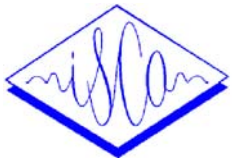

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

WITHIN-UTTERANCE CORRELATION FOR SPEECH RECOGNITION

Mats Blomberg
Dept. of Speech, Music and Hearing
KTH
Stockholm, Sweden
mats@speech.kth.se
http://www.speech.kth.se/~mats

ABSTRACT

Relations between non-adjacent parts of an utterance are commonly regarded as an important source of information for speech recognition. However, they have not been very much used in speech recognition systems. In this paper, we include this information by joint distributions of pairs of phones occurring in the same utterance. In addition to relations between acoustic events, we also have incorporated relations between spectral and prosodically oriented information, such as phone duration, position in utterance and fundamental frequency. Preliminary recognition results on N-best rescoring show 10% word error reduction compared to a baseline Viterbi decoder.

1. INTRODUCTION

The currently dominating technique for speech recognition is Hidden Markov modelling (HMM). Large vocabulary, real time recognition systems with high accuracy have been developed using this technique. One limitation, however, is its fundamental assumption of independence between observations. The consequence for recognition is that every time interval in the speech signal is regarded as statistically independent of other parts of the utterance. It is well known that this assumption is incorrect and that there exist strong relations within an utterance as a consequence of time-invariant or slowly changing characteristics of the speaker and the acoustic environment. One example of the non-realistic consequences of the independence assumption is that a speaker-independent recognition system can be regarded as assigning models from different speakers to different parts of an utterance. This problem is avoided by speaker-dependent training or adaptation of the phone models. However, the pronunciation variability of a single speaker needs to be modelled as well. The desired procedure to train speaker-dependent phone models is to include all the expected speaker variability in the training data. Similarly to the speaker model problem in the speaker-independent case, the independence assumption allows mixing several speaker-state models in a single utterance. Yet, within this interval, the health condition and the emotional state of the speaker can be regarded as constant. Avoiding this problem by training health-condition-dependent or emotional-state-dependent models for each speaker is not practically possible. Furthermore, since the goal for recognition is to identify the linguistic information in the utterance, it is not necessary to decode the speaker state as part of the recognition process. The purpose of the work in this report is to study if the use of within-utterance relations can improve the performance and robustness of speaker-independent recognition.

Certain aspects of within-utterance relations are already used to improve the recognition performance in current systems. Short-distance relations are commonly used by incorporating the first and second order time-differentiated acoustic feature vectors. However, the optimum size of the differentiation window is only around the average duration of a phoneme. Longer-range dependencies cannot be captured since variation in speech rate and variability in the phonetic context will generate large variation in the acoustic events covered by the window.

Another standard technique is to account for co-articulation effects from the immediate phonetic neighbourhood by expanding the phone inventory to context-dependent phone models. The most common units are diphones and triphones. In theory, extending the size of the phonetic context will include the effects from more distant phonemes. This is not a practical solution, though, because of the rapidly increasing number of such units.

Previous work to account for correlation between non-adjacent speech segments has been reported by [1], who studied different approaches to exploit

within-speaker correlation for vowel recognition. Techniques for simultaneous speaker adaptation and recognition have been reported by [2], [3], and [4]. The current report extends previous work in [5] and deals with aspects on using long-distance relations for speech recognition. Examples of such relations will be given. Recognition techniques under development are discussed and initial recognition results will be presented.

2. WITHIN-UTTERANCE RELATIONS

The relations between observations in an utterance arise from different sources. The obvious relations are between the acoustic observations. However, there is also inter-relation between acoustic and non-acoustic events on the phonological and linguistic levels as well as within the non-acoustic information. These three types are discussed below.

2.1 Between acoustic events

The invariant properties of the speaker's physical and behavioural characteristics give rise to relations within the acoustic realisations of his/her phonetic inventory. For example, formant frequencies are closely dependent on the vocal tract length. This results in a strong relation between the formants in a certain vowel and also between different vowels spoken by the same speaker.

Two or more occurrences of the same phone identity in an utterance are very likely to have similar spectral properties. The relation strength is reduced compared to that of within-phone due to difference in phonetic and prosodic context. Two non-identical phones belonging to the same broad phoneme class are also related.

Also, non-speech phenomena, such as environmental additive noise and transmission characteristics, result in within-utterance correlation. Provided that the training data contain these effects, compensation can, in principle, be performed in the same way as for other factors that generate within-utterance correlation. However, there already exist techniques for compensating for these and it is uncertain if correlation analysis can provide better performance.

2.2 Between acoustic and non-acoustic information

Certain types of dependence between acoustic and non-acoustic information might also be expressed in the same way as between purely acoustic observations. The requirement is that the non-acoustic information is numerically represented. By including such information in the covariance matrix, the relations to the acoustic parameters can be utilised. In this way, it is possible to capture linear relations between the lexical and the acoustic domains. For example, lexical stress and lexical duration might be represented by numerical values instead of phonetic features.

Aspects of utterance time have important influence on the pronunciation. One example is the utterance position of a phone. Normally, speech amplitude slowly decreases and voice quality gradually changes during the utterance. Another example is vowel reduction [6], in which the dependence of phone duration upon steady-state formant frequencies is expressed. These relations can be analysed by incorporating utterance position and duration to the acoustic features of a phone.

2.3 Within non-acoustic information

There are also relations within the non-acoustic information, such as co-occurrence of words, and between different phonological rules, corresponding to the assumption that a speaker is consistent in speaking style. Clarity of articulation might also be quantified into a number of levels. If the pronunciation alternatives of each word in the lexicon were assigned a clarity value, then correlation data could be used to favour those recognition hypotheses that are consistent in their pronunciation alternatives throughout the utterance.

3. COVARIANCE REPRESENTATION

In this report, phone pair relations are represented by a single mixture, full covariance, joint Gaussian distribution. A better unit for expressing within-utterance relations would be the triphone pair. However, it is practically impossible to collect a training corpus that contains a sufficient number of observations of all triphone pairs. Phone pairs are fewer and easier to train, but their discrimination power is lower. In order to avoid training a large number of triphone pairs but still to capture existing correlation, we apply an intermediate approximation between triphone pair and phone pair covariance. Triphone pair covariance matrices are tied with respect to the central phone identity, i. e., the frequency weighted average of the individual covariance matrixes of all triphone pairs that have the same mid phone identity. Note that this covariance matrix is different from the corresponding context-independent phone pair matrix.

4. RECOGNITION APPROACH

The true statistical distribution of any sentence hypothesis is in general unknown, since it has normally not been pronounced during training. Therefore, an approximate estimate has to be made. One possibility is to assemble the covariance matrix belonging to the current sentence hypothesis from trained phone-pair matrices. The dimensions of the matrix are the acoustic features for each of the recognised phones. The size of this matrix is quite large, which makes the search procedure computationally very heavy. A typical sentence length of 30 phones and a feature vector size of 24 elements results in a matrix size = 720 * 720. The processing time is too long for practical use and it is therefore necessary to reduce the matrix size. One possibility would be to lower the number of acoustic features per phone, e.g. by using the first few principal components. To reach reasonable response time in a longer sentence, it is necessary to limit the size of the feature vector to 2 or 3 elements per phone, which is too few even with principal component vectors.

4.1 Separate phone-pair observations

One way to avoid the large size of the utterance covariance matrix is to approximate the joint probability of all hypothesised phones in the utterance by the product of a number of the likelihood estimates of all combinations of observations of separate phone pairs [3]. One large matrix operation is replaced by a large number of operations on smaller matrices, which, in this case, requires much less time. This technique was adopted for the recognition experiments.

5. EXPERIMENT

Currently, the Waxholm dialog system database [7] is used for experiments. It contains around two hours of spoken dialogues from 66 speakers; 49 male and 17 female. Out of these, 56 subjects were selected for training. In this corpus, the most frequent within-utterance phone pairs have around 6000 occurrences. Phone pairs with less than 200 training occurrences were excluded from testing.

In the initial recognition experiments, the cepstral coefficients C0 through C8 have been chosen as spectral features. Two representative time samples of each parameter trajectory in each phone are selected as acoustic observations in the covariance matrix [8]. In addition to the spectral parameters, three non-spectral parameters have been included in the matrix: fundamental frequency, phone duration and time position in the utterance.

Recognition experiments have been performed by rescoring N-best sentence candidate lists produced by the speaker-independent recogniser used within the WAXHOLM project [9]. The used N-best lists contain 10 candidates on average. In order to allow correct recognition, the true sentence identity is added to the list if necessary and out-of-vocabulary words are added to the lexicon. The candidates in the N-best list are exposed to a second, more detailed Viterbi search including word boundary triphones, expanded optional phonological rules and dynamic voice source adaptation [8]. The resulting phone segments are then evaluated using the within-utterance phone pair covariance method. The final score is a weighted sum of the log likelihood scores from the second Viterbi search and the phone pair joint distributions. The weight was manually adjusted for highest performance. For reference, a baseline system gives zero weight to phone pair covariance scores.

6. RESULTS

The rescoring results with different combinations of methods and feature vectors are shown in Table 1. The phone pair covariance method performs worse than the baseline Viterbi decoder. A detailed analysis of the errors shows that the number of word deletions is doubled, while the number of insertions is relatively constant. This effect could be expected from the sparse sampling of the utterance. Combining the scores of the two methods lowered the number of errors by 10% compared to the baseline result. Adding fundamental frequency, phone duration and phone position in the utterance to the cepstral features did not change the result.

Table 1. Word error rate with different types of correlation information. B: Baseline, PP_C: Phone pair with cepstra, B+PP_C: Combined Baseline and Phone pair with cepstra, B+PP_CFDT: Combined Baseline and Phone pair with cepstra, F0, duration and time.

Method	Word Error Rate	Error Reduction
B	13.0%	---
PP_C	16.4	---
B + PP_C	11.7%	10%
B + PP_CFDT	11.7%	10%

7. DISCUSSION

The current study shows one way to utilise relations between different parts of an utterance as well as between acoustic and non-acoustic levels of information. A modest improvement has been achieved when combined with the baseline Viterbi decoder. This is encouraging considering the lower performance of the method by itself. If it can be improved, then higher performance can be expected of the combined techniques as well.

Incorporation of fundamental frequency, phone duration and time position in the utterance did not improve the result. It is possible that the prosodic features require information from detailed semantic and linguistic analysis to give a positive contribution to the recognition performance. Another contributing factor to the lack of improvement might be tracking errors in the unsupervised pitch estimation procedure. Still another possibility is that a formant or articulatory oriented representation would benefit more from this approach due to assumed higher linearity in the relations between those parameters and phone duration.

One limitation with the current implementation is the sparse time sampling of the utterance. In this report, the lack has been compensated for by the combination with frame-wise matching by the Viterbi search procedure. Another problem with the choice of two samples per phone is that these points are positioned near the phone boundaries and are, thus, sensitive to coarticulation with neighbouring phones. For these reasons, it may be possible that mid-point sampling should be included. Also, the single-mixture distributions should be replaced by multi-mixture components, which would give a better approximation of the distribution of the tied triphone pair models.

Acknowledgement

This work was performed within the Swedish competence centre CTT, Centre for Speech Technology.

REFERENCES

[1] Niyogi, P. and Zue, V. W. 1991. Correlation analysis of Vowels and their Application to Speech Recognition. Proceedings of Eurospeech91, Genova, Italy, 1253-1256.

[2] Leggetter M. J., Woodland P. C. 1995 Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9, 171-185.

[3] Hazen T. J. and Glass J.R. 1997. A comparison of novel techniques for instantaneous speaker adaptation. Proceedings of Eurospeech97, Rhodes, Greece, 2047-2050.

[4] Ström, N. 1996. Speaker Adaptation by Modeling the Speaker Variation in a Continuous Speech Recognition system. Proceedings of International Conference on Spoken Language Processing, ICSLP 96, 989-992.

[5] Blomberg M. 1998. Speech recognition using long-distance relations in an utterance. Proc. FONETIK 98, Dept. Linguistics, Stockholm University, 166-166 (Abstract. The complete article is available on Internet: http://www.speech.kth.se/~mats/papers/fon98mb.ps).

[6] Lindblom, B. 1963. Spectrographic Study of Vowel Reduction. Journal of the Acoustic Society of America. Vol 75, 945-951.

[7] Bertenstam J. Blomberg M. Carlson R. Elenius K. Granström B. Gustafson J. Hunnicutt S. Högberg J. Lindell R. Neovius L. Nord L. de Serpa-Leitao A. Ström N. 1995. Spoken dialogue data collection in the Waxholm project. STL-QPSR, KTH, 1/1995, 50-73.

[8] Blomberg M., Elenius K. 1996. Creation of unseen triphones from diphones or monophones usinga speech production approach. Proceedings of International Conference on Spoken Language Processing, ICSLP 96, 2316-2319.

[9] Ström, N. 1996. Continuous speech recognition in the WAXHOLM dialogue system, TMH-QPSR 4/1996, Speech, Music and Hearing, KTH, 67-95.