ON THE USE OF STOCHASTIC INFERENCE NETWORKS FOR REPRESENTING MULTIPLE WORD PRONUNCIATIONS

Renato De Mori, Charles Snow, and Michael Galler

McGill University School of Computer Science,
Montreal, Quebec, Canada

ABSTRACT

New techniques are introduced to manage speaker variability in large-vocabulary speaker-independent automatic speech recognition systems. Stochastic inference networks are used to model speaker intentions and automatically derive multiple phonemic transcriptions of words. The statistical inferences are generalized and managed by a Justification Maintenance System module integrated with the speech module, to dynamically adapt the system's performance to the variability of multi-speaker use. Additionally, lexical access to the database of phonemic transcriptions is mediated in multi-pass search with a new method of rejection based on beam-search derived lattices of pseudo-syllabic segments.

1. INTRODUCTION

Many words are pronounced differently by different speakers. This is particularly evident when a language is spoken by people of different mother tongues, and is highly evident for proper nouns. An important problem is that of deciding when it is useful to add a new representation for a word and how to choose this representation.

Differences in pronunciation may be seen as being due to speaker *intentions* based on a personal abstract representation of words. These intentions influence a sequence of articulatory commands which produce, under a variety of constraints, waves of air pressure and velocity which produce the observable signal.

A *Stochastic Inference Network* (SIN) [Pearl88] is a good mathematical framework for modeling speaker intentions in terms of preferences for distortions of canonical word representations and the acoustic observations they produce.

As a further step, rules can be inferred from the derived multiple pronunciations and justified using a *Justification Maintenance System* (JMS).

An application we are considering is a grammar generating sentences for an Air Traffic Control (ATC) task. This grammar, represented in network form, is used in a system for training air traffic controllers. Many controllers speak English as a foreign language, hence tend to pronounce the same word differently. This is especially true for airline and city names. A large proportion of sentence errors is due to the lack of appropriate word pronunciations.

Efficient left-to-right search algorithms have been developed under the assumption that the input signal contains a sequence of words and non-speech segments for which a model is available. This is obviously not the case for *Out of Vocabulary* (OOV) words. Furthermore, search is usually controlled by an iterative network of models in which it is difficult to represent corrections and false starts which may appear in an unpredictable position within the expected word sequence.

In order to take these problems into account a grammar of *Pseudo Syllabic Segments* (PSS) is used to control the initial search steps. A lattice of PSS is generated.

PSS hypotheses can be used to generate lexical hypotheses using partial word representations which can be learned automatically. Once lexical hypotheses have been generated, together with their time bounds computed from PSS bounds, competing phrases in given time intervals can be scored by left-to-right algorithms matching more sophisticated and complete word models with the input signal.

2. THE SYLLABIC GRAMMAR

PSSs are defined as every sequence of consonants ending with a vowel as extracted from the representation of a large lexicon. Special PSSs are also the sequences of consonants following the last vowel of a word. Models of these special PSSs can be used only to match the end of words for generating lexical hypotheses.

0-7803-2431-5/95 $4.00

0 - silb ah			
	24 - m eh		
	24 - m er		
		48 - k y	
		48 - k ih	
		48 - k uh	
		48 - k uw	
			68 - n
(ah	m er	k uh	n ...)

Table 1: Portion of a PSS lattice and max-likelihood phoneme string

The PSS grammar is graphically represented by a subnetwork containing all the non-speech models in parallel with an additional empty transition, followed by a tree representation of the PSSs with each branch labeled by a phoneme. Each phoneme is represented by a Hidden Markov Model (HMM). The type of phoneme models used in the system described in this paper are introduced in [Galler93]. Table 1 depicts a portion of a syllable lattice generated by the PSS grammar, along with the maximum likelihood phoneme string produced by the beam search. The spoken sentence corresponding to the lattice began "American one fifteen...".

A lattice of PSS hypotheses is generated as a first search step. If the PSS grammar, augmented with the non-speech models, covers all the possible descriptions of the speech signal, then it has adequate representation power although the descriptions generated by the search process may be incorrect because the models are not perfect and the search is not admissible. A PSS grammar containing 10,000 PSSs was derived from a lexicon of 100,000 English words.

3. USING STOCHASTIC INFERENCE TO MODEL INTENTIONS

We assume that generation of word hypotheses is guided by a network of word models, each word being represented by one or more pronunciations. Each pronunciation is represented by a sequence of phonemes.

An effective procedure was established as a result of substantial experimental work on data collected from students and air traffic controllers.

The initial assumption is that a new pronunciation has to be added when an error can be corrected based on justifiable distortions of the canonical form of a word or its already accepted pronunciations. A justifiable distortion is one produced by rules which have been accepted based on experimental evidence constrained by speech knowledge. Speech knowledge establishes, for example, unacceptable substitutions like a vowel with a fricative.

3.1. New Pronunciation Learning Procedure

Assume sentence S, having as most likely word sequence $S = W_1, W_2, ..., W_j, ..., W_n$, is misrecognized, and that the first (left-to-right) misrecognized word is W_j. Let $(t_b, t_e)_i$ denote the beginning and ending times (input frame numbers) for word i. The procedure for establishing accepted pronunciations is as follows:

1. determine time interval τ consisting of at least $(t_b, t_e)_j$.
2. for each alternate pronunciation suggested by the SIN (see below), assess how well that pronunciation corrects the error, iterating until the error is corrected or no further alternate candidates exist. In this latter situation, if possible, change τ to cover the previous word and perform this step once again. If it is *still* the case that no candidate corrects the error, exit with failure.
3. infer rules from the proposed form. If these rules already exist in the JMS, add the current instance as new justification for the existing rules. Otherwise, endeavor to associate new rules with existing rules which may be more detailed and/or might also have corrected the misrecognition, so as to produce more general rules.
4. update lexicon: check for entries which are similar (e.g., same place of articulation context as the phoneme the newly applied rule modified) and apply the rule(s) to them.

By counting when alternate pronunciations are used, it will be possible to associate counts and probability estimation to rules.

Examples of inferred rules are reported in Table 2. The rule `ay` → `ey` is justified by samples of '9' and 'United.' A possible generalization of this observation is that it is valid when preceded by `n`, it contains the main stress, and is followed by `n` or `t`. This makes it applicable to '19' but application can be delayed because '19' has a longer string of phonemes following `ay` than '9' or 'United.'

3.2. Evaluation of Alternate Pronunciations

Alternate pronunciations can be ordered according to a measure of the distortion they exhibit with respect to the canonical form.

Reference	Inferred Rule With Evidence
aa (ae)	ow (e) Continental
(p) ae (n)	(p) eh (n) Japan Airlines
(n) ao (r)	(n) ow (r) Northwest
ax (m)	ah (m) American
ay	ey nine United
eh	ih Northwest
ih (r)	eh (r) Air Canada
(k) ow	(k) ao Air Mexico

Table 2: Examples of inferred rules

A measure of distortion can be found by using SINs in the following way. Let A be the observed acoustic evidence of a word, W_k be a particular word which may be pronounced in a variety of ways depending upon the speaker's intentions, and W_{k_j} be the j^{th} variant, or *distortion*, of that word.

Distortions W_{k_j} can be evaluated based on a "belief"

$$\begin{aligned} B(W_{k_j}) &= P(W_{k_j}|W_k A) \\ &= \frac{P(A|W_{k_j})P(W_k)}{P(A)} \\ &= \frac{P(A|W_{k_j}W_k)P(W_{k_j}|W_k)P(W_k)}{P(AW_k)} \end{aligned}$$

$P(AW_k)$, common to all the candidates W_{k_j}, can be ignored in a comparison among candidates W_{k_j}. $P(A|W_{k_j}W_k) = P(A|W_{k_j})$ if we consider distortions W_{k_j} unique to W_k. This leads to

$$B(W_{k_j}) = cP(A|W_{k_j})P(W_{k_j}|W_k)$$

Needed is a way to determine $P(W_{k_j}|W_k)$. One way is to use the product of the probabilities of the rules for obtaining W_{k_j} from W_k. If rules not yet inferred have to be applied a floor probability is assigned to them.

Rule probabilities may have both a static and dynamic component; the latter component can capture individual speaker variations and uses a cache architecture.

A final problem remains to be addressed: the generation of candidates of W_{k_j} to be evaluated.

Given the time interval τ it is possible to extract the PSSs detected in that interval. Let $W^*_{k_j}$ be pronunciations of the word in the focus τ obtained from the PSS lattice segment. If none of $W^*_{k_j}$ corrects the observed error, then either W_k is not directly responsible for that error or the error cannot be corrected. Otherwise, candidates W_{k_j} can be found by successively reducing the differences between $W^*_{k_j}$ and the available pronunciations of W_k. Those new forms which allow correction of the error are candidates to become new pronunciations in ψ_k, the set of pronunciations of W_k. The one with maximum belief can be kept.

Candidates can be found by replacing phonemes in $W^*_{k_j}$ in a way which is consistent with constraints imposed by phonetic knowledge. Some procedural rules found useful in practice include the following:

1. remove insertions in $W^*_{k_j}$
2. align vowels in $W^*_{k_j}$ with one of the forms in ψ_k
3. substitute vowels with vowels following an order imposed by the place of articulation
4. same as (3) for consonants in the same class and with the following order: fricatives, nasals, semivowels, plosives.

Inferred rules can be applied across the entire lexicon so as to achieve the effect of correcting pronunciations based on observed distortions. Since application of the updates involves dependencies on other, previously applied updates, and retraction of an application (and its consequent applications) may be required, the JMS is employed to control update application.

It is possible that with enough 'training' on individual dialects, clusters of sets of updates may be constructed allowing a generic lexicon to be expeditiously modified to accommodate a speaker's particular dialect.

3.3. Word Hypothesization or Rejection by a Syllable Lattice

The second interesting problem being studied is word hypothesization from a syllable lattice. Here, a tree lexicon representation is used to walk through the syllable lattice. At each frame, several phone candidates are available in the syllable lattice. Some of the phonemes match perfectly with the phonemes associated with the tree branches, and some are different. Beam search can

be performed by using distances of phonemes in the lattice with respect to the phonemes in the tree. Distances are added and only the nodes whose scores remain in the beam are left in an active list, so that only the phonemes corresponding to the transitions from these nodes can be matched against the phonemes in the lattice in the following frame. Distances are learned from rules of aceptable distortions. Large distances indicate OOVs.

Out-of-vocabulary events can be used to trigger user prompts, or alternatively, context switches in which lexical access is applied to different or specialized vocabulary lists. Preliminary tests have been carried out on the ATC task described in section 1, in which the lattice-mediated recognition procedure also performs OOV-spotting.

4. EXPERIMENTAL RESULTS

Preliminary experiments were carried out with a system having the following features. The signal was sampled at 16 kHz and represented in 16 bit linear encoding on a Hewlett Packard 9000/735. For each 20 ms wide frame, the FFT and a vector of 12 Mel scaled cepstral coefficients (MCC) was computed every 10 ms (i.e., overlapping frames). Signal energy, its first derivative and the first derivative of the MCCs were also computed. Continuous density HMMs were used to model phonemes. There are 40 models trained exclusively on the TIMIT corpus [SenZue88] as described in [Galler93]. No retraining of models for this task was performed .

Phoneme models are concatenated to form word models which are compiled into a network by a grammar-builder which accepts regular expressions. For the ATC grammar used, the network contains approximately 9,000 nodes and 150 word models.

Multiple pronunciations were obtained from three professional air traffic controllers, each speaking 47 sentences. The test was performed with five (non-ATC) speakers using a test set of 15 sentences, distinct from the previous set of 47. Two recognition suites were performed, one using a canonical pronunciations based on the CMU 100,000 word dictionary. the other using multiple pronunciations which were generated manually based upon a very small number of sets of 47 sentences. Word error rates are reported in Table 3. Even though these results are highly preliminary, they show two noteworthy points: (1) speakers with heavy accents – speakers 2 (Greek) and 5 (Italian) – benefit from dramatic improvemnts, and, (2) the overall effect of multiple pronunciations is to reduce, across all

	Error Rate		
Speaker	Canonical	Multiple	Improvement
1	19.9	1.3	93.3
2	70.0	11.9	83.0
3	26.9	9.4	64.9
4	39.9	10.0	74.7
5	80.6	13.6	83.1

Table 3: Effect of Multiple Pronunciations

speakers, variability in error rates. The average error rate in the canonical test is 47.5% (s.d. 26.7), in the multiple pronunciation case it is 9.3% (s.d. 4.7).

On the two native English speakers (speakers 1 and 3), string error rates using multiple pronunciations averaged 15.8%, reflecting an average improvement of 56.3% over the corresponding canonical pronunciations. On non-native English speakers (speakers 2, 4, and 5), string error rates, after a similar improvement (51.0%) over canonical pronunciations, remained somewhat high (average 40.8%).

The ATC grammar contains sequences of digits in various places. Multiple pronunciations for digits were obtained also by the analysis of previous data consisting of sequences of digits uttered by randomly selected students.

ACKNOWLEDGEMENTS

This research was sponsored by the National Science and Engineering Council of Canada, and the Institute for Robotics and Intelligent Systems.

5. REFERENCES

[Galler93] Galler M., De Mori R., "Using Search to Improve Hidden Markov Models", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, April, 1993, pp. II-303 – II-307.

[Pearl88] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inferece* Morgan Kauffman, 1988

[SenZue88] Seneff S., and Zue V.W., "Transcription and Alignment of the TIMIT Database", (distributed with the DARPA TIMIT CD-ROM by the National Bureau of Standards), April 9, 1988.