A DETECTION BASED APPROACH TO ROBUST SPEECH UNDERSTANDING

Kuansan Wang

Speech Technology Group, Microsoft Research, Redmond, WA 98052, USA

ABSTRACT

Field speech data pose great challenges to statistical modeling because the speech signal is often intermixed with extraneous sounds and other environmental noises that are either too difficult to compensate dynamically or too expensive to collect sufficient data for proper offline training. In this paper, we propose a detection based method in which the speech recognizer can sharply tune to only the "meaningful" speech and gracefully ignore the "unwanted" audio segments. The method is designed to be integrated with the frame synchronous search for a single pass processing. In contrast to the conventional keyword spotting techniques, this integration allows the use of the language model for better predicting the detection targets during the search. To study its efficacy, we apply the framework to a spontaneous speech understanding application where cohesive phrases congruent to the domain semantics and application context are used as the salient feature for selective hearing. Experimental results on the effectiveness of the system in dealing with out of domain phrases and other spontaneous speech effects are encouraging.

1. INTRODUCTION

Despite the rapid evolvements in automatic speech recognition and understanding technology, the information theoretical model of speech communications, proposed three decades ago [7], remains largely unchanged. As illustrated in [7, Fig. 1][12, Fig. 2.1], the speaker is often viewed as employing source and channel coding mechanisms that "encode" a message into the speech waveform. As a result, a speech recognizer, be it human or machine, often is viewed as a "decoder" because its job is merely to reverse the coding process and recover the message encoded in the speech signal. A widely adopted principle in designing the decoder follows the *maximum a posteriori* (MAP) decision rule. There the task of the decoder is to search for an outcome *S'* that maximizes the posterior probability, i.e.

$$S' = \arg\max P(S \mid x) = \arg\max P(x \mid S)P(S) . \quad (1)$$

To pursue (1), the majority of the recognizers [6][12] use what we call the *classification* based approach, in which every acoustic frame in the speech signal $x = (x_1, x_2 ... x_t ...)$ is classified and its probabilistic score is accumulated unselectively. More specifically, we have

$$P(x \mid S) = \sum_q \prod_t P(x_t \mid q_t) P(q_t \mid S, q_{t-1}) \quad (2)$$

where $q = (q_1, q_2 ... q_t \ldots)$ is a hidden, first order Markov process that attempts to classify each acoustic frame x_t into a corresponding modeling unit that can yield a highest likelihood score $P(x_t \mid q_t)$. Implied in (2) is an assumption that acoustic observations can be treated as statistically independent given the hidden process, an assumption known to be problematic. This assumption is relaxed by segmental modeling techniques [2][3] where each hidden variable represents a segment of the acoustic waveform rather than an individual frame, i.e.,

$$P(x \mid S) = \sum_q \prod_i P(x_{t_i}, ... x_{t_{i+1}-1} \mid q_i) P(q_i \mid S) \quad (3)$$

These formulations have been widely demonstrated as effective in tackling large vocabulary continuous speech recognition (LVCSR) problems [6][12].

The classification based approach, though widely adopted, indeed poses great challenges to statistically modeling. First, because every frame counts, one needs to model all sort of acoustic conditions, including in and out of domain speech as well as non-speech sounds like lip and throat noises. For spontaneous speech, the modeling effort becomes even more demanding as speakers may stutter, hesitate, correct, or insert extraneous sounds such as *uh*, *um*, *well*, *you know*, etc., anywhere in an utterance. A common solution to this problem, known as keyword spotting [2][9], is to device "garbage" models to account for these audio segments. However, because of the diverse nature of the unwanted sounds, methods of obtaining high quality garbage models remain elusive. Secondly, a hallmark of the classification based approach, as highlighted in (2) or (3), is that every frame contributes *equally* to the overall score of a hypothesis *S*. As a result, when an utterance is largely composed of unwanted sounds, as is often the case in the field applications (e.g., "*um….yes*" where the "*yes*" makes up only a small portion of the signal), the score for the speech portions of the audio is often overwhelmed by their longer counterparts accounted for by the garbage models. Effectively, this leads to the hypotheses being compared against one another based on how well they match garbage models, a situation that is highly undesirable and potentially may serve as a major source of errors given the poor quality of

the scores produced by the garbage models.

Although there is no reason why (1) may not be applied to explain human speech recognition as well, there is no evidence that the auditory system follows the same classification based approach as in (2) or (3). In fact, many studies have suggested [1][5][8][13] that the remarkable robustness exhibited by the auditory system may be attributed to the use of a detection based rather than a classification based mechanism. Applying detection based methods for automatic speech recognition has been attempted with various degrees of success [4][9][10][14]. In the previous works, however, detection techniques are often included as a disjoint processing stage where the detection algorithms are not necessarily designed to optimize the overall system performance described by (1), and therefore a multi-pass architecture is often used. In this paper, we propose a new formulation that tightly integrates the detection based algorithm into the MAP decision. As a result of the tight integration, we show realizing the detection based recognition in a single pass architecture is possible. In the following, we first elaborate the two contributions of the paper in Sec. 2 and 3, and in Sec. 4 describe some experimental results.

2. DETECTION BASED DECISION

A distinctive auditory processing principle is that a signal is often decomposed into components but the components are not always used together. For speech, it is as if only the relevant portions are detected and selected out of the waveform. This principle, when contrasted with the classification oriented keyword spotting techniques, highlights the appeal of a detection based method in that salient segments of speech can be extracted automatically without the need of garbage models. Applying the principle for the MAP decision rule of (1), we have

$$\begin{aligned} P(x \mid S) &= \sum_{\pi} P(x, \pi \mid S) = \sum_{\pi} \prod_i P(x_i \mid \pi) P(\pi \mid S) \\ &\approx \max_{\pi} \prod_i P(x_{t_i}, \ldots, x_{t_i + d_i} \mid \pi) P(\pi \mid S) \end{aligned} \quad (4)$$

where $\pi = \{(t_i, d_i): i = 0, 1, 2, \ldots\}$ represents a partition of the signal x into a sequence of landmarks, where the i^{th} landmark $x_i = (x_{ti}, \ldots x_{ti+di})$ occurrs at time t_i with a duration d_i. The above equation bears resemblances to the classification based recognition based on segmental models described by (3) in that the units assumed to be statistically independent are much larger than adjacent acoustic frames. However, a notable difference between a detection based approach of (4) and a classification based approach of (3) is that a partition here does not have to cover every acoustic frame in the utterance, i.e., $t_i + d_i \leq t_{i+1}$.

As indicated in (4), a detection based recognizer must address how the landmarks composing a partition are detected, and how the likelihood of each partition for a given hypothesis can be evaluated. In this work, sequential detection and language model techniques, respectively, are used for these purposes.

3. RECURRENT SEQUENTIAL DETECTION

Sequential detection is a technique that addresses whether an ongoing observation $x_i = (x_t, x_{t+1}, x_{t+2}, \ldots)$ has provided sufficient evidence to accept or reject a hypothesis, or such decision should be postponed pending more observations. Sequential detection is suitable for applications where immediate decisions amidst continuing observations are desired. In addition to reducing the false acceptance rate P_F and false rejection rate P_M as in regular detection problems, the latency to decision is also a factor to be considered for sequential detection.

The mathematical foundation of sequential detection is well established. It can be shown [11] that the optimal decision rule for sequential detection is to conduct a sequential probabilistic ratio test (SPRT) as

$$\text{SPR} = \frac{P(x_i \mid \pi) P(\pi)}{P(x_i \mid \pi^*) P(\pi^*)} : \begin{cases} > A & \text{accept } \pi \\ < B & \text{reject } \pi \\ \text{otherwise} & \text{defer} \end{cases} \quad (5)$$

where π^* denotes the alternative hypothesis of π. The choice for the acceptance and rejection thresholds is bounded by

$$A \leq \frac{1 - P_M}{P_F}, \quad B \geq \frac{P_M}{1 - P_F} \quad (6)$$

and is independent of the true probability distributions of $P(x_i \mid \pi)$ and $P(x_i \mid \pi^*)$. It is therefore possible to choose these two thresholds to match however low error rates at the price of lengthening the latency to decision d_i.

As sequential detection is designed to process ongoing observations, it is straightforward to integrate SPRT into the word or frame synchronous search process commonly implemented in LVCSR and realize (1) with (4) in a single pass architecture. The integration leads to *recurrent* sequential detection for landmarks: Whenever the signal x_i is appended with a new observation, the SPR for every hypothesis is computed. The approximation of producing the distribution of the alternative hypothesis for utterance verification [9] can be adapted here, i.e., by using the weighted n^{th}-order mean of the competing hypotheses as the probability of the alternative hypothesis

$$P(x_i \mid \pi^*) P(\pi^*) = \sqrt[n]{\sum_{\sigma \neq \pi} P(\sigma) P(x_i \mid \sigma)^n} \,. \quad (7)$$

The search process consequently applies (5) to prune out all the hypotheses below the rejection threshold B, and continues to draw observation until a single winner surpassing the acceptance threshold A emerges, at which time all the lingering competitors are rejected. Note that (5) implies that, with sufficiently large n, there is likely to be only one winning hypothesis with SPR > 1 because

$$P(x_i \mid \pi^*) \to \max_{\sigma \neq \pi} P(x_i \mid \sigma) \quad \text{as } n \to \infty .$$

Each SPRT detects one landmark in a hypothesis π. The statistically independent assumption made in (4) implies that the SPRT be recurrently applied to detect successive landmarks composing the hypothesis π.

It is theoretically possible that a signal instance x_i can cause the sequential detection to be indecisive indefinitely. In practice, an upper bound on the landmark duration can be introduced so that $P(\pi \mid S) = 0$ if some d_i in π exceeds the limit. The timeout mechanism, together with the winner-take-all nature of the recurrent sequential detection, plays a key role in alleviating the reliance on the garbage models. For sharply tuned models, an extraneous sound segment, being a poor match to any of the speech models, produces no clear winner, resulting in the SPR for contending landmark hypotheses to wander between the decision thresholds. These hypotheses are eventually turned away either when the timeout expires or when the observations progress to the "meaningful" acoustic segment and a corresponding landmark is detected.

4. CASE STUDY: MIPAD EXPERIMENTS

Although the detection based framework described in Sec. 2 and 3 may be potentially applied to speech recognition in general, we first assessed its effectiveness on a speech understanding task. In this paper we further report the experimental results based on the MiPad scenarios described in [15]. MiPad is a mobile device with a personal information management (PIM) application. The recognition target S for MiPad is the user's intention for the PIM tasks, which usually consists of a collection of semantic objects representing the command (e.g. "*send email*") and the parameters for the command (e.g. "to *Alex* with subject: *progress report*").

The choice of landmarks in (4) is obviously application dependent. Since MiPad is a speech understanding task, we choose the phrase segments composing the semantic objects as the landmarks for detection. For example, the landmarks for the email creation task include the expressions for the email creation command, the recipients, the subject, the body of the message, but not the semantic objects such as date, time, location, etc., that are relevant to calendar or meeting tasks. This domain knowledge is crafted into the semantic language model $P(\pi \mid S)$ using the unified probabilistic context free grammar (PCFG) and N-gram technique [15] and is woven into the frame synchronous search process described in Sec. 3. The semantic language model dynamically predicts for each hypothesis which semantic objects to detect during the decoding processing. As an example illustrated in Fig. 1., when the landmark of check schedule command is detected, the hypothesis will automatically tune off the detection of irrelevant semantic objects such as email

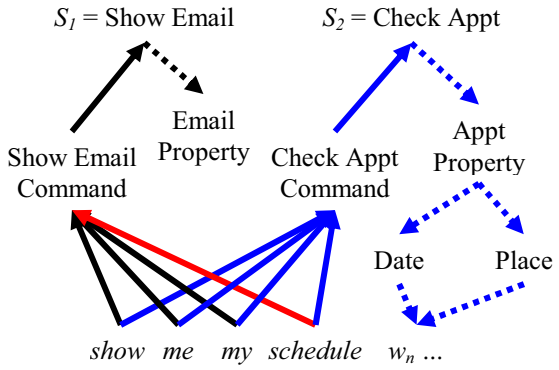

Fig. 1. An illustration of two competing hypotheses being composed as words are recognized. Dotted lines represent the semantic objects predicted by the semantic language model. When the tie breaker, in this example the word "schedule", emerges, the SPR of S_2 trumps the alternative hypotheses, leading to the search process to detect only the landmarks for S_2 for the rest of the utterance.

subject, recipients, etc., considerably narrowing the search space dynamically. The use of the prediction power of a semantic language model and the tight integration of the domain knowledge during search manifest themselves as significant contrasts to the conventional keyword spotting based methods for detection based recognition (e.g., [9]).

Two versions of MiPad, one using the classification based and the other, detection based recognition[1], were implemented for user studies. The complexity of the acoustic models, the language models, the graphical layouts, and the development time devoted to fine tune the free parameters were kept roughly in par with each other. Both versions ran on a Toshiba 3500 Tablet PC using the built-in microphone located on the lower left corner of the display. Non-stationary noises, such as those from the hard disk spinning and the pen tapping the screen, were audible in the recordings. The experiments were all conducted in a moderately noisy office for which no prior speech data were available for acoustic training or adaptation. A generic gender and speaker independent acoustic model with the online cepstral mean normalization was used for all the experiments. For the classification based recognizer, a simple phone loop was used as the garbage model. The prior $P(S)$ in (1) was set to assume the uniform distribution for the experiments. Despite the efforts to make the two systems comparable, the nature of a classification versus a detection algorithms, however, does introduce distinctive behaviors. On the classification based system, MiPad shows the understanding result only after the user has finished the whole utterance, while the sequential detection algorithm allows the detection based system to display understanding outcome immediately

[1] Video available at http://research.microsoft.com/srg/videos under MIPAD demo 2003.

after a semantic object is accepted. As a result, the detection based MiPad can employ a dynamic prompting strategy and solicit fewer out of domain utterances [16].

The experiments previously conducted for the user interface studies [16] also provide some insights into the efficacy of the detection based approach to recognition. We utilized (6) to determine the thresholds for a false acceptance rate P_F at 5% and the decision timeout (Sec. 3) at 1.5 seconds. This leads to an average per user semantic object recognition accuracy at 57.03% with the standard deviation 12.78%. The low accuracy is primarily due to a high false rejection rate as the accuracy rate is the percentage of the correctly recognized semantic objects minus the substitution, the false acceptance, and the false rejection errors. The large fluctuations among test subjects are reflected in the standard deviation. In contrast, the classification based system has an average accuracy 65.7% with standard deviation 10.75% after the poor recognition on full sentences instigating a change in user's speaking patterns (see discussion below). The *t*-test shows the difference in accuracy rate is *not* statistically significant (t=0.729, p=0.253, df=4) due to the large variances in recognition accuracy across users.

The resilience of the detection based system to the spontaneous speech effects seems to have enticed the users to employ longer, and hence fewer, utterances to complete a task. The average number of semantic objects per utterance is 6.17 and 2.02 for the detection and the classification based systems, respectively. With the standard deviations at 1.26 and 0.36, the *t*-test shows the difference is statistically significant (t=5.49, p=0.0025). The average number of utterances used to complete a task is 1.33 versus 6.75, with standard deviations 0.144 and 1.80, respectively. The difference is also statistically significant under *t*-test (t=5.187, p=0.003). These data provide a quantitative support for the observations that users for the classification based system would quickly switch to short utterances narrowly targeted at individual input fields after several attempts of using longer and more naturally phrased sentences failed. In contrast, users of the detection based system often would repeat, in the same utterance, the phrase segments that are not detected, making the sentence structure even more non grammatical and more spontaneous. Surprisingly, users did not seem to be bothered by including corrections in their sentences, even though the overall recognition accuracy is not high in the detection based system. This is supported objectively from the data that users did not switch to shorter utterances as with the classification system, and subjectively by the feedback that they like the highly interactive nature of the detection based system. Users only pointed out that the latency to decision (d_i) seemed unpredictable and the occasionally long latency could be irritating. The unwieldy latency, based on the theory of sequential detection (Sec. 2) might originate from the poor models of $P(x \mid \pi)$ and $P(x \mid \pi^*)$ used in the experiments. This highlights the need of good acoustic models because of the role they play in the latency, even though they do not influence the choice of detection thresholds as indicated in (6).

REFERENCES

[1] Allen J. "How do humans processing and recognize speech?" *IEEE Trans. Speech and Audio Processing*, pp. 567—577, October, 1994.

[2] Gish H., Ng K. "A segmental speech model with applications to word spotting." in *Proc. IEEE ICASSP-1993,* Minneapolis, MN, 1993.

[3] Hon H.-W., Wang K. "Unified frame and segment based models for automatic speech recognition." in *Proc. IEEE ICASSP-2000*, Istanbul, Turkey, 2000.

[4] Hori T., *et al*. "Paraphrasing spontaneous speech using weighted finite state transducers." in *Proc. ISCA & IEEE Workshop on spontaneous speech processing and recognition*, Tokyo, Japan, 2003.

[5] Houtsma A. J. M, Rossing T. D., and Wagenaars W. M. Auditory demonstrations, Institute for Perception Research (IPO), Eindhoven, Netherlands and the Acoustical Society of America, New York, NY, 1987.

[6] Huang X. D., Acero A., and Hon H.-W. Spoken Language Processing, Prentice Hall, NJ, 2001.

[7] Jelinek F., Bahl L. R., and Mercer R. L. "Design of a linguistic statistical decoder for the recognition of continuous speech." *IEEE Trans. Information Theory*, pp. 250–256, May 1975.

[8] Juang B. H., Furui S. "Automatic recognition and understanding of spoken language – a first step toward natural human-machine communication." *Proceedings of IEEE*, pp. 1142 – 1165, August, 2000.

[9] Kawahara T., Lee C.-H., and Juang B.-H. "Flexible speech understanding based on combined key-phrase detection and verification." *IEEE Trans. Speech and Audio Processing*, pp. 558–568, November, 1998.

[10] Niyogo R., Mitra P., and Sondhi M. M. "A detection framework for locating phonetic events." in *Proc. ICSLP-98,* Sydney Australia, 1998.

[11] Poor H. V. An introduction to signal detection and estimation, Springer-Verlag, New York, NY, 1988.

[12] Rabiner L. R., Juang B.-H. Fundamentals of Speech Recognition, Prentice Hall, NJ, 1993.

[13] Wang K., Shamma S. A. "Spectral shape analysis in the central auditory system." *IEEE Trans. Speech and Audio Processing*, pp. 382—395, September, 1995.

[14] Wang K., Goblirsch D. M. "Extracting dynamic features using the stochastic matching pursuit algorithm for speech event detection." in *Proc. IEEE ASRU Workshop*, Santa Barbara, CA, 1997.

[15] Wang K., "Semantic object synchronous decoding in SALT for highly interactive speech interface", in *Proc. Eurospeech-2003*, Geneva, Switzerland, 2003.

[16] Wang K. "A study on semantic synchronous understanding on speech interface design." in *Proc. UIST-2003*, Vancouver, BC, 2003.