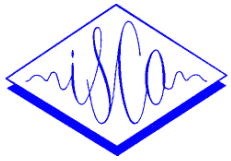

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

Integrating Speaker and Speech Recognizers: Automatic Identity Claim Capture for Speaker Verification

Larry Heck, Dominique Genoud

Nuance Communications, Menlo Park, CA 94025, USA

heck@nuance.com genoud@nuance.com

Abstract

This paper presents a novel approach to the integration of a speech and speaker recognizer for the purpose of automatically capturing an identity claim of a user. The approach integrates the speaker recognition score into the search process of the speech recognizer resulting in a best hypothesis that jointly optimizes the probability of the word sequence and the speaker. This facilitates the use of a natural speech-based interface, where the identity claim can be ambiguous and relatively difficult to recognize (e.g., names). This paper presents a theoretical framework for the integration of speech and speaker recognition systems. In addition, experimental results are presented that show a 35% reduction in the NL-error rate of an over-the-telephone speech recognition task, where the testset consists of users from a US city of size 1 million identifying themselves by simply speaking their name.

1. Introduction

One of the most challenging tasks for commercial speaker verification systems is to design a natural, convenient interface for capturing the identity claim of a user. For telephone applications, many of the current systems rely on a DTMF-based approach, where the user claims their identity by entering their account number through the telephone's touch-tone keypad. After the identity claim is established, the system verifies the claim by asking the user to speak a phrase (e.g., password, random phrase) and then scores this utterance against a speaker model of the user created in a previous enrollment session.

While a DTMF-based approach to capturing the identity claim of a user has been widely adopted, we have observed in our trials/deployments that a majority of people prefer a more natural, speech-based interface that allows them to simply speak their identity claim. Automatic speech recognition systems can be used to capture the identity claim, but often recognition performance is poor over large populations. This is particularly true with personal names as well as, in some cases, telephone numbers (in high noise environments). In addition, the identity claim is often not unique over large populations (e.g., John Smith), which introduces a further complication of how to resolve this ambiguity without requiring the user to provide additional information.

Typical speech-based approaches to capture the identity claim have solely relied on general speech recognition technology. However, given that the the identity claim is spoken by the individual associated with the claim, speaker recognition/verification technology could by utilized to determine if the voice of the talker matches the speaker model associated with the identity claim. Somewhat related work exists in [4, 5], where the approaches attempted to use speaker recognizers to effectively "quantize" the speaker into speech recognition acoustic models built on similar sounding speakers. However, these approaches were focused on reducing speaker mismatch in the speech recognizer, and not focused on solving the problem of using the speaker recognizer to improve the performance of identity claim capture.

This paper presents a novel approach to the integration of a speech and speaker recognizer for the purpose of automatically capturing an identity claim of a user. Section 2 formulates the problem of identity claim capture in the more general context of integrating a speech and speaker recognition system. Sections 3 and 4 briefly describe the speaker and speech recognition systems used in this study. Section 5 presents the computational impact of integrating a speech and speaker recognizer for identity claim capture. Finally, experiments for both digit and name-based identity claim over large populations are presented in Section 6.

2. Mathematical Formulation

The problem of capturing the identity claim of a user through a speech utterance can be expressed in the more general context of integrating a speech and speaker recognition system. The goal is to find the word sequence and the speaker with the highest joint probability among all possible word sequences W and speakers S, which is conditioned on a feature vector sequence $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_T$. Every identity claim in the dictionary is mapped to a sequence of HMMs which themselves consist of *states* q, such that every word is equiva-

lent to a Markov state sequence $Q = q_1, q_2, \ldots, q_{t-1}, q_T$ [3]. Using Bayes' rule and the product rule of probability, the conditional sequence probability $P(W, S|X)$ can be broken down into four terms and simplified as:

$$\begin{aligned}
\{\hat{W}, \hat{S}\} &= \underset{W,S}{\operatorname{argmax}} P(W, S|\mathbf{X}) \\
&= \underset{W,S}{\operatorname{argmax}} P(\mathbf{X}|W, S)P(S|W)P(W) \\
&= \underset{W,S}{\operatorname{argmax}} \sum_{Q} P(\mathbf{X}|W, Q, S)P(S|W, Q)P(W, Q) \\
&\approx \underset{W,S}{\operatorname{argmax}} \sum_{Q} P(\mathbf{X}|Q, S)P(S|Q)P(W, Q) \\
&\approx \underset{W,S}{\operatorname{argmax}} \underset{Q}{\operatorname{MAX}} P(\mathbf{X}|Q, S)P(S|Q)P(W, Q) \\
&= \underset{W,S}{\operatorname{argmax}} \underset{Q}{\operatorname{MAX}} P(\mathbf{X}|Q, S)P(S|Q)P(W)P(Q|W) \\
&= \underset{W,S}{\operatorname{argmax}} \underset{Q \in Q_W}{\operatorname{MAX}} P(\mathbf{X}|Q, S)P(W)P(Q)P(S|Q) \\
&= \underset{W,S}{\operatorname{argmax}} \underset{Q \in Q_{W,S}}{\operatorname{MAX}} P(\mathbf{X}|Q, S)P(W)P(Q)P(S) \quad (1)
\end{aligned}$$

The first expression $P(X|Q, S)$ is the observation likelihood given the state sequence and the speaker model and can be computed as

$$\begin{aligned}
P(X|Q, S) &= \prod_{t=1}^{T} P(\mathbf{x}_t|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}, q_1^T, S) \\
&\approx \prod_{t=1}^{T} P(\mathbf{x}_t|q_t, S). \quad (2)
\end{aligned}$$

This component is the most expensive to compute of the four and therefore greatly influences the overall time of the search. In addition, speaker-dependent models for each state are not likely to be available due to lack of sufficient examples from each individual speaker. To complete the search in reasonable time as well as avoid using poorly trained state-dependent speaker models, we use the following approximation:

$$P(X|Q, S) \approx P(X|Q) \cdot P(X|S)^{\alpha} \quad (3)$$

where α is a scaling factor that can be empirically optimized to balance the contribution of the speaker-independent speech recognizer's score, $P(X|Q)$, and the speaker recognizer's score, $P(X|S)$.

The scores of the speech and speaker recognition systems can be combined in the search of the recognizer at either the frame/state-level (e.g., forward pass of a Viterbi search) or at the utterance-level (e.g., multi-pass rescoring of the N-Best hypotheses list). This paper will study the combination of scores at the utterance level with a multi-pass rescoring approach. With this approach, we generate a combined speaker and speech recognition score S_T for each entry of each N-Best list,

$$\begin{aligned}
S_T = \log P(X|Q) \; + \; \alpha \cdot \log P(X|S) \; + \\
\beta \cdot \log P(W) \; + \; \log P(Q) \; + \; \log P(S) \quad (4)
\end{aligned}$$

where α is a weighting on the speaker recognition score, and β is a weighting on the language model score. The other three terms in Equation (1) include the transition probability, $P(Q)$, the language model probability, $P(W)$, and finally the prior probability of the speaker, $P(S)$. The transition and language model probabilities can be computed using standard approaches in the speech recognition literature[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 be set equal for all speakers if data is unavailable.

After the combined scores S_T are computed, the N-Best list can be resorted. The *new* top hypothesis can then be selected as the most likely identity claim.

3. Speaker Recognition System Description

The speaker recognition system 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:

$$\log P(X|S) = \frac{1}{T} \sum_{t=1}^{T} \log \; p(\mathbf{x}_t|\lambda_s) \;\; - \;\; \log \; p(\mathbf{x}_t|\lambda) \quad (5)$$

where $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ denotes the set of feature vectors extracted from the utterance by the feature extraction front-end, λ_s is the speaker model (corresponding to the speaker that the caller claims to be), and λ is the background model used for normalizing the likelihood scores. Probability density functions of both speaker and background models are modeled as Gaussian mixture models (GMMs) as follows:

$$p(\mathbf{x}_t|\lambda) = \sum_{i=1}^{L} \; w_t \; p(\mathbf{x}_t|b_i) \quad (6)$$

where b_i are multi-dimensional Gaussian densities, and w_i are the corresponding weights. Each Gaussian is represented by a mean μ, and a variance σ^2.

The background models are channel- and gender-dependent. Previous work has shown that this gives improved performance over channel-independent background models for channel mismatched conditions[2]. Each background model is derived from a channel- and gender-independent root model using Bayesian adaptation.

Channel-specific speaker models are obtained by Bayesian adaptation. During the enrollment phase, the gender and channel of the speaker are first detected and then the corresponding channel and gender dependent background model is adapted using the speaker enrollment data. During the testing phase, the channel-dependent models corresponding to the test utterance are scored.

4. Speech Recognition System Description

The speech recognizer used for the experiments utilizes context dependent triphones that are modeled with clusters of Gaussian mixtures called Genones[1]. The system was trained with over a million digit strings, stock quote requests, and phonetically rich utterances collected over the telephone from various sources.

5. Computations

A major consideration in the efficacy of the approaches described in this paper is whether they can be implemented efficiently. We will focus on the multi-pass N-Best rescoring approach in the following analysis. In the original approach, recognition occurs on a single utterance (identity claim) followed by a verification against a single speaker model corresponding to the identity claim. The total computations can be expressed as

$$C_T = C_R \ + \ C_V \tag{7}$$

where C_R denotes the computations of the recognizer required to generate an N-Best list, and C_V denotes the computations required by the speaker recognition system.

The computations required by the speaker recognition system can be approximated as the number of Gaussian distance computations. To compute a speaker recognition score for a single speaker model, we have (for each 20ms frame of data)

$$C_V = 2GK \tag{8}$$

where G is the number of Gaussians in the root model and K is the number of channel-dependent models used in the system. A faster approach was described in [6] where the all channel- and gender-dependent GMMs are adapted from a common root GMM, and only the top M scoring Gaussians from the root model were scored in the adapted GMMs. This approach yields the following computations for the speaker recognition system

$$C_V^1 = G \ + \ 2MK. \tag{9}$$

For the N-Best rescoring technique described in this paper, the added computations are due to the scoring of additional speaker models on the N-Best list of the recognizer (for the multi-pass approach). Also, additional speaker models will need to be computed for each entry on the N-Best list where the identity claim is ambiguous (e.g., "John Smith"). With our new approach, the computations for the speaker recognition component can be expressed as

$$C_V^2 = G \ + \ 2MKNA \tag{10}$$

where N is the size of the N-Best list of the recognizer, and A is the level of ambiguity for the identity claim (e.g., $A = 2$ if two persons share the same identity claim). However, this can be sped up considerably by noting that the Gaussians in the background models only need to be computed once for all entries in the N-Best list. Using this fact gives the following

$$C_V^3 = G \ + \ (1 \ + NA)MK. \tag{11}$$

We will use an example to illustrate the computations of the multi-pass rescoring approach. Let the number of Gaussians in the root model $G = 2000$, the number of top Gaussians in the decoding approach $M = 5$, the number of channels $K = 6$ (2-gender X 3-handset-types), the size of the N-Best list $N = 10$, and the average level of ambiguity $A = 2$ (approximate ambiguity level for first+last names in the white pages of a US city of 1-million persons). For this example, we compare the computations of the original approach (C_V)

$$\begin{aligned} C_V &= 2GK \\ &= 2 * 2000 * 6 \\ &= 24000 \end{aligned} \tag{12}$$

to the fast scoring approach described above

$$\begin{aligned} C_V^1 &= G \ + \ 2MK \\ &= 2000 + 2 * 5 * 6 \\ &= 2060 \end{aligned} \tag{13}$$

This is a savings of 91% in computations. Building on this fast scoring approach, the N-Best rescoring technique of this paper would have computations of

$$\begin{aligned} C_V^3 &= G \ + \ (1 \ + NA)MK \\ &= 2000 \ + \ (1 \ + 10 * 2)5 * 6 \\ &= 2630 \end{aligned} \tag{14}$$

which still compares favorably to the direct approach in Equation (12), especially given that the new approach is scoring an additional $(10 * 2 - 1) = 19$ speaker models.

6. 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.

6.1. Digit-based Identity Claim Experiments

The digit-based identity claim database used for these experiments is composed of 1000 10-digit utterances spoken over long distance telephone lines in noisy environments. The utterances serve both as the customers claim of identity (i.e., their home telephone number) as well as

their verification utterance. The telephones used in the data collection include landline, portable, and cellular. The background noise in the utterances includes loud TV audio, competing talkers, and other household sounds. For these experiments, the 10-digit telephone numbers were constructed to be unique to an individual. There are 449 unique speakers in the test set (400 female, 49 male). Each person has a speaker model that was trained on a previous phone call. The speaker models were trained in a single session on two repetitions of the same 10-digit phone number. The verification task is especially difficult, with 345 of the 1000 utterances in mismatched conditions to the enrollment session (including training on landline and verifying on cellular).

The grammar used by the recognizer is a simple digit loop across 11 possible digits (zero through nine and "oh"). Each utterance has exactly 10-digits. All of the 1000 utterances in the testset are in grammar.

After completing a first pass, the speech recognizer produced an N-Best list of the top (unique) hypotheses according to the speech recognition score (log-likelihood). A speaker recognition score was computed for the correct entry in the N-Best list by scoring the spoken utterance against the speaker model associated with the correct 10-digit string. For the other entries in the N-Best list, if speaker models existed for these identities, then a score was computed for the utterance against this model. However, given the extremely large number of possible digit strings on the N-Best list (11^{10}), we did not have speaker models for most of the entries. Therefore, to simulate the performance of the new technique where every entry is a competitive identity claims, we generated a speaker recognition score by randomly choosing from the correct speakers' distribution of impostor scores.

Table 1 shows an example of an N-Best list with 10 entries. The actual spoken utterance was "8162311831". 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 scores from the speech recognizer (normalized as the difference from the top hypothesis score) 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. Here, the value of α in Equation (4) was optimized to be $\alpha = 725$. As can be seen, the combined score gives a different ranking than the speech recognizer score, with the second ranking hypothesis rising to the top. With the combined score, the system yields a correct result which would have otherwise been an error.

Figure 1 shows the NL-error rate of the combined speech and speaker recognizers as compared to the theoretical limit ("N-Best error rate") for a given size N-Best list. As discussed above, the acoustic environment is challenging for this test, with the NL-error rate for the speech recognizer at 26.4% ($N = 1$), and the equal error rate of the speaker recognizer at 5.7%. 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 10-digit number 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 with only 3 hypotheses. With an N-Best list of size 3 or greater, the new approach gives an error rate of 19.0-20.6%, which is a 22-28% improvement over the best performance of the speech recognizer alone.

Table 1: *Example of a digit-based identity claim capture with an N-Best rescoring approach. The correct transcript is "***8162311831***". The normalized score from the speech recognizer (Speech Score) is used to determine the order of the table, with the top entry having the best score. 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	Speech Score	Speaker Score	Combined Score
1	816**3**311831	0	-1.46	-1055.7
2	**8162311831**	-134	+0.19	3.4
3	816**8**311831	-232	-0.96	-925.9
4	8163311810	-329	+0.14	-227.0
5	8163311331	-375	-3.68	-3046.2
6	8163319831	-387	+0.09	-324.3
7	8160311831	-404	-1.70	-1635.5
8	8163311818	-443	-1.29	-1379.5
9	8162311810	-463	-0.35	-716.4
10	8163311821	-485	-0.67	-972.4

6.2. Name-based Identity Claim Experiments

The name-based identity claim database used for these experiments is composed of 1000 utterances of personal names (first and last name) spoken over long distance telephone lines on a relatively clean channel. 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.

Table 2 shows an example of an N-Best list for the name-based identity claim task with 10 entries. The actual spoken utterance was "chris craft". Here, the value of α in Equation (4) was optimized to be $\alpha = 325$. As shown with the digit-based identity claim task, this example has the correct hypothesis in the second position using

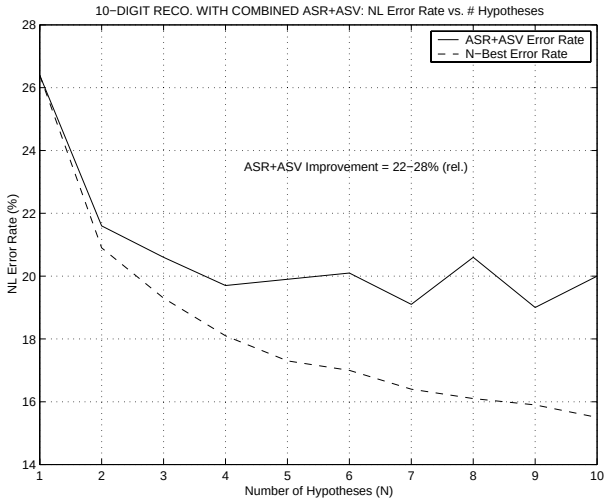

Figure 1: *Performance of the combined speech (ASR) and speaker (ASV) recognition system for the 10-digit identity claim task. This is compared to the "N-Best error rate" (theoretical limit of performance). The combined system (ASR+ASV) gives a 22-28% relative reduction in NL-error rate as compared to the speech recognition system alone.*

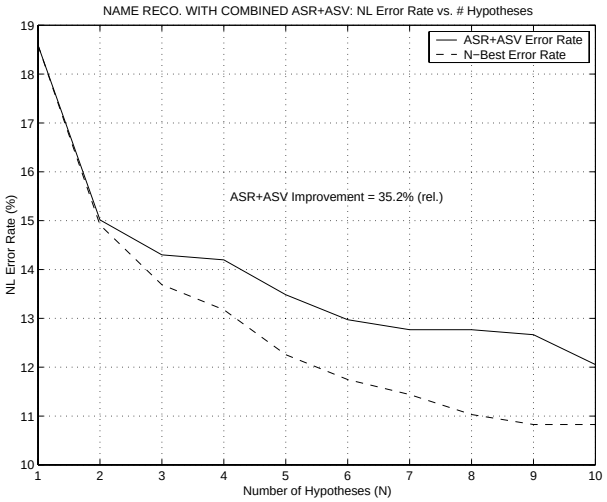

Figure 2: *Performance of the combined speech (ASR) and speaker (ASV) recognition system for the name-based identity claim task compared to the "N-Best error rate". The combined system (ASR+ASV) gives a 35.2% relative reduction in NL-error rate as compared to the speech recognition system alone.*

the recognizer score alone, but in the first (highest) position when using the combined speech and speaker recognizers.

Figure 2 shows the NL-error rate of the new rescoring approach for the name-based identity claim task as compared to the N-best error rate. The NL-error rate for the speech recognizer is 18.6%, and the equal error rate of the speaker recognizer is 4.85%. For the name task, the rescoring approach improves as the number of hypotheses on the N-Best list is increased. This suggests richer N-Best lists than in the 10-digit recognition task described earlier. For $N = 10$, the new rescoring approach with combined speech and speaker recognizers yields a 35.2% improvement over the best performance of the speech recognizer alone.

To determine the sensitivity of the rescoring approach to the combination "weight" α in Equation (4), we varied α and computed the NL-error rate of the combined system at $N = 10$, as shown in Figure 3. As can be seen, the rescoring approach is not very sensitive to the combination weight. Based on the plot, one approach could be to use the speech recognizer to compute the N-Best list, and then throw out the scores, replacing them with the scores from the speaker recognizer.

Finally, we examine the sensitivity of the rescoring approach to the accuracy of the speaker recognition system. This is accomplished by computing the distributions of the claimant and impostor test utterances, and, for a fixed false reject rate (fixed claimant score distribution), we vary the false accept rate (FAR) by moving the impos-

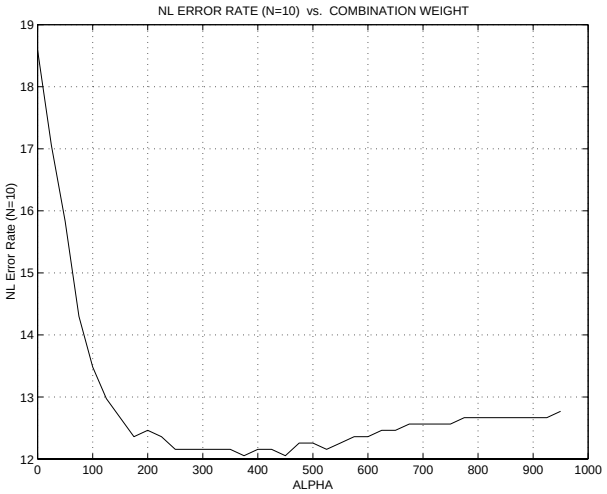

Figure 3: *Sensitivity of the combination "weight" α on the speaker recognition system to the N-Best error rate. The plot indicates that the system is not very sensitive to the combination weight.*

Table 2: *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	Speech Score	Speaker Score	Combined Score
1	chris graf	0	+1.32	428.8
2	chris craft	-60	+4.72	1475.3
3	chris krauss	-209	-0.86	-490.0
4	chris kress	-359	+1.96	278.3
5	christi crouse	-461	-1.04	-800.2
6	bruce graf	-529	-0.61	-726.6
7	craig kraft	-564	+0.18	-506.6
8	chris groves	-613	-0.98	-930.2
9	christine craft	-640	-1.33	-1073.9
10	curtis craft	-651	+0.71	-420.6

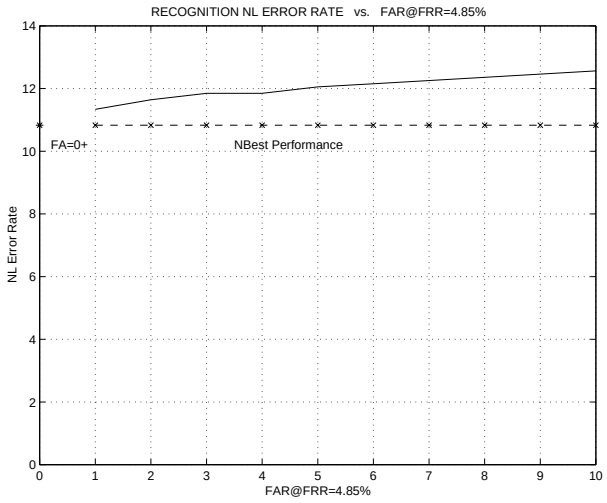

Figure 4: *Recognition versus FAR.*

tor score distribution. For each FAR, we compute the NL-error rate of the name-based identity claim with the combined speech and speaker recognizers. Figure 4 shows this NL-error rate as compared to the optimal N-Best error rate where $N = 10$. As can be seen, the rescoring approach is relatively insensitive to the accuracy of the speaker recognition system, with the speaker recognizer still contributing to the overall performance even with a FAR=10%.

7. Conclusions

This paper presented a novel approach to the integration of a speech and speaker recognizer for the purpose of automatically capturing an identity claim of a user. The approach integrates the speaker recognition score into the search process of the speech recognizer resulting in a best hypothesis that jointly optimizes the probability of the word sequence and the speaker. This facilitates the use of a natural speech-based interface, where the identity claim can be ambiguous and relatively difficult to recognize (e.g., names). This paper presented a theoretical framework for the integration of speech and speaker recognition systems. In addition, experimental results were presented that show a 35% reduction in the NL-error rate of an over-the-telephone speech recognition task, where the testset consisted of users from a US city of size 1 million identifying themselves by simply speaking their name.

Future work will investigate the use of the speaker recognizer to reject out of grammar utterances, since it is hypothesized that the speaker recognizer is a more effected rejection mechanism than the speech recognizer.

8. References

[1] 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.

[2] L.P. Heck and M. Weintraub. Handset dependent background models for robust text-independent speaker recognition. *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, 1997.

[3] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Mass., 1997.

[4] D.A. Reynolds and L.P. Heck. Integration of speaker and speech recognition systems. In *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, pages 869–872, Toronto, Canada, 1991.

[5] A.E. Rosenberg and K.L. Shipley. Speaker indentifcation and verification combined with speaker independent word recognition. In *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, pages 184–187, Atlanta, GA, 1981.

[6] R. Teunen, B. Shahshahani, and L.P. Heck. A model-based transformational approach to robust speaker recognition. In *Proc. International Conf. Spoken Language Processing*, Bejing, China, 2000.