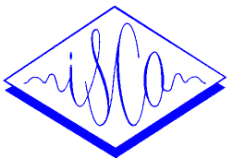


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

PROBABILISTIC COMPENSATION OF UNRELIABLE FEATURE COMPONENTS FOR ROBUST SPEECH RECOGNITION

Cyan L. Keung<keunglui@ust.hk>, Oscar C.Au<eeau@ust.hk>,
Chi H. Yim<eeyim@ust.hk>, Carrson C. Fung< c.fung@ieee.org>

Department of Electrical and Electronic Engineering,
Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

ABSTRACT

Missing feature theory is well studied in robust ASR context, many works have been done on additive noise of different colors. These are based mainly on classical spectral subtraction and marginal density techniques. This paper addresses the problem of temporal distortion of feature components, that is all about time domain instead of frequency one. No specific noise model and extract computation needed. We showed that the digit words recognition rate is above 95%, given test samples are clean with 10dB white noise added to middle 30% portion of speech along the time axis.

1. INTRODUCTION

Missing feature topics are interesting, people investigated many different kinds of noise and tried to develop a series of classification rules to determine whether the input features are clean or noisy or in-between. If they are noisy, the distorted group of feature vectors are assumed distributing uniformly[1] or some other forms we familiar with. Then obtain the marginal density of the whole distorted group of feature components by integration. Not much better than classic spectral subtraction and HMM adaptation techniques[2], white noise in any case cannot be improved further. Some different ideas in recent years like Class probability imputation[3], certain and uncertain factorization techniques[4] are proposed Temporal correction techniques[5] can be found in weighted viterbi algorithms paper[6] but we have a different approach here. In this paper, a compensation method target for improvement of temporal distorted speech recognition task based on the framework of Hidden Markov Models (HMMs) continuous density Gaussian mixture model is tested. Noise added speech information, if not badly damaged, still contain useful information in contributing to recognition process. Classification of reliable and unreliable part of speech data is not an issue here because we assume it is known and given in recognition.

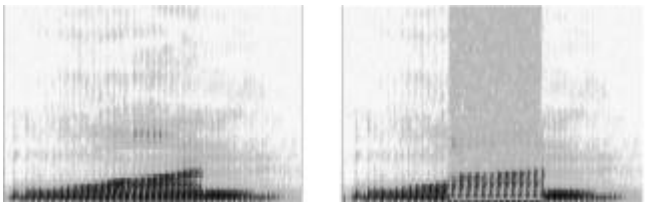


Fig 1. Clean and temporal distorted speech 30% in middle.

Trust worthiness of any given piece of observation is fed from external. We ‘tell’ the recognizer when to treat the observations reliable and when to treat them less-reliable from time to time. Such an indicator could be the field strength signal directly reported by a radio receiver or whatever system giving update of SNR estimate at real-time. We defined a cost function which is a variant of log likelihood function, weighted by a factor. It is continuous and bounded by 0 and 1. If the signal-to-noise-ration (SNR) of the observation is high enough, we treat it absolutely reliable and then set it to 1. If SNR increased, decrease it to weight the distorted observation less, anyway it cannot fall beyond zero.

2. TEMPORAL COMPENSATION

Non-stationary or impulse of noise happens very often in real life acoustic or telecommunication environment. If we abort everything in-between the noisy region of speech data, we lose information. If count them all, they may upset the recognition very badly. A compromised way to deal with the transition between clean and very noisy data is to assign a weight factor to certain observation vectors we sure they are reliable or not. This is done by modifying the log likelihood function, multiply by a factor and we call this a cost function.

2.1 Log likelihood and Cost function

Recall the conventional log likelihood function which is modeled as a mixture of *K* multivariate Gaussian densities, *w* is the mixture weight:

$$\log f(x \mid S_j) = \log \sum_{i=1}^{K} w_{ij} N(x, m_{ij}, \Sigma_{ij}) \qquad (1)$$

x is observation vector at state *j*
Gaussian density:

$$N(x, m, \Sigma) = \frac{1}{(2p)^{0.5n}\sqrt{\Sigma}} \exp\left[-0.5(x-m)^T C^{-1}(x-m)\right] \qquad (2)$$

We introduce that gives a cost function *g*.

$$g(x \mid S_j) = a \cdot \log f(x \mid S_j) \qquad (3)$$

In probability computation, just replace the conventional log likelihood *f* by this cost function *g*. changes all along the time.

3. EXPERIMENTS AND RESULTS

3.1 Setting up

TIDIGIT corpus single digit utterances were used for training and testing. White noise was carefully adjusted to desired magnitude and individually mixed using CoolEdit package. Endpoint detection algorithm[7] is used to accurately remove the silence portion of every sample. Clean, 20dB and 10dB test set are tested to verify the methods we proposed. Feature sets are all MFCC-39 with 12 DCT coefficients, 12 Deltas and 12 accelerations plus energy, delta and acceleration energy. Recognition engine is custom programmed to fit our needs, recognition results are almost identical to HTK[8] in the case of conventional log likelihood computation.

3.2 External signal of reliability

We created all test set ourselves, we signal the recognizer all information it needs say, the SNR and the exactly time location of the noise added.

3.3 Results

Baseline is done using chopped version of speech sample, that is simply done by removing the middle 30% portion along time axis and merge the head and tail together. Ignore any part of observation up to 50% does not upset the performance very much but take noisy data into count can worse off the overall performance. Isolated sngle digit recognition task itself is easy, random guess gives 1/11 (9%) accuracy.

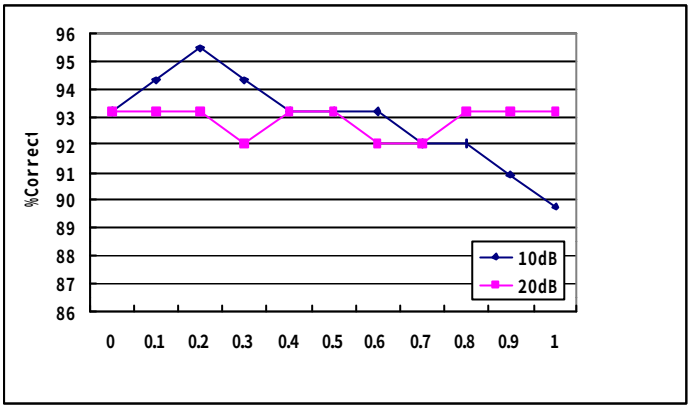
	%Correct		
	Clean	20dB	10dB
100%	97.7	89.8	57.9
60%	N/A	90.5	71.1
30%	N/A	93.1	89.7

Table 1. Recognition accuracy on temporally distorted speech samples up to 100%, 60% and 30% of time line in middle.

	%Correct
0%	97.7
30%	92.3
60%	89.7

Table 2. Baseline, the recognition accuracy on chopped speech samples, 0% 30% and 60% chopped away in middle.

Fig 2. Temporal compensation results on 10dB and 20dB cases.



92.3% accuracy is obtained from baseline experiment, we can see temporal compensation boosts up the accuracy in both 10dB and 20dB cases.

4. CONCLUSIONS AND FUTURE WORKS

Viterbi and forward algorithm are both tested in recognition, results are almost the same, difference is unnoticeable.

Temporal feature compensation outperformed baseline results, the best recognition rate in 10dB case is above 95%. Unfortunately it does not work obviously in 20dB case.

There might be explicit relationship between speech SNR and the variable , it has to be found out in the coming papers.

In this paper, feature compensation only involved Gaussian mixture density function, compensate also HMM transitional probability matrix[2] is very likely to do the job better.

Isolated single digit recognition is not very useful and practical in real-life, connected digit string and phone-base recognition has to be investigated with this technique also.

5. REFERENCES

[1] Philippe Renevey and Adrzej Drygajlo "Missing Feature Theory and Probabilistic Estimation of Clean Speech Components for Robust Speech Recognition" Eurospeech' 99.

[2] J.A.Nolazco Flores and S.J. Young "Continuous Speech recognition in Noise using Spectral Subtraction and HMM adaptation" ICASSP 94 Vol1 pages 409-412

[3] MartinCooke, Andrew Morris and Phil Green "Missing Data Techniques for Robust Speech Recognition" ICASSP 97 Vol2 pages 863-866

[4] Andrew C.Morris, Martin P. Cooke, Phil D Green "Some solutions to the Missing Feature problem in Data Classification with Application to Noise Robust ASR" ASSP98 Proceedings 1998 IEEE intenational conference Vol2 pages 737-740

[5] N.B. Yoma, L.L. Ling and Sandra Dotto Stump "Temporal constraints in viterbi alignment for speech recognition in noise" Eurospeech' 99

[6] N.B. Yoma, F. McInnes, and M.Jack. "Weighted viterbi algorithm and state duration modelling for speech recognition in noise" ICASSP' 98 pages 709-712

[7] L. F. Lamel, L. R. Rabiner, A. E. Roserberg and J. G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", *IEEE ASSP*, Vol. 29, No. 4, pp 777 – 785, Aug. 1981.

[8] S.J. Young and P.C. Woodland. "HMM Toolkit User reference and programmer manual" Cambridge University Engineering Department, 93