7th International Conference on Spoken
Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

# COMBINING SPEAKER AND SPEECH RECOGNITION SYSTEMS

*Larry P. Heck and Dominique Genoud*

Nuance Communications, 1380 Willow Rd., Menlo Park, CA 94025
{heck,genoud}@nuance.com

## ABSTRACT

This paper presents a general framework for the integration of speaker and speech recognizers. The framework poses the problem of combining speech and speaker recognizers as the joint maximization of the *a posteriori* probability of the word sequence and speaker given the observed utterance. It is shown that the *posteriori* probability can be expressed as the product of four terms: a likelihood score from a speaker-independent speech recognizer, the (normalized) likelihood score of a text-dependent speaker recognizer, the likelihood of a speaker-dependent statistical language model, and the prior probability of the speaker. Efficient search strategies are discussed, with a particular focus on the problem of recognizing and verifying name-based identity claims over very large populations (e.g., ”My name is John Doe”). The efficient search approach uses a speaker-independent recognizer to first generate a list of top hypotheses, followed by a resorting of this list based on the combined score of the four terms discussed above. Experimental results on an over-the-telephone speech recognition task show a 34% reduction in the error rate where the test-set consists of users speaking their first and last name from a grammar covering 1 million unique persons.

## 1. INTRODUCTION

The speech signal conveys several levels of information, including the spoken message (words), the identity of the talker, and the language spoken. The aim in automatic speech processing is to extract this information for use in a variety of applications, including information access (database queries), services, and communications. Typically, the speech and speaker recognition problems are treated as separate extraction goals and, as a result, are often designed to ignore or remove the other information conveyed in the signal. For example, speaker-independent speech recognizers often reduce speaker variability through speaker normalization techniques in an attempt to improve performance, while speaker recognition systems often treat the text of the speaker as unwanted variability.

For some applications, however, combining the information from the speech and speaker recognizers can provide substantial benefits to the application goal. For example, combined speech and speaker recognizers could be beneficial in the class of problems where the goal of the automatic system is to recognize what *a particular person* is saying. A specific problem in this category is the automatic recognition of spoken identity claims (e.g., ”My name is John Doe”). For this problem, the system can utilize a speaker recognition system to help guide the speech recognition search for the identity claim that corresponds to the person whose voice characteristics most closely matches that of the user. Preliminary results in [1] and extended in this paper show significant promise in this approach.

Another class of problems benefiting from a combined speech and speaker recognizer involves identifying or detecting a particular person by how they speak. Techniques to address this problem have been based both on the acoustic characteristics of the speaker's voice, as well as the particular words and phrases a user speaks that helps differentiates them from other talkers. Applications include surveillance, verification of a customer by monitoring a conversation between the agent and customer, and speaker tracking. We presented an approach to improved speaker detection performance based on speaker-dependent word/phrase choices in [2].This technique relied on speaker-dependent statistical language models (N-grams) to represent the idiosyncratic differences between individual speakers. A similar approach was also demonstrated in [3].

Finally, combined speech and speaker recognizers can be used to simultaneously perform speaker and knowledge verification. Knowledge (or verbal information) verification is the authentication of a person's identity based on their ability to provide answers to questions only known the correct user. Answers to the question can be verified by performing speech recognition followed by a comparison of the text (or natural language interpretation) to the correct answer stored in a database. Successful techniques that simultaneously perform speaker and knowledge verification have been presented in [4, 5].

This paper presents a framework for the combination of speech and speaker recognizers. In Section 2, we present a formulation of the problem as one of jointly maximizing the probability of the word sequence and the speaker given the observed acoustics. This formulation leads to a clear identification of the necessary subsystem components that contribute to the objective function. An efficient search strategy is then discussed based on a rescoring of a speaker-independent recognizer's N-best list of hypotheses. Section 3 explores the application of this formulation and efficient search strategy to the problem of large-scale identity claim recognition. The speech and speaker recognition subsystems are described, and new experimental results are presented that extend our previous work on this problem. Combining speech and speaker recognizers reduce the error rate on this problem by 34%.

## 2. FORMULATION

To address the applications described in the previous section, a framework is required for combining speech and speaker recognition systems. Stated mathematically, the goal is to find the word sequence and the speaker that maximizes the joint probability among all possible word sequences $W$ and speakers $S$, conditioned on a feature vector sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_T\}$

$$\begin{aligned}
\{\hat{W}, \hat{S}\} &= \operatorname*{argmax}_{W,S} P(W,S|\mathbf{X}) \\
&= \operatorname*{argmax}_{W,S} \frac{P(\mathbf{X}|W,S) \cdot P(W,S)}{P(\mathbf{X})} \\
&= \operatorname*{argmax}_{W,S} P(\mathbf{X}|W) \cdot \frac{P(\mathbf{X}|W,S)}{P(\mathbf{X}|W)P(\mathbf{X})} \cdot P(W,S) \\
&= \operatorname*{argmax}_{W,S} \underbrace{P(\mathbf{X}|W)}_{SI-Speech} \cdot \underbrace{\frac{P(\mathbf{X}|W,S)}{P(\mathbf{X}|W)}}_{TD-Speaker} \cdot \underbrace{P(W|S)}_{SD-Lang} \cdot P(S) \qquad (1)
\end{aligned}$$

The four components represent the contribution from a speaker-independent recognizer (SI-Speech), a (normalized) text-dependent speaker recognizer (TD-Speaker), a speaker-dependent language model (SD-Lang), and a prior for the given speaker, $P(S)$. The speaker-dependent language model represents how a particular person chooses their words. For example, if a person has a habit of saying certain phrases ("Yes, exactly") or uses particular disfluencies ("um"), and/or filler words indicating speaker turns in a dialog ("yeah"), then this speaker-dependent language model will help to identify the individual person by theses word sequences. This has been implemented with a standard N-gram statistical language model[2, 3]. The prior probability of the speaker, $P(S)$, can be estimated from the application if data is available (e.g., frequency of calling and/or ANI for telephony applications), or can simply set to a constant if data is unavailable.

The three likelihood terms and the prior in (1) can be combined in the search at various time/state resolutions, from the frame-level to the utterance-level. The combination at the frame-level could be accomplished for example in the forward pass of a Viterbi search. However, for large numbers of speakers and possible word sequences, the search space implemented in the forward pass of Viterbi will be very large, $\mathcal{O}(\text{Words} \times \text{Speakers})$. In this case, efficient search strategies are required.

One such search strategy uses a multi-pass rescoring approach, where the speaker-independent recognizer is used first to generate a list of N-Best hypotheses of spoken word sequence. Once the N-Best list is specified, then the combined score from the three likelihood terms and the prior in (1) can be used to resort the N-Best list. While this approach is suboptimal, it greatly reduces the complexity of the search.

## 3. EXAMPLE APPLICATION: NAME-BASED IDENTITY CLAIM RECOGNITION

Many current applications of automatic speech/speaker recognition technologies are designed to provide over-the-telephone access to personal accounts and services. Examples include voice-activated access to brokerage and bank accounts, telephone calling card accounts, insurance and medical records, and personalized voice portals. Each of these applications requires that the users *identify* themselves so that the user's personal profile (e.g., favorite stock portfolio, sports teams, etc) can be loaded and made readily available. Application developers have a number of ways to elicit the user's identity, but perhaps the most convenient and user-friendly is to have the system simply prompts the user to speak their name.

However, the automatic recognition of spoken names over very large name grammars is a significant challenge, preventing the use of names as the identity claim. But the problem can be simplified by using the added knowledge that the system will recognize the name from *the particular persons who has that name*. Posing the problem this way facilitates the use of the framework developed in Section 2.

### 3.1. System Description

Referring to Equation (1), the name-based identity claim system reported here makes the following simplifications:

1. the choice of words when speaking a name is not indicative of the speaker's identity (i.e., the language model term $P(W|S) = P(W)$), and
2. the prior probability of the speaker is not known, so the term $P(S)$ is constant and therefore can be dropped from the search.

The resulting maximization problem for name-based identity claim can be expressed as

$$\{\hat{W}, \hat{S}\} = \operatorname*{argmax}_{W,S} P(\mathbf{X}|W) \cdot P(W) \cdot \frac{P(\mathbf{X}|W,S)}{P(\mathbf{X}|W)} \qquad (2)$$

#### *3.1.1. Speech Recognition Subsystem*

The speech recognition system used is described in [6]. The acoustic models use context dependent triphones states that are clustered using bottom-up agglomerative clustering. Each state cluster shares a set of Gaussians (called genones).The system was trained with over a million digit strings, stock quote requests, and phonetically rich utterances collected over the telephone from various sources. The output score of the recognizer is composed as follows (with $\beta$ scaling the language model score)

$$\Lambda_{speech} = \log P(\mathbf{X}|W) + \beta \log P(W) \qquad (3)$$

#### *3.1.2. Speaker Recognition Subsystem*

The speaker recognition subsystem used in the following experiments is based on a likelihood ratio detector. The score of an utterance is obtained by computing the average log-likelihood ratio as follows:

$$\begin{aligned}
\Lambda_{speaker} &= \log \frac{P(\mathbf{X}|W,S)}{P(\mathbf{X}|W)} \\
&\approx \frac{1}{T} \sum_{t=1}^{T} \log\ p(\mathbf{x}_t|W, \lambda_s) \ -\ \log\ p(\mathbf{x}_t|\lambda) \qquad (4)
\end{aligned}$$

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ denotes the set of feature vectors extracted from the utterance by the feature extraction front-end, $\lambda_s$ is the speaker model (corresponding to the speaker that the caller claims to be), and $\lambda$ is the text-independent background model used for normalizing the likelihood scores.

#### *3.1.3. Combined System*

We utilize the efficient multipass search strategy described in the previous section to solve the maximization problem in (2), where the the combined score that will be used to resort the N-Best list, $\Lambda_T$, is computed by summing the speech and speaker scores as defined above, i.e.,

$$\Lambda_T = \Lambda_{speech} + w \cdot \Lambda_{speaker}. \qquad (5)$$

The scalar weight $w$ provides a tunable parameter depending on the relative quality of the speech and speaker recognition systems. (Before combining the above scores, the log likelihoods from the speech and speaker recognizers are normalized to compensate for difference in the dynamic range (on a held-out testset).)

### 3.2. Experiments

The goal of these experiments is to determine the potential impact of integrating speaker recognition scores into the speech recognition search process. These experiments will focus on a multi-pass approach, where the N-Best list of the recognizer is rescored with the speaker recognition system.

The test-set of name-based identity claims used for these experiments is composed of 1000 utterances of personal names (first and last name) spoken over long distance telephone lines. There are 500 unique speakers in the testset. The grammar consists of approximately 1 million first+last names from the white pages of a United States city telephone directory. That means that the potential size of the group in which the speakers have to be identified is 1 million.

After completing a first pass, the speech recognizer produced an N-Best list of the top (unique) hypotheses according to the speech recognition score. For every entry in the N-best list, an associated speaker recognition score was computed by scoring the spoken utterance against the speaker model associated with this entry. If the entry did not belong to the correct user and a model existed for that entry, then a score was computed against this model. However, given the large number of possible first+last names on the N-Best list we did not have speaker models for many of the entries. Therefore, to simulate the performance of the new technique where every entry is a competitive identity claim, we generated a speaker recognition score by randomly choosing a score from a pre-computed distribution of impostors estimated on a similar task.

**Table 1**. *Example of a name-based identity claim capture with an N-Best rescoring approach. The correct transcript is "chris craft". Combining the speech and speaker recognition scores identifies the second entry in the N-Best list as the best scoring hypothesis, which corrects the original error made by the speech recognizer.*

| N | Hypothesis | $\Lambda_{speech}$ | $\Lambda_{speaker}$ | $\Lambda_T$ |
|---|---|---|---|---|
| 1 | chris graf | 1.01 | 0.91 | 2.08 |
| **2** | **chris craft** | **1.00** | **1.00** | **2.18** |
| 3 | chris krauss | 0.47 | 0.01 | 0.48 |
| 4 | chris kress | 0.18 | 0.96 | 1.32 |
| 5 | christi crouse | 0.10 | 0.00 | 0.10 |
| 6 | bruce graf | 0.06 | 0.02 | 0.08 |
| 7 | craig kraft | 0.04 | 0.18 | 0.25 |
| 8 | chris groves | 0.02 | 0.00 | 0.02 |
| 9 | christine craft | 0.01 | 0.00 | 0.01 |
| 10 | curtis craft | 0.01 | 0.55 | 0.66 |

Table 1 shows an example of an N-Best list for the name-based identity claim task with the top 10 entries. The actual spoken utterance was "chris craft". The first column shows the ranking of the hypotheses, with the first row being the best hypothesis of the speech recognizer alone. The corresponding hypotheses are shown in the second column, with the (normalized) scores from the speech recognizer shown in the third column. The scores from the speaker recognizer are shown in the fourth column. The last column of the table shows the combined score of the speech and speaker recognizers. This example has the correct hypothesis in the second position using the recognizer score alone, but in the first (highest) position when using the combined speech and speaker recognizers.

To determine the sensitivity of the NL error to the combination weight $w$ in Equation (5), we varied $w$ and plotted the NL-error rate of the combined system at $N = 10$, as shown in Figure 1. The overall performance improves significantly, even for a very small contribution from the (normalized) speaker recognition score. The original NL error rate of the recognizer alone was 18.6% and the equal error rate of the speaker recognizer is 4.85%. At the best combination weight of $w = 1.19$ (i.e., the speaker recognition score is weighed 1.19 times that of the speech recognition score), the NL error rate drops to approximately 12.2%, a relative improvement of **34%**. For large combination weights, the performance degrades from the optimal weight but levels out to an NL error rate of approximately 15%, still better than that obtained with the speech recognition score alone. This shows that improvements can even be obtained with an approach that simply uses the speech recognizer to generate the N-best list, discards the speech recognition scores, and resorts the N-best list with the speaker recognition scores.

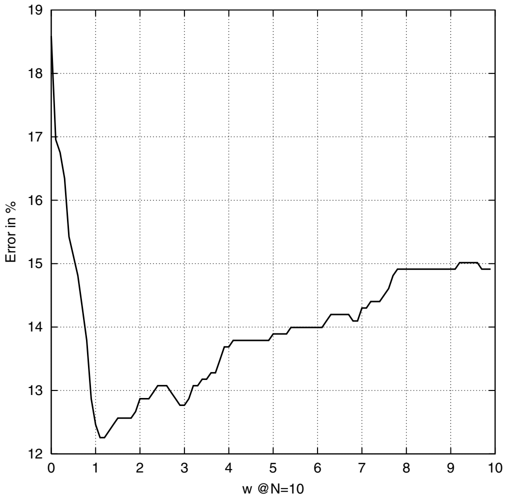

**Fig. 1**. *Sensitivity of the system error rate with an N-best size of 10 to the combination weight w between the speech and speaker recognizer scores (normalized). As can be seen, the minimum error rate is at* $w = 1.19$.

Using the combination weight of $w = 1.19$, the top curve in Figure 2 shows the NL-error rate of the combined speech and speaker recognizers ("AllSystemErr after resorting without rejection") as compared to the theoretical limit ("Min nbest error") for a given size N-Best list. Given that we are using a multi-pass rescoring approach, the improvement to this error rate from the speaker recognition system is bounded by the N-Best performance. The N-Best performance is a theoretical measure that counts an utterance as correctly recognized if the correct name appears anywhere in the top N hypotheses generated by the recognizer. As the size of the N-Best list increases, the integration of the speech and speaker recognition systems shows significant improvement, even

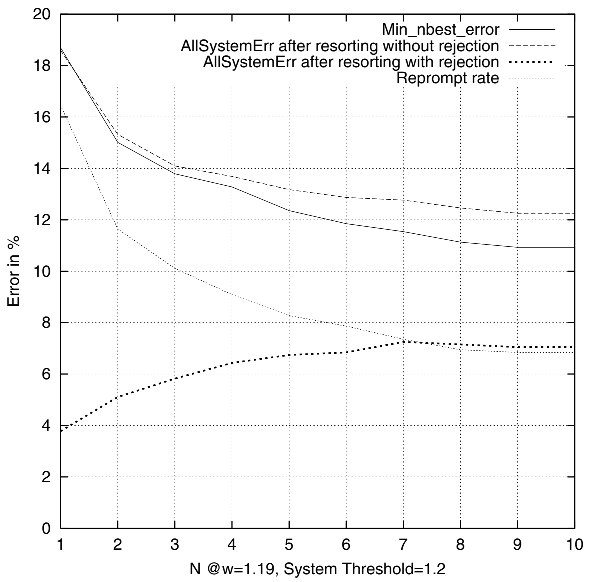

**Fig. 2**. *The top curve "AllSystemErr after resorting without rejection" shows the performance of the combined speech and speaker recognition system vs. the size of the N-best list. This is compared to the theoretical limit "Min nbest error". The combined system gives a 34% relative reduction in NL-error rate as compared to the speech recognition system alone. The bottom two curves show the performance of the system when a rejection threshold is used. The first curve labeled "AllSystemErr after resorting with rejection" shows significant gains in performance with the penalty shown in the last curve of a non-zero "Reprompt Rate".*

with only 3 hypotheses.

Figure 3 shows the system performance when the rejection threshold is varied. All the combined scores $\Lambda_T$ below the threshold are rejected and the utterance is reprompted. Below a rejection threshold of 1.7, the system error decreases but with of course an increase of the reprompt rate. Rejection thresholds greater than 1.7 hurt performance due to previously correctly accepted utterances being rejected.

The ability to reject an utterance that would otherwise be an error, and then prompt the user to repeat the utterance is critical when deploying a speech recognition application. Plotting performance with rejection more accurately demonstrates the performance that the user would experience. The third curve of Figure 2 shows the performance of the combined speech and speaker recognizers with rejection, labeled as "AllSystemErr after resorting with rejection". The last curve shows the corresponding reprompt rate, or the percentage of times the users are reprompted (equal to the percentage of utterances rejected). This curve is labeled as "RepromptRate". It is interesting to note that with reprompting, the NL error rate can be reduced to 6.9% with a modest 7% reprompt/reject rate.

## 4. CONCLUSIONS

This paper presented a general framework for the integration of speaker and speech recognizers. By formulating the problem as a joint maximization of the *a posteriori* probability of the word sequence and speaker, an expression can be derived that clearly shows how to properly combine a speaker-independent speech recognizer, a text-dependent speaker recognizer, a speaker-dependent language model, and a term representing the prior probability of the speaker. An efficient multipass N-Best search method was developed to implement the maximization, and it was applied to an over-the-telephone, name-based speech recognition task, resulting in a 34% reduction in the NL error rate.

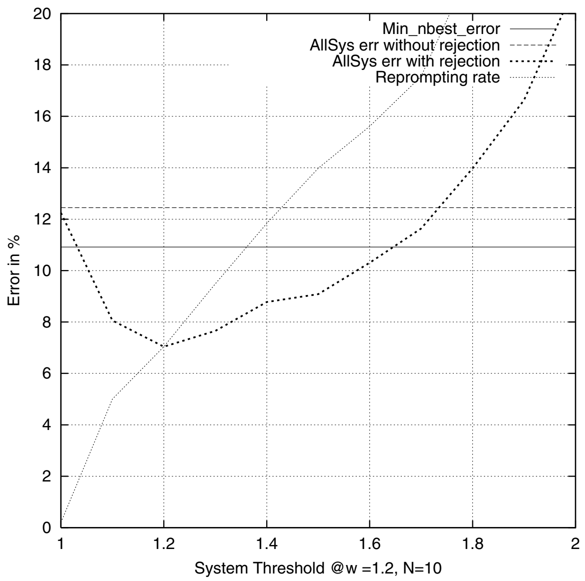

**Fig. 3**. *System Performance Versus Rejection Threshold. The system performance for an N-best size of 10 entries is shown for varying rejection thresholds. All the combined scores $\Lambda_T$ below the threshold are rejected and the utterance is reprompted. As can be seen, the benefits of the rejection threshold are present until the rejection threshold exceeds 1.7.*

Future work will explore the use of the framework developed in this paper for a variety of other applications, with a particular focus on the development of application-specific search strategies to solve the joint maximization problem.

## 5. REFERENCES

[1] L.P. Heck and D. Genoud. Integrating Speaker and Speech Recognizers: Automatic Identity Claim Capture for Speaker Verification. *Odyssey Speaker Recognition Workshop*, 2001.

[2] L.P. Heck. The Role of LVCSR in Speaker Detection: Speaker-Dependent Word Usage. *DoD Site Report*, Feb 1998.

[3] G. Doddington. Speaker Recognition Based on Idiolectal Differences Between Speakers. In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, pages 2521–2524, 2001.

[4] L.P Heck and R. Teunen. Secure and Convenient Transactions with Nuance Verifier. *Nuance Users Conference*, April 1998.

[5] Q.Li and B. Juang. Speaker Verification Using Verbal Information Verification for Automatic Enrollment. In *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, May 1998.

[6] V. Digalakis, P. Monaco, and H. Murveit. Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers. *IEEE Trans. on Speech and Audio Proc.*, pages 281–289, July, 1996.