise, and some real-world noises including a ding, a door slam, a telephone ring and a gunshot. In the experiments, each utterance is divided into 10 segments (i.e. $N = 10$), and a union model with an order $M = 3$ is used. Fig.2 shows the average performances of the union model and product model (which is equivalent to a baseline HMM in this case) subject to the white noise corruption, as a function of the SNR and noise duration. Similar relationships have been obtained for those real-world noise cases [4].

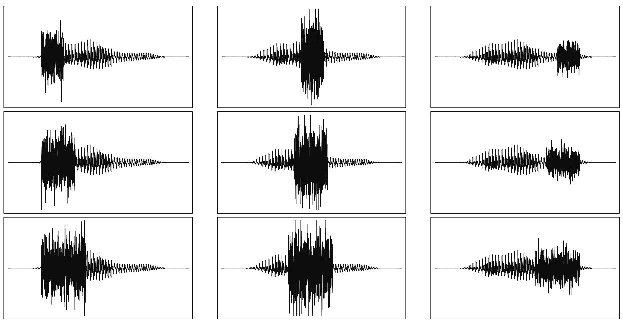

Figure1: A speech utterance, "one", with 20%, 30% and 40% (top to bottom) corruption by white noise at the beginning, middle and end (left to right), respectively.

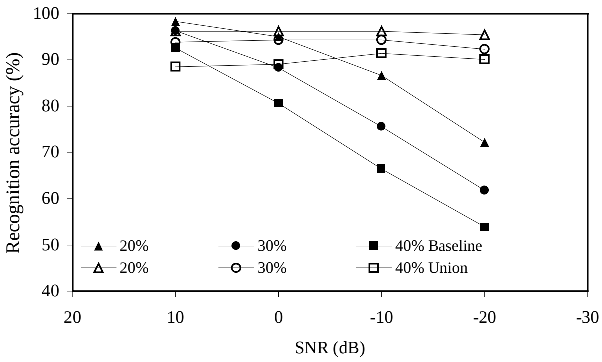

Figure 2: Average performances of the union model and baseline HMM with white noise, as a function of the SNR and noise duration shown in percentages relative to the speech duration.

As shown in Fig.2, the recognition accuracy was affected by a number of factors, particularly the SNR and the duration of the noise. While the baseline HMM was sensitive to both factors, the union model showed strong robustness to the variations of SNR, which thus was affected only by the duration of the noise. We can see that there was no significant performance degradation for the union model as the SNR was decreased.

5. APPLICATION TO PARTIAL FEATURE COMPONENT CORRUPTION

Consider the feature set $o=\{o_1, o_2 \ldots, o_N\}$ represents N different feature streams. Then we may use the union model to select a sub-set of features that provide the most discriminative information. In particular, we apply this to the selection of static spectral features and dynamic spectral features. Because the static features are more sensitive to background noise than the dynamic features, they should play a less significant role if they are affected. However, this is difficult to decide without knowledge about the environment (i.e. clean or noisy). This uncertainty can be dealt with by using the union model.

We have included this feature selection union model into the sub-band union model described in Section 3, for modeling both band corruption and feature-stream corruption. Experiments are based on the TiDigits database, corrupted by stationary narrow-band noise of the same type as in Section 3, and by some wide-band noise (e.g. pub and railway station). Table 2 presents the results, showing the performance of the sub-band union model with and without feature selection. This feature selection is particularly significant for the tests without endpoint detection, as in these cases the static features corresponding to the silence parts before and after each utterance will contain pure noise.

6. SUMMARY

This paper provided a summary of our recent work on the use of the probabilistic union model for speech recognition subjected to partial unknown corruption in frequency band, in time duration and in feature steams. We have conducted isolated-word recognition experiments using both the TiDigits database and BT E-set database. The results have shown the great potential of the new model for dealing with partial, unknown, time-varying noise. Current work is focused on the application of the new model to continuous speech recognition.

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

7. REFERENCES

[1] Ming, J., and Smith, F.J. “Union: a new approach for combining sub-band observations for noisy speech recognition,” *Robust’99*, pp. 175-178.

[2] Ming, J., Stewart, D., Hanna, P. and Smith, F.J. “A probabilistic union model for partial and temporal corruption of speech,” *IEEE Workshop on ASRU*, pp. 43-46, 1999.

[3] Ming, J., and Smith, F.J. “A probabilistic union model for sub-band based robust speech recognition,” *ICASSP’2000*, pp. 1787-1790.

[4] Ming, J., Hanna, P., Stewart, D., Jancovic, P., and Smith, F.J. “Union: a model for speech recognition subjected to partial and temporal corruption with unknown, time-varying noise statistics,” to be presented in *EUSIPCO’2000*.

[5] Ming, J., and Smith, F.J. “Union: a new approach for combining sub-band observations for noisy speech recognition,” *Speech Communication*, to appear.

[6] Bourlard, H. “Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR,” *Robust’99*, pp. 1-10.

[7] Hermansky, H., Tibrewala,S. and Pavel. M., “Towards ASR on partially corrupted speech”. *ICSLP’96*, pp. 462-465.

Table 2: Summary of performance of the sub-band union model without feature selection / with feature selection

Noise condition	SNR (dB)	With endpoint detection		Without endpoint detection	
		Union model	Product model	Union model	Product model
Clean		–	–	98.4 / 99.4	99.2
Narrow band	0	87.3 / 90.3	65.4	47.4 / 86.2	25.1
Pub	10	81.0 / 84.1	83.5	37.1 / 69.9	38.5
Railway station	10	80.2 / 84.1	84.4	28.4 / 75.1	53.3