A CHUNK-BASED PHONETIC SCORE FOR MOBILE VOICE SEARCH

Rohit Prabhavalkar[*]

prabhava@cse.ohio-state.edu
Department of Computer Science and Engineering
The Ohio State University

Jasha Droppo

jdroppo@microsoft.com
Speech Technology Group
Microsoft Research

ABSTRACT

We propose a *chunk-based* phonetic score for re-scoring word hypotheses for the mobile voice search task. The score is based on a novel technique for aligning decoded phone sequences with forced-alignments of hypothesized word sequences and exploits phone-boundary timing information. In experimental results, we find that the proposed approach results in relative a word error rate reduction of 4.4% and a relative sentence error rate reduction of 2.3% for the Windows live search for mobile task [1].

Index Terms— voice search, phonetic score, pronunciation modeling

1. INTRODUCTION

As mobile voice search services become widespread, increasingly large amounts of acoustic data become available. The most straightforward use of this data is to train larger acoustic models. However, the gains obtained from such an approach have been shown to result in diminishing returns in previous studies [2].

Recognizing large vocabulary conversational speech continues to remain a challenging problem for automatic speech recognition (ASR) systems. Conversational speech is characterized by disfluencies and pronunciation variability. As a result of this "sloppy" articulation, the surface pronunciation can differ markedly from the expected canonical pronunciation. Evidence from a study by McAllaster et al. [3] indicates that pronunciation variation is one of the main causes of poor performance of ASR systems on conversational speech: in oracle experiments, when data is sampled from acoustic models, error rates were found to be significantly lower (5-10%) when the sampled data corresponds to the dictionary pronunciation than when the data was sampled according to phone transcripts corresponding to the utterance (40%).

There is reason to believe that adding more training data to a traditional speech recognition system will not address this problem. While context-dependent modeling (for example, using triphones) can account for some of the variation, such an approach has its limitations. It has been shown previously by Jurafsky et al. [4] that while triphone-based systems capture phone substitutions, they do not model insertions and deletions well.

One approach to modeling pronunciation variation is to add variant pronunciations to the lexicon; the variant pronunciations may be learned automatically from the data or else based on prior linguistic knowledge [5]. In this work, however, we take an alternative approach. We use a large (4.5M+) corpus of transcribed utterances to learn patterns of variation between the forced-alignment of the reference word hypothesis, corresponding to the canonical pronunciation, and a phonetic decoding of the utterance, which can be thought of as an estimate to the surface pronunciation. In our system, the standard acoustic-model (AM) and language-model (LM) scores are augmented with a *phonetic score* that is computed over phone sequences corresponding to the forced-alignment of a hypothesized word sequence ($\mathbf{q}^{\mathrm{w}}$) and a phonetic decoding of the utterance ($\mathbf{q}^{\mathrm{x}}$).

Our approach is similar, in spirit, to previous work on learning non-parametric pronunciation models [6], learning context-dependent string edit distances [7] and phone-to-word transduction [8]. The novelty of our approach, distinguishing it from previous methods [6, 7], lies in the fact that we do not treat $\mathbf{q}^{\mathrm{w}}$ and $\mathbf{q}^{\mathrm{x}}$ simply as sequences of phonetic labels: we explicitly exploit timing information associated with phone-boundaries in these sequences, which may be easily obtained as part of the forced-alignment or phonetic decoding process. In addition, the proposed approach naturally incorporates long-span context since it is defined in terms of contiguous sequences of phonetic labels - which we term as *chunks* - similar to the multigram model of Deligne et al. [9].

2. NOTATION AND PRELIMINARIES

Let $\mathbf{q}^{\mathrm{w}} = (q_1^{\mathrm{w}}, q_2^{\mathrm{w}}, \cdots, q_M^{\mathrm{w}})$ and $\mathbf{q}^{\mathrm{x}} = (q_1^{\mathrm{x}}, q_2^{\mathrm{x}}, \cdots, q_N^{\mathrm{x}})$ be the phone sequences corresponding to a forced-alignment of the word hypothesis and the phonetic decoding respectively, where each $q_i^{\mathrm{w}}, q_j^{\mathrm{x}} \in \mathcal{Q}$, the set of phones for the task. Each phone q, is associated with a start-time - start(q) - and a corresponding end-time - end(q) - representing the hypothesized time-boundaries for the phone.

Given a particular decoded phone sequence, $\mathbf{q}^{\mathrm{x}}$, we denote a segmentation $\mathbf{s}$ of $\mathbf{q}^{\mathrm{x}}$ by $\mathbf{s} \sim \mathbf{q}^{\mathrm{x}}$. Informally, a segmentation splits up $\mathbf{q}^{\mathrm{x}}$ into contiguous sequences of phonetic labels which are non-overlapping and span the entire length of $\mathbf{q}^{\mathrm{x}}$. Stated formally, a segmentation $\mathbf{s}$ of $\mathbf{q}^{\mathrm{x}}$ can be written as,

$$\mathbf{s} = \langle s_1, s_2, \cdots, s_K \rangle = \langle (r_1, t_1), (r_2, t_2), \cdots, (r_K, t_K) \rangle \quad (1)$$

where each component of the segmentation, s_i, is termed as a segment and consists of a starting index r_i and an ending index t_i that refer to phones in $\mathbf{q}^{\mathrm{x}}$, with $|\mathbf{s}| = K$ denoting the number of segments in the segmentation. In any valid segmentation, the first segment must begin at the first index ($r_1 = 1$), the last segment must end at the last index ($t_K = N$), consecutive segments must follow each-other and be non-overlapping (Equation 2 below) and segments must have positive lengths (Equation 3).

$$r_i = t_{i-1} + 1 \qquad 1 < i \leq K \quad (2)$$

$$r_i \leq t_i \qquad 1 \leq i \leq K \quad (3)$$

[*]This work was performed as part of a summer internship at Microsoft Research.

Thus, a segmentation, $\mathbf{s} \sim \mathbf{q}^{\text{x}}$, splits $\mathbf{q}^{\text{x}}$ into $|\mathbf{s}|$ *chunks*: the ith chunk of $\mathbf{q}^{\text{x}}$ is defined as $\mathbf{q}^{\text{x}}_{s_i} = \left\langle q^{\text{x}}_{r_i}, q^{\text{x}}_{r_i+1}, \cdots, q^{\text{x}}_{t_i} \right\rangle$. Finally, we denote by $\mathbf{q}^{\text{x}}_{\mathbf{s}}$ the entire sequence of segments corresponding to the decoded phone sequence : $\mathbf{q}^{\text{x}}_{\mathbf{s}} = \left\langle \mathbf{q}^{\text{x}}_{s_1}, \mathbf{q}^{\text{x}}_{s_2}, \cdots, \mathbf{q}^{\text{x}}_{s_K} \right\rangle$.

In defining our phonetic score, given a segmentation of the phonetic decoding, we would like to associate chunks, $\mathbf{q}^{\text{x}}_{s_i}$, of the phonetic decoding with corresponding chunks from $\mathbf{q}^{\text{w}}$. In previous work, these have been computed directly using Levenshtein alignments between the two sequences [7] or using similar dynamic programming-based string alignments with learned edit costs [6]. In contrast to these approaches, our model directly utilizes the timing information associated with individual phones in the segmentations in order to determine corresponding chunks. Observe that a segmentation $\mathbf{s} \sim \mathbf{q}^{\text{x}}$ induces a segmentation in $\mathbf{q}^{\text{w}}$ as well: we define the ith chunk induced in $\mathbf{q}^{\text{w}}$ as the minimal contiguous sequence of phones in $\mathbf{q}^{\text{w}}$ whose time-extent completely contains the ith segment $\mathbf{q}^{\text{x}}_{s_i}$. More formally, we define a corresponding segmentation of $\mathbf{q}^{\text{w}}$ as,

$$\mathbf{s}^{\text{w}} = \left\langle s^{\text{w}}_1, \cdots, s^{\text{w}}_{|\mathbf{s}|} \right\rangle = \left\langle (u_1, v_1), \cdots, (u_{|\mathbf{s}|}, v_{|\mathbf{s}|}) \right\rangle \tag{4}$$

where, u_i and v_i are the start and end indices for the ith component of the segmentation, satisfying the following constraints:

$$u_i = \underset{1 \le j \le M}{\text{argmax}} \left\{ \text{start}(q^{\text{w}}_j) \le \text{start}(\mathbf{q}^{\text{x}}_{s_i}) \right\} \qquad 1 \le i \le |\mathbf{s}| \tag{5}$$

$$v_i = \underset{1 \le j \le M}{\text{argmin}} \left\{ \text{end}(q^{\text{w}}_j) \ge \text{end}(\mathbf{q}^{\text{x}}_{s_i}) \right\} \qquad 1 \le i \le |\mathbf{s}| \tag{6}$$

We use the notation $\mathbf{q}^{\text{w}}_{s^{\text{w}}_i}$ to denote the ith chunk of $\mathbf{q}^{\text{w}}$ corresponding to the ith induced segment $\mathbf{s}^{\text{w}}_i$. Finally, we denote the entire sequence of induced chunks as $\mathbf{q}^{\text{w}}_{\mathbf{s}^{\text{w}}} = \left\langle \mathbf{q}^{\text{w}}_{s^{\text{w}}_1}, \mathbf{q}^{\text{w}}_{s^{\text{w}}_2}, \cdots, \mathbf{q}^{\text{w}}_{s^{\text{w}}_K} \right\rangle$. Unlike segments in $\mathbf{s} \sim \mathbf{q}^{\text{x}}$ which are non-overlapping by definition (Equation 2-3), the induced segments in $\mathbf{s}^{\text{w}}$ (and thus chunks in $\mathbf{q}^{\text{w}}$) need not be non-overlapping. We refer to $\mathbf{q}^{\text{x}}_{s_i}, \mathbf{q}^{\text{w}}_{s^{\text{w}}_i}$ as the ith *chunk pair* induced by the segmentation $\mathbf{s} \sim \mathbf{q}^{\text{x}}$. This is illustrated in Figure 1.

3. CHUNK-BASED PHONETIC SCORE

In defining our chunk-based phonetic score, we make the simplifying assumption that the chunk pairs are conditionally independent given the segmentation $\mathbf{s}$.

$$p(\mathbf{q}^{\text{w}}_{\mathbf{s}^{\text{w}}}, \mathbf{q}^{\text{x}}_{\mathbf{s}} | \mathbf{s}) = \prod_{i=1}^{|\mathbf{s}|} p(\mathbf{q}^{\text{w}}_{s^{\text{w}}_i}, \mathbf{q}^{\text{x}}_{s_i} | s_i) \tag{7}$$

We now define the phonetic score, denoted by $\varphi(\mathbf{q}^{\text{w}}, \mathbf{q}^{\text{x}})$, as the posterior probability of the word hypothesis conditioned on the phonetic decoding under the chunk-based model,

$$\varphi(\mathbf{q}^{\text{w}}, \mathbf{q}^{\text{x}}) = p(\mathbf{q}^{\text{w}} | \mathbf{q}^{\text{x}}) = \frac{p(\mathbf{q}^{\text{w}}, \mathbf{q}^{\text{x}})}{p(\mathbf{q}^{\text{x}})} \tag{8}$$

$$= \frac{\sum_{\mathbf{s}} \prod_{i=1}^{|\mathbf{s}|} p(\mathbf{q}^{\text{w}}_{s^{\text{w}}_i}, \mathbf{q}^{\text{x}}_{s_i} | s_i) p(\mathbf{s})}{\sum_{\mathbf{s}} \prod_{i=1}^{|\mathbf{s}|} p(\mathbf{q}^{\text{x}}_{s_i} | s_i) p(\mathbf{s})} \tag{9}$$

3.1. Efficiently computing the phonetic score

The phonetic score can be computed efficiently if the prior over segmentations $p(\mathbf{s})$ decomposes over individual segments s_i. Writing, $p(\mathbf{s}) = Z(\mathbf{s}) \prod_{i=1}^{|\mathbf{s}|} f(s_i)$, where $f(s_i) \ge 0$ and $Z(\mathbf{s})$ is a normalization term, from Equation 9 we have,

$$\varphi(\mathbf{q}^{\text{w}}, \mathbf{q}^{\text{x}}) = \frac{\sum_{\mathbf{s}} \prod_{i=1}^{|\mathbf{s}|} p(\mathbf{q}^{\text{w}}_{s^{\text{w}}_i}, \mathbf{q}^{\text{x}}_{s_i} | s_i) f(s_i)}{\sum_{\mathbf{s}} \prod_{i=1}^{|\mathbf{s}|} p(\mathbf{q}^{\text{x}}_{s_i} | s_i) f(s_i)} \tag{10}$$

The phonetic score can be computed using a dynamic programming algorithm by accumulating sums corresponding to partial segmentations of the first j phones of $\mathbf{q}^{\text{x}}$, which we denote by $\mathbf{q}^{\text{x}}_{(1,j)}$. Accumulating partial sums corresponding to the numerator of Equation 10 in $\alpha(j)$ we have,

$$\alpha(j) = \sum_{\mathbf{s} \sim \mathbf{q}^{\text{x}}_{(1,j)}} \prod_{i=1}^{|\mathbf{s}|} p(\mathbf{q}^{\text{w}}_{s^{\text{w}}_i}, \mathbf{q}^{\text{x}}_{s_i} | s_i) f(s_i) \tag{11}$$

$$= \sum_{l=j-L^{\max}+1}^{j} \alpha(l-1) p(\mathbf{q}^{\text{w}}_{s^{\text{w}}_{(l,j)}}, \mathbf{q}^{\text{x}}_{s_{(l,j)}} | s_{(l,j)}) f(s_{(l,j)}) \tag{12}$$

where, $L^{\max}$ is the maximum allowed length of any segment in the segmentation and $s_{(l,j)}$ denotes a segment starting at index l and ending at index j in $\mathbf{q}^{\text{x}}$ and $s^{\text{w}}_{(l,j)}$ is the corresponding segment in $\mathbf{q}^{\text{w}}$. We refer to $L^{\max}$ as the *order* of the phonetic score. Larger values of $L^{\max}$ allow the model to learn the correlations present in larger contexts, at the expense of memory and speed. The denominator sum in Equation 10 can be computed analogously, by accumulating partial sums in $\beta(j)$,

$$\beta(j) = \sum_{\mathbf{s} \sim \mathbf{q}^{\text{x}}_{(1,j)}} \prod_{i=1}^{|\mathbf{s}|} p(\mathbf{q}^{\text{x}}_{s_i} | s_i) f(s_i) \tag{13}$$

$$= \sum_{l=j-L^{\max}+1}^{j} \beta(l-1) p(\mathbf{q}^{\text{x}}_{s_{(l,j)}} | s_{(l,j)}) f(s_{(l,j)}) \tag{14}$$

Once $\alpha(j)$ and $\beta(j)$ have been computed for $1 \le j \le N$, the required phonetic score can be computed as $\varphi(\mathbf{q}^{\text{w}}, \mathbf{q}^{\text{x}}) = \frac{\alpha(N)}{\beta(N)}$.

3.2. Discussion of proposed phonetic score

Intuitively, the one-best phonetic decoding, $\mathbf{q}^{\text{x}}$, can be thought of as an approximation to the *surface* phonetic sequence uttered by the speaker. The phonetic score implicitly models pronunciation variability, since the score can be interpreted as a measure of the degree to which a given word hypothesis (in terms of its phonetic forced-alignment) is consistent with the observed surface pronunciation (as represented by the phonetic decoding of the acoustics).

The intuition behind our method of forming corresponding chunk pairs in the two phone sequences is based on the following observation: the phone boundaries of the decoded phone sequence and the forced-alignment of the reference word hypothesis tend to align perfectly for phones that are produced canonically. On the other hand, mis-alignments tend to occur when the decoded phones do not correspond to the phones in the canonical pronunciation; in such cases, the chunks $\mathbf{q}^{\text{w}}_{s^{\text{w}}_i}$ provide additional contextual information for the phonetic environment in which the mis-match occurs. Thus, our proposed score which also incorporates variable-length long-span chunk-pairs allows for the computation of a phonetic score that captures context that is intermediate between the phonetic edit distance of [7] and the non-parametric model of [6].

4. ESTIMATING DISTRIBUTIONS OVER CHUNK-PAIRS

Following [6], we estimate the distributions over chunk pairs, $p(\mathbf{q}^{\text{w}}_{s^{\text{w}}_i}, \mathbf{q}^{\text{x}}_{s_i} | s_i)$, non-parametrically given a corpus of training examples which have been annotated with the reference word transcription. We begin by computing a forced-alignment of the phones corresponding to the reference hypothesis $\mathbf{w}$ to obtain the phone

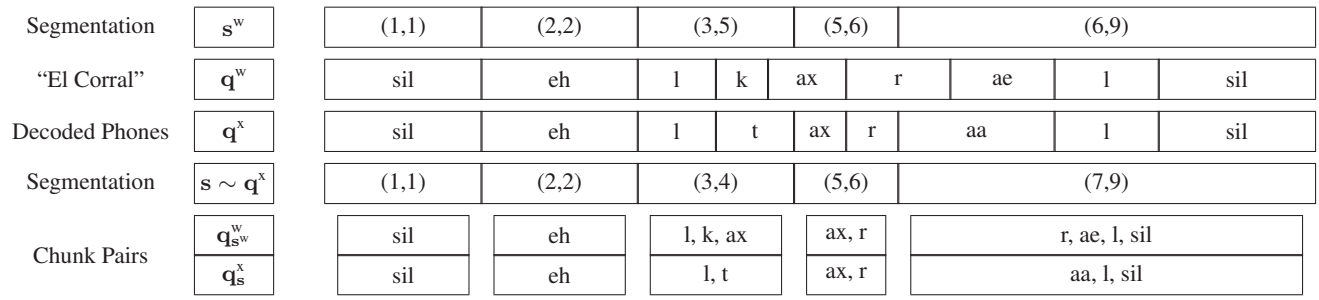

Fig. 1. An example illustrating *chunk pairs* induced in the forced-alignment of the word hypothesis $\mathbf{q}^{\mathrm{w}}$ and the phonetic decoding of the utterance $\mathbf{q}^{\mathrm{x}}$. The segmentation $\mathbf{s} = \langle (1,1),(2,2),(3,4),(5,6),(7,9) \rangle \sim \mathbf{q}^{\mathrm{x}}$ splits the decoded phone string into five chunks. Each of these chunks has a corresponding chunk in $\mathbf{q}^{\mathrm{w}}$ that is determined based on the hypothesized phone boundaries in the two sequences. Notice that although chunks in $\mathbf{q}^{\mathrm{x}}$ are non-overlapping, the same is not true for chunks in $\mathbf{q}^{\mathrm{w}}$.

sequence $\mathbf{q}^{\mathrm{w}}$ and the corresponding phonetic decoding $\mathbf{q}^{\mathrm{x}}$ for the utterance. Assuming that $|\mathbf{q}^{\mathrm{x}}| = N$, we can find $O(N^2)$ segments and thus chunks in $\mathbf{q}^{\mathrm{x}}$. Each of these chunks has a corresponding chunk in $\mathbf{q}^{\mathrm{w}}$ as described in Section 2. If we denote by $\mathcal{C}(\mathbf{q}^{\mathrm{w}}_{s^{\mathrm{w}}}, \mathbf{q}^{\mathrm{x}}_{s})$ the count of the number of times a chunk $\mathbf{q}^{\mathrm{w}}_{s^{\mathrm{w}}}$ occurs corresponding to a chunk $\mathbf{q}^{\mathrm{x}}_{s}$ in the training examples, then we can estimate the required distribution as,

$$p(\mathbf{q}^{\mathrm{w}}_{s^{\mathrm{w}}}, \mathbf{q}^{\mathrm{x}}_{s} | s) = \frac{\mathcal{C}(\mathbf{q}^{\mathrm{w}}_{s^{\mathrm{w}}}, \mathbf{q}^{\mathrm{x}}_{s})}{\sum_{\substack{\mathbf{q}^{\hat{\mathrm{w}}}_{\hat{s}^{\hat{\mathrm{w}}}}, \mathbf{q}^{\hat{\mathrm{x}}}_{\hat{s}} \\ s.t. |\hat{s}| = |s|}} \mathcal{C}(\mathbf{q}^{\hat{\mathrm{w}}}_{\hat{s}^{\hat{\mathrm{w}}}}, \mathbf{q}^{\hat{\mathrm{x}}}_{\hat{s}})} \tag{15}$$

where, the denominator in Equation 15 sums over the counts of all chunk pairs observed in the training examples, where the chunk corresponding to the phonetic decoding has length $|s|$. We emphasize here that $\mathbf{q}^{\hat{\mathrm{x}}}_{\hat{s}}$ represents a chunk of length $|s|$ and $\mathbf{q}^{\hat{\mathrm{w}}}_{\hat{s}^{\hat{\mathrm{w}}}}$ represents a corresponding chunk; these do not represent entire chunk sequences.

5. EXPERIMENTS

In order to determine the effectiveness of the proposed chunk-based phonetic score, we conducted n-best re-scoring experiments on the Windows live search for mobile voice search task [1]. This task consists of approximately 5000 hours of training data, 9 hours of development data, and 13 hours of test data. For our experiments, the entire training set was used to train both the acoustic model and the chunk-based phonetic score model.

The overall architecture of the proposed system is presented as Figure 2. We first perform an n-best decoding using the baseline system to get a set of n-best word hypotheses, ranked by the acoustic ($\log p(\mathbf{x}|\mathbf{q}^{\mathrm{w}_n})$) and scaled language model ($\gamma \log p(\mathbf{q}^{\mathrm{w}_n}, \mathrm{w}_n)$) scores derived from the baseline system, where γ is the language model scaling parameter. The baseline system is a hidden Markov model-based (HMM) system whose acoustic model contains 135K diagonal Gaussian components shared by 22K states, which are in turn shared by 9.7K HMMs. The system utilizes a trigram language model with 4.7M n-grams over a lexicon of 65K words. For each of the word sequence hypotheses in the n-best list, we perform a forced-alignment of the word using its dictionary pronunciation, to obtain a sequence of phones, $\mathbf{q}^{\mathrm{w}_n}$, corresponding to each word hypothesis sequence, w_n.

In parallel, we also perform a less-constrained phone decoding of the input utterance using a phone-based language model to obtain a decoded phone sequence, $\mathbf{q}^{\mathrm{x}}$ corresponding to the utterance. The system utilizes the same acoustic model as is used to perform the decoding and forced-alignment described in the previous step. The language model is a trigram model with 16K n-grams over the phone-set consisting of 45 phones. It was trained on a forced-alignment of in-domain data.

The phonetic score $\varphi(\mathbf{q}^{\mathrm{w}_n}, \mathbf{q}^{\mathrm{x}})$ between the pair of phone sequences is computed according to Equation 10, with a suitable choice of prior distribution $p(\mathbf{s})$ on the segmentations. The chunk-pair probabilities, required in the computation are assumed to be pre-computed as described in Section 4.

Finally, the n-best word hypotheses w_n are re-ranked by computing a new score, $\psi(\mathrm{w}_n, \mathrm{x})$, obtained by linearly interpolating the baseline LM-AM scores with the logarithm of the phonetic score,

$$\begin{aligned} \psi(\mathrm{w}_n, \mathrm{x}) =& (1-\lambda)\{\log p(\mathbf{x}|\mathbf{q}^{\mathrm{w}_n}) + \gamma \log p(\mathbf{q}^{\mathrm{w}_n}, \mathrm{w}_n)\} \\ &+ \lambda\{\log \varphi(\mathbf{q}^{\mathrm{w}_n}, \mathbf{q}^{\mathrm{x}})\} \end{aligned} \tag{16}$$

where $0 \leq \lambda \leq 1$ is the interpolation weight and is tuned on development set.

5.1. Effect of ($L^{\max}$) and segmentation prior ($p(\mathbf{s})$)

The choice of the order, $L^{\max}$, of the phonetic score has an impact on the amount of contextual information available to the system with the caveat that the model size increases as $L^{\max}$ increases. Although there is a danger of sparsely estimating higher-order chunk pair distributions, in practice we did not see any negative effect while increasing $L^{\max}$ as large as 10; performance improved steadily as the order of the phonetic score was increased to about 7, where it plateaued. The experiments in this section use $L^{\max} = 8$.

The model presented in Section 3.1 allows for a non-uniform prior distribution over the possible segmentations. In pilot experiments, we experimented by setting $f(s_i) = e^{\delta}$ to bias the average segment length. However, system performance was found to be insensitive to various choices of δ. Consequently, the experiments presented here use $\delta = 0$, which corresponds to a uniform prior.

5.2. Smoothing over n-best phone decodings

In Section 3, we describe the system as computing a phonetic score for each word hypothesis with respect to the 1-best decoded phone sequence, $\mathbf{q}^{\mathrm{x}}$. As we have mentioned previously, the 1-best decoded phone sequence represents an approximation of the true phonetic sequence uttered by the speaker. Errors in the phonetic decoding

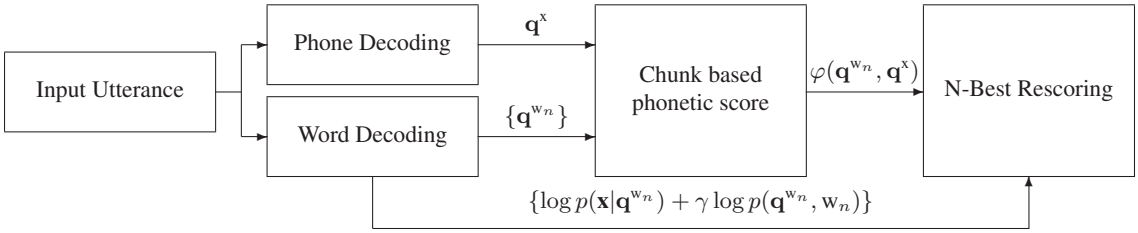

Fig. 2. Architecture of proposed system. The chunk based phonetic score augments the acoustic and language model scores output from the baseline word decoder.

System	Development Set		Test Set	
	WER (%)	SER (%)	WER (%)	SER (%)
Baseline	33.0	33.5	33.8	34.8
unsmoothed	31.5	32.9	32.6	34.4†
smoothed	31.1	32.5	32.3	34.0

Table 1. Performance obtained by rescoring 100-best word hypotheses from the baseline system using either the phonetic score computed using the 1-best phone decodings as defined in Equation 10 (unsmoothed) or the smoothed phonetic score averaged over the 10-best phone decodings as defined in Equation 17. The experiment marked with a † represents a significant ($p \leq 0.05$) improvement over the corresponding baseline, all other improvements are significant with $p \leq 0.01$.

will result in poor approximations to the true phonetic sequence, and potentially a poor estimate for the phonetic score. In such situations, the n-best decoded phone sequences, $\mathbf{q}^{\mathrm{x}_1}, \mathbf{q}^{\mathrm{x}_2}, \cdots, \mathbf{q}^{\mathrm{x}_n}$ may more accurately reflect the uncertainty in the surface phonetic sequence. We therefore consider the use of a *smoothed* phonetic score, obtained by averaging the phonetic score from Equation 10 over the n-best decoded phone sequences,

$$\varphi^{(\mathrm{n})}(\mathbf{q}^{\mathrm{w}}, \mathbf{q}^{\mathrm{x}_1}, \cdots, \mathbf{q}^{\mathrm{x}_n}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\mathbf{q}^{\mathrm{w}}, \mathbf{q}^{\mathrm{x}_i}). \qquad (17)$$

Once the smoothed phonetic score has been computed, it can be interpolated with the baseline AM-LM scores as before.

5.3. Results

Based on the results obtained in pilot experiments, we evaluated both smoothed and unsmoothed versions of the phonetic score obtained by rescoring the 100-best word lists from the baseline using either the 10- or 1-best phonetic decodings respectively. In Table 1 we present the results obtained on both the development and test sets. As can be seen from the table, systems employing either the smoothed or unsmoothed phonetic scores result in significant improvements (measured using a binomial sign test) over the corresponding baselines. The use of the smoothed phonetic score results in significant performance improvements in terms of word error rate (WER) over the baseline ($p \leq 0.01$); it also significantly outperforms the system employing the unsmoothed phonetic score ($p \leq 0.025$). The smoothed phonetic score also significantly outperforms the baseline in terms of sentence error rate (SER) on both the development as well as test set ($p \leq 0.01$). Overall, the best performing system achieves a relative WER reduction of 4.4% over the baseline and a relative SER reduction of 2.3% on the test set.

6. CONCLUSIONS

We presented a chunk-based phonetic score based on a novel technique for computing corresponding chunks between the phonetic decoding and the forced-alignment of the word hypothesis. The proposed framework is conceptually simple, and easy to incorporate within an n-best rescoring framework. The proposed phonetic score implicitly captures pronunciation variation, while simultaneously incorporating long-range context through higher-order chunks. In experimental results, the proposed system resulted in a 4.4% relative WER reduction and a 2.3% relative SER reduction on Windows live search for mobile task [1].

7. REFERENCES

[1] A. Acero, N. Bernstein, R. Chambers, Y. C. Ju, X. Li, J. Odell, P. Nguyen, O. Scholz, and G. Zweig, "Live search for mobile: Web services by voice on the cellphone," in *Proc. ICASSP*, 2008.

[2] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, pp. 1513–1525, 2006.

[3] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *Proc. ICSLP*, 1998.

[4] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. ICASSP*, 2001.

[5] M. Wester, "Pronunciation modeling for ASR - knowledge-based and data-derived methods," *Computer Speech & Language*, vol. 17, pp. 69–85, 2003.

[6] B. Hutchinson and J. Droppo, "Learning non-parametric models of pronunciation," in *Proc. ICASSP*, 2011.

[7] J. Droppo and A. Acero, "Context dependent phonetic string edit distance for automatic speech recognition," in *Proc. ICASSP*, 2010.

[8] G. Zweig and J. Nedel, "Empirical properties of multilingual phone-to-word transduction," in *Proc. ICASSP*, 2008.

[9] S. Deligne, F. Yvon, and F. Bimbot, "Variable-length sequence matching for phonetic transcription using joint multigrams," in *Proc. Eurospeech*, 1995.