INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken
Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Automatic Speech Recognition of Co-Channel Speech: Integrated Speaker and Speech Recognition Approach

Larry P. Heck[⋆] *and Mark Z. Mao*[†⋆]

[⋆] Nuance Communications, Menlo Park, CA, USA
[†] Department of Electrical Engineering, Stanford University, CA, USA
heck@nuance.com, markmao@stanford.edu

Abstract

This paper presents a novel Bayesian approach to the problem of co-channel speech. The problem is formulated as the joint maximization of the *a posteriori* probability of the word sequence and the target speaker given the observed speech signal. It is shown that the joint probability can be expressed as the product of six terms: a likelihood score from a speaker-independent speech recognizer, the (normalized) likelihood score of a speaker recognizer, the likelihood of a sequence of prosodic events, the likelihood of a speaker-dependent statistical language model, a prior representing the channel usage patterns of a speaker, and the prior probability of the speaker. An efficient single-pass Viterbi search strategy is presented. Experimental results on over-the-telephone recognition of co-channel speech show a 45% reduction in word error rate of a 10-digit telephone number task.

1. Introduction

Co-channel speech occurs when two or more talkers are speaking at the same time and their speech is summed into one signal. Co-channel speech is common in hands-free voice-enabled applications. Examples include information kiosks using speech recognition, information and services access from the car, and voice interaction with any application using a speaker-phone.

Traditional approaches to co-channel speech processing have focused on enhancing the target speech, attenuating the interfering speech, or a combination of both [1, 2, 3]. While the motivation for some of this previous work on co-channel speech processing was to improve automatic speech recognition (ASR), few of the previous approaches explicitly considered ASR performance in their design. It was assumed that enhancing the speech to improve SNR or human intelligibility would automatically translate into maximize speech recognition performance. But, as has been demonstrated in the robust speech recognition literature, this assumption often does not hold.

In contrast, the approach presented in this paper is designed to specifically and directly maximize speech recognition performance. Section 2 presents a novel Bayesian approach to the problem of co-channel speech, formulated as the joint maximization of the *a posteriori* probability of the word sequence and the target speaker given the observed speech signal. Based on this formulation, Section 3 describes the implementation of the resulting co-channel speech recognition system. Finally, Section 4 shows several experiments that demonstrate the effectiveness of the approach.

2. Formulation

Our goal is to develop a speech recognition system that will transcribe speech from the desired (target) speaker while ignoring speech from the interfering speaker. Stated mathematically, the goal is to find the word sequence from the target speaker that maximizes the joint probability among all possible word sequences W and speakers S, conditioned on the observations O. The observations can take many forms, such as a sequence of mel-frequency cepstral coefficient (MFCC) feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T-1}, \mathbf{x}_T\}$. Other observations that have recently shown promise in speech and speaker recognition problems include prosodic events [4, 5] represented here as a sequence of prosodic feature vectors $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{Q-1}, \mathbf{f}_Q\}$. Also, the channel C of the speaker (e.g., handset-type on the telephone) has shown to be an important observation with significant dependencies on the acoustic models in speaker recognition [10, 11]. Given this set of observations $O = \{\mathbf{X}, \mathbf{F}, C\}$, we can express the joint maximization problem as

$$\{\hat{W}, \hat{S}\} = \operatorname*{argmax}_{W,S} P(W, S_t|\mathbf{X}, \mathbf{F}, C)$$

$$= \operatorname*{argmax}_{W,S} \frac{P(\mathbf{X}, \mathbf{F}, C|W, S_t) \cdot P(W, S_t)}{P(\mathbf{X}, \mathbf{F}, C)}$$

$$= \operatorname*{argmax}_{W,S} P(\mathbf{X}, \mathbf{F}, C|W, S_t) \cdot P(W, S_t)$$

$$= \operatorname*{argmax}_{W,S} P(\mathbf{X}|\mathbf{F}, C, W, S_t) \cdot P(\mathbf{F}, C|W, S_t) \cdot P(W, S_t)$$

$$= \underset{W,S}{\operatorname{argmax}} \; P(\mathbf{X}|\mathbf{F}, C, W, S_t) \cdot P(\mathbf{F}|C, W, S_t) \cdot P(C|W, S_t) \cdot P(W|S_t) \cdot P(S_t)$$

$$= \underset{W,S}{\operatorname{argmax}} \underbrace{P(\mathbf{X}|W)}_{SI-Speech} \cdot \underbrace{\frac{P(\mathbf{X}|\mathbf{F}, C, W, S_t)}{P(\mathbf{X}|W)}}_{Speaker} \cdot \underbrace{P(\mathbf{F}|C, W, S_t)}_{Prosody} \cdot \underbrace{P(C|W, S_t)}_{Channel} \cdot \underbrace{P(W|S_t)}_{SD-LM} \cdot P(S_t) \quad (1)$$

The six components represent separate *knowledge sources*, including a speaker-independent recognizer (SI-Speech), a (normalized) speaker recognizer (Speaker), a prosody-based subsystem (Prosody), a channel detector (Channel), a speaker-dependent language model (SD-LM), and a prior for the given speaker, $P(S_t)$. Each knowledge source processes the speech at different temporal/frequency resolutions.

The features F used in the prosody-based subsystem can be at either the frame, segment, utterance, or session resolution and include statistics of pitch, and session-based speaking-rate, pause rate, and timing (see [8]). The temporal resolution of the speaker- and text-dependent channel is typically at the session level, and indicates the channel usage patterns of a user. The speaker-dependent language model uses phoneme/word/phrase resolution features, representing how a particular person chooses their words. This has been implemented with a standard N-gram statistical language model [6, 7]. The prior probability of the speaker, $P(S)$, is at the application level, and 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.

Finally,the speaker-independent speech recognition score and the speaker recognition score in (1) can be combined in the search at various resolutions, from the frame-level to the utterance-level. The combination at the frame-level could be accomplished in the forward pass of a Viterbi search. This one-pass approach would be appropriate in applications where the number of speakers is small, such as co-channel speaker separation. For other applications with 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 such as multipass rescoring are required [9].

3. System Description

Referring to Equation (1), we can make the following simplifications for the co-channel speech problem:

1. Given that the channel type is not likely to depend on the particular word-sequence that is spoken, we can assume that the channel type is only dependent on the usage patterns of the speaker, $P(C|W, S) \approx P(C|S)$.
2. The prior probability of the speaker is often not known, so the term $P(S)$ is constant and therefore can be dropped from the search.
3. While prosody is likely a useful source of information for co-channel speech recognition, we will not explore its use in this paper.
4. Likewise, while the idiosyncratic choice of words (speaker-dependent language model) is likely useful for co-channel speech recognition, we will defer this to future work.

With the simplifications,the resulting maximization problem for co-channel speech recognition can be expressed as

$$\{\hat{W}, \hat{S}_t\} = \underset{W,S}{\operatorname{argmax}} P(\mathbf{X}|W) \cdot P(W) \cdot \frac{P(\mathbf{X}|W, S)}{P(\mathbf{X}|W)} \quad (2)$$

3.1. Speech Recognition Subsystem

The speech recognition system used is described in [12]. 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 β scaling the language model score)

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

3.2. Speaker Recognition Subsystem

The score for frame i of the speaker recognition subsystem is computed as

$$\Lambda_{spkr}(\mathbf{x}_t) = \frac{1}{T} \sum_{k=i-D/2}^{k=i+D/2} \log p(\mathbf{x_t}|\lambda_{tgt}) \; - \; \log p(\mathbf{x_t}|\lambda_{\mathbf{bkg}}) \quad (4)$$

where the window size D trades off resolution for an improved (smoothed) estimate of the speaker recognition score, and λ_{tgt} and λ_{bkg} are the speaker models for the target and background talkers, respectively. The availability of prior knowledge regarding the target and background talkers depends on the application. Often, the target speaker is known while the background talker is not. In this case, a speaker-independent background model λ can be substituted for λ_{bkg}. In cases where neither talker is known *a priori*, clustering techniques can be used to initialize models directly from the test utterance. Finally, note in Equation (4) that a text-independent speaker recognizer is used. This facilitates a combination of the speech and speaker recognizers at the frame level *during the first pass Viterbi search* as detailed below.

3.3. Combined System

To combine the speaker and speech recognition subsystems for recognition of co-channel speech, the speaker recognition score is treated as a measure of *reliability*: if the speaker recognition score for a frame is low, then the frame is viewed as unreliable and its contribution to the Viterbi search is discounted. On the other hand, if the speaker recognition score for a frame is high, then the frame is viewed as reliable and it contributes fully to the search.

Expressed mathematically, we use the speaker recognition score to *weight* the frame likelihoods from the speech recognition system. The weighting is implemented by mapping the speaker recognition scores onto a sigmoid function with a range of [0,1], i.e.,

$$\Lambda_T(\mathbf{X}_t) = W(\Lambda_{spkr}(\mathbf{X}_t)) \cdot \Lambda_{speech}(\mathbf{x}_t) \tag{5}$$

where

$$W(y) = \frac{1}{1 + e^{-b(y-\theta)}} \tag{6}$$

4. Experiments

The testset consists of 1142 females speaking their 10-digit home telephone number over long-distance lines. To simulate co-channel speech, a 10-digit telephone number spoken by a male background talker was added to each of the test utterances. The 10-digit telephone numbers of the foreground and background talkers are different. The target-to-interferer ratio (TIR) of energy across the target and background utterances was set to 10dB for these experiments.

For these experiments, we used the values of $b = \infty$ (a unit step function) in Equation (6), and conducted several experiments over the entire testset to optimize the values of D and θ. First, Table 1 shows results when we have a speaker model for both the target and background talkers. The speaker model for the target and background speakers were trained on separate enrollment phone calls using 3 repetitions of the speaker's phone number. The best window size is $D = 130ms$ and the threshold is $\theta = 0.0$, giving a 45.1% error rate reduction (ERR) over the baseline.

In Table 2, we assumed we only had a speaker model of the target speaker and therefore used a speaker-independent model for the background talker in Equation (4). The best window size is shorter at $D = 70ms$ (or 5 frames in our system) and the best threshold is lower at $\theta = -0.25$. These parameters yielded a 28.6% improvement in word error rate (WER) as compared to the baseline speech recognition system.

Table 1: *The table shows results when speaker model exist for both the target and background talkers. For this case, the best window size is $D = 130ms$ and the threshold is $\theta = 0.0$, giving a 45.1% error rate reduction (ERR) over baseline*

	D (ms)	θ	ASR Errors Ins/Del/Sub	WER (%)	ERR (rel. %)
Baseline	-	-	971/971/2253	36.7	-
	30	0.0	853/853/1469	27.8	24.4
	70	0.0	780/820/1408	26.3	28.4
	130	0.0	538/568/1200	20.2	45.1
	170	0.0	557/617/1241	21.2	42.5
	520	0.0	790/1140/1732	32.1	12.8
	130	-0.5	567/577/1257	21.0	42.8
	130	-0.25	534/554/1223	20.2	45.0
	130	0.0	538/568/1200	20.2	45.1
	130	0.25	552/612/1238	21.0	42.8
	130	0.5	577/727/1282	21.6	38.4

Table 2: *The table shows results when a speaker model exists only for the target speaker. For this case, the best window size is shorter at $D = 70ms$ and the threshold is lower at $\theta = -0.25$, giving a 28.6% error rate reduction (ERR) over baseline.*

	D (ms)	θ	ASR Errors Ins/Del/Sub	WER (%)	ERR (rel. %)
Baseline	-	-	971/971/2253	36.7	-
	30	-0.25	876/876/1460	28.1	23.5
	70	-0.25	809/819/1369	26.2	28.6
	130	-0.25	789/809/1424	26.5	28.0
	170	-0.25	819/889/1476	27.9	24.2
	270	-0.25	881/1131/1542	31.1	15.4
	70	-0.5	846/856/1374	26.9	26.7
	70	-0.25	809/819/1369	26.2	28.6
	70	0.0	780/820/1408	26.3	28.4
	70	0.25	785/875/1441	27.2	26.2
	70	0.5	815/1045/1472	29.2	20.64

Tables 3 and 4 show results for an alternative approach which uses the speaker recognition scores to splice out the occluded (low speaker recognizer score) portions of the target talker's speech. In both tables, the verification score is computed exactly as was done in Table 2 using the best window size of $D = 70ms$. As can be seen, the error rate reductions were as high as 19.6% and 17.9% for waveform cutting and frame removal, respectively. These approaches were not as effective as integrating the verification scores directly into the Viterbi search as was done in Tables 1 and 2. Also, the improvement for speech removal methods degraded rapidly as the threshold on the speaker recognition score (confidence) increased.

Table 3: *The table shows results for the simple waveform cutting approach, where portions of the waveform are simply removed if the speaker verification score is below the threshold θ. As can be seen, the waveform cutting approach has an error rate reduction of up to 19.6%.*

	D (ms)	θ	ASR Errors Ins/Del/Sub	WER (%)	ERR (rel. %)
Baseline	-	-	971/971/2253	36.7	-
	70	-0.5	841/911/1622	29.5	19.6
	70	-0.25	863/1083/1699	31.9	13.0
	70	0.0	830/1500/1700	35.3	3.9
	70	0.25	836/2076/1717	40.5	-10.4
	70	0.5	799/2879/1676	46.9	-27.8

Table 4: *The table shows results for a frame removal approach, where frames of feature vectors are simply removed if the speaker verification score is below the threshold θ.As can be seen, the frame removal approach has an error rate reduction of up to 17.9%.*

	D (ms)	θ	ASR Errors Ins/Del/Sub	WER (%)	ERR (rel. %)
Baseline	-	-	971/971/2253	36.7	-
	70	-0.5	892/1002/1546	30.1	17.9
	70	-0.25	848/1168/1604	31.7	13.6
	70	0.0	882/1552/1564	35.0	4.6
	70	0.25	829/2299/1518	40.7	-10.9
	70	0.5	809/3199/1427	47.6	-24.3

5. Conclusions

This paper addressed the problem of ASR of co-channel speech by jointly maximizing the *a posteriori* probability of the word sequence and target speaker given the observed utterance. An efficient single-pass search strategy was presented. Experimental results on an over-the-telephone 10-digit recognition task with co-channel speech show up to a 45.1% reduction in word error rate when both the target and background talker had previously trained speaker recognition models. The improvement was up to 28.6% when a speaker model existed only the target speaker. An alternative approach was presented by simply using the speaker recognition scores to splice out the occluded (low speaker recognizer score) portions of the target talker's speech. The error rate reductions were as high as 19.6% and 17.9% for waveform cutting and frame removal, respectively. However, the improvement for both methods degraded rapidly as the threshold on the speaker recognition score (confidence) increased.

6. References

[1] D.P. Morgan, B. George, L. Lee, and S. Kay. Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Trans. on Speech and Audio Proc.*, 5:407–424, 1997.

[2] T.F. Quatieri and R.G. Danisewicz. An approach to co-channel talker interference suppression using sinusoidal model for speech. *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, 38:56–69, 1990.

[3] B. Smolenski and R. Yantorno. Co-channel speaker segment separation. In *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, May 2002.

[4] M.K. Sönmez, E. Shriberg, L.P. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proc. International Conf. Spoken Language Processing*, Sydney, Australia, 1998.

[5] E. Shriberg and A. Stolcke. Prosody modeling for automatic speech recognition and understanding. In *Proceedings of the Workshop on Mathematical Foundations of Natural Language Modeling*, Minneapolis, Minnesota, 2002.

[6] L.P. Heck. The role of lvcsr in speaker detection: Speaker-dependent word usage. *DoD Site Presentation*, Feb 1998.

[7] G. Doddington. Speaker recognition based on idiolectal differences between speakers. In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, pages 2521–2524, 2001.

[8] L.P. Heck. Integrating high-level information for robust speaker recognition. In *Johns Hopkins Workshop on SuperSID (http://www.cslp.jhu.edu/ws2002/groups/supersid)*, Baltimore, Maryland, 2002.

[9] L.P. Heck and D. Genoud. Integrating speaker and speech recognizers: Automatic identity claim capture for speaker verification. *Proc. Odyssey Speaker Recognition Workshop*, 2001.

[10] D.A. Reynolds. The effects of telephone transmission degradations on speaker recognition performance. *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, pages 329–332, 1995.

[11] 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.

[12] 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.