

A SURFICIAL PRONUNCIATION MODEL

Eric Sven Ristad¹

Peter N. Yianilos²

¹Mnemonic Technology, Inc., Princeton, NJ, USA

²NEC Research Institute, Princeton, NJ, USA

ABSTRACT

We argue for a *surficial* pronunciation model: a model without underlying forms. The surficial model outperforms a traditional generative model by a significant margin on conversational speech (Switchboard) as well as on read speech (TIMIT). Our results suggest that the true mapping from underlying forms to surface forms is too complex to be accurately modeled using current techniques, and that we would be best served to model the surface forms directly.

1 INTRODUCTION

Following the information-theoretic approach pioneered by the IBM Speech Recognition group in the 1970's [7, 1], and lead by the generative phonology revolution [2, 6], the pronunciation models in modern speech recognition systems typically consist of a phonological lexicon coupled with a statistical transducer. The phonological lexicon maps each syntactic word to a small set of underlying pronunciations; it is typically designed by hand. The statistical transducer maps each underlying form to a larger set of surface variants; it is induced from data.

The original IBM speech recognition system generated its underlying forms using a hand-crafted pronunciation dictionary and a manually designed context sensitive grammar. Next, each segment in the underlying form was replaced by a stochastic finite state automaton, which generated surface variants for that segment.

Little has changed in the intervening decades. Current speech technology still employs a sparse pronouncing lexicon of hand-crafted underlying forms. When the vocabulary is large or contains many proper nouns, then the pronouncing lexicon may be generated by a hand-crafted text-to-speech system [8]. In most systems, the mapping from underlying forms to surface forms is left to the acoustic models. In more advanced systems, underlying segments are mapped to their surface realizations using a statistical decision tree [9].

In this paper, we argue for a *surficial* pronunciation model: a model without underlying forms. We demonstrate that a surficial model outperforms a generative model by a significant margin, both on conversational speech (Switchboard) and on read speech (TIMIT). Indeed, our surficial model is within 4% of the best performance achievable by any pronunciation recognition system on Switchboard. Our results suggest that the true mapping from underlying forms

to surface forms may simply be too complex to be accurately modeled using current statistical techniques. If so, then we would be best served to model the surface forms directly.

The remainder of this paper consists of four sections. Section 2 describes our pronunciation model. Section 3 presents recognition results on the Switchboard corpus of conversational speech. Section 4 presents recognition results on the TIMIT corpus of read speech. Section 5 concludes with some speculative remarks.

2 HIDDEN PRONUNCIATION MODEL

Our hidden pronunciation model consists of two trivial probability models: (i) the probability $p(w, x|L)$ that a word w will have hidden pronunciation x according to the lexicon L , and (ii) the probability $p(x, y|M)$ that the stochastic transducer M will generate the hidden pronunciation x and the observed pronunciation y . From these two models, we derive the joint probability

$$p(w, y|M) = \sum_{x \in A^*} p(w, x, y|M) = \sum_{x \in L(w)} p(w|x, L)p(x, y|M)$$

of the word w and the observed pronunciation y . This is a giant finite mixture model consisting of $|L|$ components. When the stochastic transducer M is memoryless and independent of the syntactic word – as it is for our experiments – then this model has only $O(|L| + k^2)$ parameters for a phonetic alphabet A of k symbols. The model parameters may be optimized using expectation-maximization [10, 11].

For each observed pronunciation y , the minimum error rate classifier outputs \hat{w}

$$\begin{aligned} \hat{w} &\doteq \operatorname{argmax}_w \{p(w|y, M, L)\} \\ &= \operatorname{argmax}_w \{p(w, y|M, L)\} \\ &= \operatorname{argmax}_w \left\{ \sum_{x \in A^*} p(w, x, y|M, L) \right\} \\ &= \operatorname{argmax}_w \left\{ \sum_{x \in L(w)} p(w, x, y|M, L) \right\} \end{aligned}$$

where $L(w)$ is the set of pronunciations for the word w according to the lexicon L . This decision rule correctly aggregates the similarity between an observed pronunciation y and all hidden pronunciations for a given word.

The central question raised by the surficial model is how to construct the pronouncing lexicon L . In the generative approach, we carefully craft a sparse lexicon of underlying forms and then rely on the transducer M to map each lexical entry to its surface variants. In the surficial approach, we

crudely construct a rich lexicon of surface forms from the entire training corpus, and rely on the transducer M to generate subtle variations in the hidden pronunciation. The surficial lexicon contains every observed pronunciation for every word in the training corpus.

It is important to distinguish hidden variables from underlying forms. The generative and surficial approaches both use hidden variables; only the generative approach uses underlying forms. In the generative approach, an underlying form represents the irreducible information content of a given surface form and the transducer encodes the predictable information of a given language's phonology. In the surficial approach, the hidden pronunciation x is an actual pronunciation (i.e., a surface form), and the transducer encodes only the variability across pronunciations. Our use of hidden variables is a modeling technique, employed principally to overcome the weakness of the stochastic transducer.

The surficial approach enjoys a number of advantages in addition to its superior recognition performance. Firstly, the lexicon is constructed without costly human intervention. Indeed, it may be constructed entirely automatically from the aligned output of a speech recognition system. Secondly, the model has very few parameters and is not prone to overfitting. Thirdly, the surficial approach outperforms the generative approach by a significant margin on the most difficult speech corpus (Switchboard). Fourthly, the surficial approach offers the promise of instantaneous learning of new words and new pronunciations in real-time (i.e., without batch-optimizing any parameters in the model), because the power of the model comes from the pronunciation lexicon rather than the transducer.

3 SWITCHBOARD

We have tested our approach on Switchboard [4]. Over 200,000 words of Switchboard have been manually assigned phonetic transcripts at ICSI [5]. We partitioned the available transcripts 9:1 into 192,879 training samples and 21,431 test samples. We report recognition results for three experiments; see [10, 11] for additional details. In experiment E1, we used the standard Switchboard pronouncing lexicon. In experiment E3, we built a pronunciation lexicon directly from the training corpus. The test corpus contains 512 samples whose truth value does not appear in the E3 lexicon. In experiment E5, we merged the E1 and E3 lexicons in order to obtain a surficial model that includes at least one pronunciation for every word in the test corpus.

The following table presents the essential characteristics of the lexicons used in the three Switchboard experiments.

	entries	entries /word	novel forms	entries /sample
E1	70,952	1.070	2908	1.895
E3	22,140	2.583	1773	9.434
E5	93,092	1.404	1307	11.329

The first two fields of the table pertain to the lexicon alone. 'Entries' is the number of entries in the lexicon and 'entries/word' is the mean number of entries per word. The final two fields characterize the relation between the lexicon and the test corpus. 'novel forms' is the number of samples in the test corpus whose phonetic forms do not appear in the

lexicon, and 'entries/sample' is the mean number of lexical entries that exactly match the phonetic form of a sample in the test corpus.

The minimum error rate achievable by any decision function on the Switchboard test corpus alone is 7.55%. If the decision function must be consistent across the entire Switchboard corpus, then the minimum error rate achievable on the test corpus is 8.65%.

Our techniques give a 18.61% word error rate using the E1 lexicon and a 12.19% word error rate for the E3 lexicon if we drop the 512 out-of-vocabulary samples from the test corpus. If we merge the E1 and E3 lexicons, then the error rate is 12.63% on the full test corpus. Since the minimum error rate is 8.65%, the surficial approach reduces the error rate of the generative approach by a factor of 2.5.

4 TIMIT

We have also tested our approach on TIMIT [3]. We partitioned the TIMIT transcripts into 30,132 training samples and 11,025 test samples according to the TIMIT instructions. Experiment E1 used the standard TIMIT lexicon, while experiment E3 used the training corpus for its lexicon. Due to the small size of the training corpus and the artificial design of the TIMIT protocol, the test corpus contains 2,897 samples (26.38%) whose words do not appear in the training corpus.

The following table presents the essential characteristics of the lexicons used in the three TIMIT experiments.

	entries	entries /word	novel forms	entries /sample
E1	6,233	1.001	8,564	0.243
E3	11,623	2.376	4,312	1.309
E5	17,856	2.869	3,958	1.551

The minimum error rate achievable by any decision function on the TIMIT test corpus alone is 4.91%. If the decision function must be consistent across the entire TIMIT corpus, then the minimum error rate achievable on the test corpus is 5.69%.

Our techniques give a 17.36% word error rate using the E1 lexicon and a 14.19% word error rate for the E3 lexicon if we drop the 2,897 out-of-vocabulary samples from the test corpus. If we merge the E1 and E3 lexicons, then the error rate is 16.68% on the full test corpus. Although the surficial model outperforms the generative model by a significant amount, its advantage is not as great as it is for Switchboard. We believe that this is due to the small size of the TIMIT corpus and the artificial nature of the TIMIT protocol.

5 CONCLUSION

Given the simplicity of our pronunciation model, and its surprisingly high performance on these difficult tasks, our results argue for the elimination of underlying forms in pronunciation models. The true mapping from underlying forms to surface forms may simply be too complex to be accurately modeled using current techniques. If so, then we would be best served to model the surface forms directly.

REFERENCES

- [1] BAHL, L. R., AND JELINEK, F. Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Trans. Inform. Theory IT-21*, 4 (1975), 404–411.
- [2] CHOMSKY, N., AND HALLE, M. *The Sound Pattern of English*. Harper & Row, New York, 1968.
- [3] GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., PALLETT, D. S., AND DAHLGREN, N. L. The darpa timit acoustic-phonetic continuous speech corpus. Tech. Rep. Speech Disc CD1-1.1, NIST, Gaithersburg, MD, 1986.
- [4] GODFREY, J., HOLLIMAN, E., AND MCDANIEL, J. Switchboard: telephone speech corpus for research and development. In *Proc. IEEE ICASSP* (Detroit, 1995), pp. 517–520.
- [5] GREENBERG, S., HOLLENBACH, J., AND ELLIS, D. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *Proc. IC-SLP* (Philadelphia, October 1996).
- [6] HALLE, M. Phonology in generative grammar. *Word* 18, 1–2 (1962), 54–72.
- [7] JELINEK, F., BAHL, L. R., AND MERCER, R. L. The design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Trans. Inform. Theory IT-21*, 3 (1975), 250–256.
- [8] RILEY, M., LJOLJE, A., HINDLE, D., AND PEREIRA, F. The AT&T 60,000 word speech-to-text system. In *Eurospeech'95: ECSA 4th European Conference on Speech Communication and Technology* (Madrid, Spain, September 1995), J. M. Pardo, E. Enríquez, J. Ortega, J. Ferreiros, J. Macías, and F.J.Valverde, Eds., vol. 1, European Speech Communication Association, pp. 207–210.
- [9] RILEY, M. D., AND LJOLJE, A. Automatic generation of detailed pronunciation lexicons. In *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Kluwer Academic, Boston, March 1996, ch. 12.
- [10] RISTAD, E. S., AND YIANILOS, P. N. Learning string edit distance. Tech. Rep. CS-TR-532-96, Department of Computer Science, Princeton University, Princeton, NJ, October 1996. Revised November 1997.
- [11] RISTAD, E. S., AND YIANILOS, P. N. Learning string edit distance. *IEEE Trans. PAMI* (to appear). Preliminary version as Princeton CS-TR-532-96.