

EXPLICIT MODELLING OF COMMON ACOUSTIC FEATURES FOR CHARACTER RECOGNITION

Mario E. Munich[†] and Qiguang Lin[†]

Evolution Robotics
Pasadena, CA 91103
mariomu@vision.caltech.edu

AOL Voice Services
Mountain View, CA 94043
QiguangLin@aol.com

ABSTRACT

This paper presents a novel approach for robust, isolated character recognition. A major challenge of character recognition is that some characters are acoustically confusing and that no language modeling can be resorted to resolve the confusion. In the proposed approach, we attempt to explicitly model the common acoustic structures among different, confusing characters through state tying. As a result, decoding decision is made only by states modeling distinct sound segments. We first describe the training procedure of the new approach, then present recognition results from three character databases. Compared with the baseline system (which is a whole word/character model), the new approach is 45% better when evaluated using true telephone speech.

1. INTRODUCTION

Character recognition is an important sub-problem of speech recognition that arises in recognition of spelled words. Conventional speech recognizers fail to provide correct recognition results for words that are not in the vocabulary. The natural alternative in these cases is spelling of the word. Hence, character recognition becomes essential for many real-world applications that deal with arbitrary names of addresses such as e-commerce, automated directory assistance, or database search over the telephone.

Several groups of characters are highly confusing, making character recognition particularly challenging. The E-set (B, C, D, E, G, P, T, V, Z) that share the /iy/ sound, the A-set (A, J, K) that share the /ey/ sound or the M-N and F-S pairs are examples of such groups. Telephone speech makes the recognition task even harder due to the bandpass and time-varying nature of the telephone channel as well as the lower sampling rate used to capture speech. The difference among letters in the mentioned groups become so minimal that it is difficult even for humans to perform a reliable recognition. A perceptual experiment on telephone speech [4] carried out at the Oregon Graduate Institute (OGI) showed that human performance on spoken letter recognition ranged from 90 to 95%, with an average performance of 93%. This performance provides a standard of comparison for computer letter recognition results.

A number of approaches have been proposed for character recognition in the past [10, 5, 9]. Loizou and Spaniards [10] presented a speaker-independent, isolate alphabet recognition system based on context-dependent phoneme HMMs. The system used the OGI spelled and spoken telephone alphabet corpus [1] for training and evalua-

tion. They achieved 85% recognition rate on the alphabet and 71.38% recognition rate on the E-set. Hamaker et. al. [5] used context-independent syllables for alphabet recognition. The system yielded an error rate of 12.8% on the alphabet portion of the OGI Alphadigits corpus. Karnjanadecha and Zahorian [9] proposed non-conventional signal modeling for alphabet recognition. They achieved 89.6% recognition rate on synthetic telephone speech, obtained by adding noise and band limiting utterances from the ISOLET database. Several systems [4, 6, 8] based on time-delay neural networks and on HMM's have also been proposed for recognition of spelled names over the telephone.

This paper presents a novel approach for robust, isolated character recognition. Conventional recognition systems use whole-character models. Our system explicitly models common acoustic structure among different, confusing characters through state tying. As a result, utterance classification is made only by states modeling distinct sound segments. The models obtained in this way would be shown to be more robust and accurate than conventional ones.

The paper is organized as follows. Section 2 describes our character recognition system. Experimental results are shown on section 3 and conclusion are drawn on section 4.

2. CHARACTER RECOGNITION

We investigated the character recognition performance of three different systems. The first one is the baseline system, that follows conventional whole-word modeling techniques. The second one includes silence models that are shared by all characters. The third system uses state-tying to model common acoustic structures across groups of characters.

2.1 System I: whole-character models (baseline system)

Conventional approaches for isolated word recognition [11, 2] use a model per word in the vocabulary. Our baseline system follows this approach by using whole-character models, as shown in figure 1. Each model is trained independently with the Baum-Welch algorithm. Classification of a test utterance is done by computing the models' likelihood scores and choosing the model with highest likelihood.

The baseline system is simple and easy to build, but presents a noticeable drawback. Portions of the models that represent similar acoustic content differ due to the context-dependent nature of the training. Figure 2 shows one utterance from the TI46 database. This utterance is composed by three different segments: an initial silence portion, a speech portion, and an ending silence portion. All silence segments in all utterances have similar acoustic characteristics; therefore, they should be statistically indistinguishable from the

[†] The major part of the work was performed while the authors were with VocalPoint, Inc., San Francisco, CA

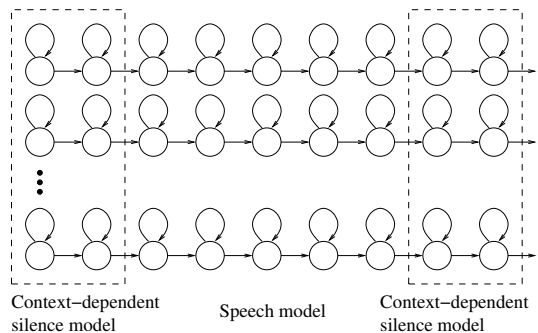


Figure 1: Conventional isolated character recognition models. The silence portions of the models are context-dependent since each model uses a different set of examples for training.

recognition point of view. However, given that different training sets are used to train each character model, the silence portion of the models would differ.

The computation of the likelihood score used for classification is performed with Baum-Welch algorithm. This technique performs a soft correspondence between states of the model and frames of the utterance. Character models whose silence portions differ may generate different state alignments and may produce different likelihoods values for the corresponding silence parts. Therefore, the overall likelihood used for classification would account not only for the difference between speech models but also for the difference between silence models, leading to possible misrecognition.

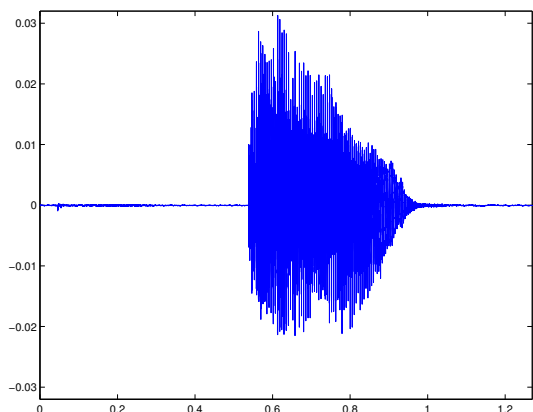


Figure 2: Example utterance from the TI46 database. The utterance can be divided in three clearly differentiable parts: an initial silence segment, a speech segment, and an ending silence segment.

2.2 System II: context-independent silence models

The baseline system could be improved by training the silence parts of the models in a context-independent way. In other words, the silence parts of the models would be trained using all data from all utterances, while the speech part of the model would be trained using only the corresponding training set. Figure 3 shows the topology of the resulting models. The differences in classification likelihoods in this case would correspond only to differences between the speech models. The technique of sharing sub-models across various

models is known as state-tying in the speech recognition literature [7] and has been successfully applied in continuous speech recognition systems. The complete network of figure 3 amounts to nothing more than a larger HMM. The system needs to segment all utterances into silence and speech parts in order to perform training. Once again, the Baum-Welch algorithm provides the soft-segmentation of the utterances, leading to a maximum likelihood estimation of the silence and speech models.

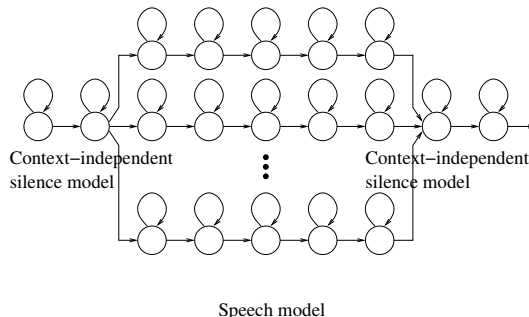


Figure 3: Context-independent silence models are shared by all character models.

The training algorithm works as follows. Given a desired topology for the silence and speech models like the one presented in figure 3, the Baum-Welch algorithm provides the probability of a given frame of speech being produced by a particular state of the model. This computation is carried out only using the speech model that corresponds to the utterance's transcription. This probability indicates a soft-assignment between states and frames that is used to reestimate the parameters of the models. This probability and the corresponding feature vector would be accumulated to perform the re-estimation. Since the silence models would be shared by all character models, all the frames in all the utterances, along with the corresponding probabilities for the states of the silence models, would be used to reestimate the parameters of these models. Only the utterances corresponding to a particular speech model would be used to reestimate the parameters of this speech model.

2.3 System III: State-tying for groups of confusing letters

As mentioned before, some groups of characters share acoustic content, such as the /iy/ sound in the E-set or the /ey/ sound in the A-set. Therefore, the corresponding models obtained in section 2.2 are refined by tying a few states of the speech models, as shown in figure 4. As a result, classification within the group is made only by states modeling distinct sound segments. Re-estimation of the parameters of these new speech models is performed using all utterances from a particular group of characters, while the silence models are kept constant. State-tying is performed in the back portion of the speech model for some groups of characters, e.g., E-set and A-set, while state-tying is done in the front part of the speech model for some other groups of characters, e.g., M-N and F-S pairs. The number of states to tie in each group is selected experimentally.

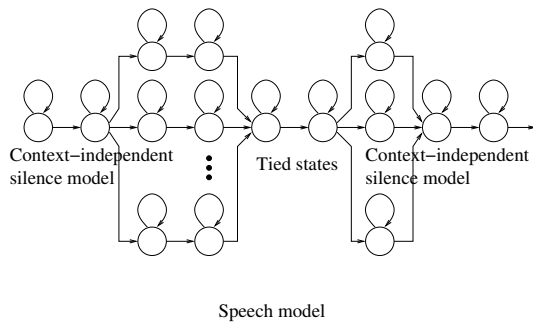


Figure 4: State-tying for groups of letters. Common acoustic structure among various characters would be represented with the same sub-model.

3. EXPERIMENTS

Three different speech corpora were used in the experiments. The first one was the publicly available TI46 word speech database. The second one is a modified version of the TI46 database such that it resembles telephone bandwidth. The third one was a proprietary telephone speech corpus collected at VocalPoint Technologies.

3.1 TI46 database

The Texas Instruments 46-Word Speaker-Dependent Isolated Word Corpus (TI46) is a database of speech which was originally designed and collected at Texas Instruments, Inc. (TI) in 1980, and used initially in performance assessment tests of isolated-word speaker-dependent technology (see [3]). The speaker-dependent condition of the corpus arises from the fact that utterances from the same subjects were placed in both the training and test sets.

The 46-word vocabulary consists of two sub-vocabularies: (1) the TI 20-word vocabulary (consisting of the digits zero through nine plus the words "enter", "erase", "go", "help", "no", "rubout", "repeat", "stop", "start", and "yes", and (2) the TI 26-word "alphabet set" (consisting of the letters "a" through "z"). Only the latter sub-vocabulary was used in the experiments presented in section 3.5. The "alphabet set" corpus contains read utterances from 16 speakers (8 males and 8 females). Each speaker provided 26 examples of each vocabulary word, 16 examples were assigned to the training set and 10 examples were assigned to the test set. The corpus was collected at Texas Instruments in a quiet acoustic enclosure using an Electro-Voice RE-16 Dynamic Cardioid microphone at 12.5kHz sample rate with 12-bit quantization.

3.2 Transformed TI46 database

As mentioned before, telephone speech increases the challenge of the recognition task, so the TI46 database was modified in order to simulate telephone speech. The original TI46 sampling rate of 12.5kHz was converted to 8kHz by filtering the original data with a 10-tap Kaiser-window based low-pass FIR filter and by re-sampling it using a polyphase implementation. The band-pass characteristic of the telephone channel was simulated by filtering the re-sampled data with a Butterworth filter of order 4 and cut-off frequencies of 280Hz and 6.6kHz.

3.3 VPT database

The third corpus used in the experiments is a database of speech data collected at VocalPoint Technologies. The database reflected data acquired in a variety of conditions. The data was actual telephone speech captured evenly from both an analog phone line and a digital phone line. Two different phones were used, the first one was a cordless phone that was attached to the public telephone network while the second one was a cellular phone attached to the cellular telephone system. Two-thirds of the data were captured with the cordless phone while the other third was obtained with the cellular one. Each subject provided ten examples of each character.

Two groups of subjects were used for this corpus. The first group was composed of 50 subjects. The corresponding utterances were separated into a training set and a test set following the speaker-dependent characteristic and the data proportions of the TI46 dataset in order to make a fair comparison; out of the ten utterances per character provided by each subject, six were assigned to the training set and the remaining four to the test set. The second group was composed of 16 subjects. All the utterances of the second group were used for testing the system in a speaker-independent setting. Both the speaker-dependent and the speaker-independent test sets have roughly the same number of utterances in order to compare results.

3.4 System implementation

The utterances were pre-emphasized with a first-order filter of the form $(1 - 0.97z^{-1})$. A Hamming windows was used to segment the utterances into frames. The window was 25 ms long and the frame rate was 10 ms. The short-time spectrum of each frame was computed using a 512-point fast Fourier transform (FFT). The short-time spectrum was filtered with a 24-channel triangular filter bank. Mel-frequency cepstrum coefficients were obtained by taking a 24-point discrete cosine transform of the 24 log-filter bank energies. The initial set of feature vectors was composed of 12 cepstrum coefficients and energy. First and second order differences of these 13 features were added to form a 39-dimensional feature vector. No cepstral mean normalization was used.

All character models were composed of 9-state left-to-right HMMs. The state output pdf was modeled with a mixture of Gaussian pdfs. The baseline system used sixteen mixtures per state. System II assigned the two beginning states and the two ending states to the initial and ending silence models. The five remaining central states were assigned to the speech model. Thirty-two mixtures per state were used for the silence models and sixteen mixtures per state were used for the speech models. System III tied the three ending states of the speech model for the E-set, the A-set, and the Q-U pair. The beginning state of the speech model was tied in the case of the M-N and F-S pairs.

Training of the models was conducted incrementally as follows. Each training utterance was divided in nine segments of equal length. A mixture of Gaussian was fit to all corresponding segments from each utterance. These mixtures were used to initialize the baseline system models. Four iterations of the Baum-Welch algorithm were used to obtain the final character models.

Context-independent silence models were built on top of the baseline system models. Each training utterance was di-

vided into states using Viterbi alignment. Speech segments from all utterances corresponding to the beginning two states of the models were grouped together. These segments are assigned to the initial silence model of system II. These speech segments were divided in two equal-length parts and a mixture of Gaussians was fit to each part. These mixtures provide the starting parameters for the Baum-Welch re-estimation. A similar procedure was used to obtain the starting parameters for the ending silence model and for the speech models. Four iterations of the Baum-Welch algorithm were used to obtain the final models.

System II models were used as the initial models for system III. The silence models were kept unmodified, while the speech models were re-estimated following the desired state-tied topology. Four iterations of the Baum-Welch algorithm were used to obtain the final models. Testing of system III was carried out as follows. All test utterances were classified with system II. Each utterance assigned to one of the groups of characters modeled by system III was re-classified within the group. For example, an utterance classified as “B” by system II would be re-classified using the system III models for the E-set. This re-classification provided the final decoding decision to be used in the computation of the error rates.

3.5 Experimental results

Recognition experiments were conducted in speaker-dependent and -independent scenarios. The error rates of the systems presented in section 2 are shown on table 1. We observe a clear improvement in performance when using state-tying. System III outperforms the baseline system by roughly 60% for speaker-dependent conditions and by 45% for speaker-independent ones.

	TI46	transf. TI46	VPT spk. dep.	VPT spk. ind.
I	3.24%	16.88%	18.27%	24.24%
II	2.22%	4.30%	10.46%	17.12%
III	1.40%	2.51%	7.10%	13.35%

Table 1: Performance comparison of the three systems presented in the paper.

Table 2 shows the difference in error rates obtained by state-tying of different sets of characters. We observe that the improvement in performance is much more pronounced for the speaker-dependent case than for the speaker-independent one. Still, the speaker-independent performance on the E-set is 30% better than the results presented on reference [10].

	TI46	transf. TI46	VPT spk. dep.	VPT spk. ind.
E-set	5.51%	9.98%	20.65%	29.90%
tied E-set	3.42%	5.02%	12.96%	19.89%
A-set	0.63%	0.63%	5.85%	7.05%
tied A-set	0.21%	0.00%	2.82%	6.59%
M-N pair	2.52%	3.14%	9.09%	11.15%
tied M-N	1.89%	2.52%	7.79%	11.15%
F-S pair	0.0%	5.97%	12.50%	17.78%
tied F-S	0.0%	6.60%	11.15%	17.45%
Q-U pair	0.31%	0.31%	2.60%	3.70%
tied Q-U	0.31%	0.31%	1.62%	3.70%

Table 2: Comparison of error rates with and without state-tying for different sets of characters.

4. CONCLUSIONS

This paper has presented a novel approach to character recognition based on tying states of the HMMs that correspond to common acoustic contents. The experiments have shown that our system outperforms whole-character models by roughly 60% in speaker-dependent conditions, and by 45% in speaker-independent conditions. The overall system performance is 11% better than the best performance presented in the literature [10] for speaker-independent scenarios.

A weakness of the proposed method is the lack of an automatic technique for deciding which states need to be tied in order to accurately represent common acoustic patterns. This decision is currently based on experimental results.

REFERENCES

- [1] R. Cole, K. Roginski, and M. Fanty. A telephone speech database of spelled and spoken names. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 891–893, 1992.
- [2] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-time Processing of Speech Signals*. Macmillan Publishing Company, 1993.
- [3] G.R. Doddington and T.B. Schalk. Speech recognition: Turning theory to practice. *IEEE Spectrum*, 18(9), September 1981.
- [4] M. Fanty, R.A. Cole, and K. Roginski. English alphabet recognition with telephone speech. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 199–206, New York, NY, 1992. Springer-Verlag.
- [5] J. Hamaker, A. Ganapathiraju, J. Picone, and J. Godfrey. Advances in alphasdigit recognition using syllables. In *Proc. of Int. Conf. Acoust. Speech and Signal Processing ICASSP98*, 1998.
- [6] H. Hild and A. Waibel. Recognition of spelled names over the telephone. In *Proceedings of the International Conference on Speech and Language Processing*, volume 1, pages 346–349, 1996.
- [7] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, 1997.
- [8] J-C. Junqua. Smartspell: a multipass recognition system for name retrieval over the telephone. *IEEE Trans. On Speech and Audio Processing*, 5(2):173–182, 1997.
- [9] M. Karnjanadecha and S. A. Zahorian. Signal model for isolated word recognition. In *Proc. of Int. Conf. Acoust. Speech and Signal Processing ICASSP99*, 1999.
- [10] P. C. Loizou and A. S. Spanias. High performance alphabet recognition. *IEEE Trans. Speech and Audio Processing*, 4(6):430–445, 1996.
- [11] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Inc., 1993.