Chapter 7
A Novel Way to Start Speech Dialogs in Cars by Talk-and-Push (TAP)

Balázs Fodor, David Scheler, and Tim Fingscheidt

Abstract The obligation to press a push-to-speak button before issuing a voice command to a speech dialog system is not only inconvenient but it also leads to decreased recognition accuracy if the user starts speaking prematurely. In this chapter, we investigate the performance of a so-called talk-and-push (TAP) system, which permits the user to begin an utterance within a certain time frame before or after pressing the button. This is achieved using a speech signal buffer in conjunction with an acoustic echo cancelation unit and a combined noise reduction and start-of-utterance detection. In comparison with a state-of-the-art system employing loudspeaker muting, the TAP system delivers significant improvements in the word error rate.

Keywords Acoustic echo cancellation • Frequency-domain adaptive filter (FDAF) • Noise reduction • Automatic speech recognition • In-car speech dialog • Push-to-speak

7.1 Introduction

Modern in-car speech dialog systems require the user to press a push-to-speak (PTS) button to initiate a dialog. The button press is normally followed by an acoustic acknowledgment tone indicating that the user may start speaking.

In practice, this procedure often causes degraded system performance due to nonconforming user behavior. For example, an inexperienced user cannot be expected to wait for the acknowledgment tone before they start speaking. Instead, the start of utterance (SOU) is likely to occur before the beep or, even worse, before

B. Fodor (✉) • D. Scheler • T. Fingscheidt
Technische Universität Braunschweig, Institute for Communications Technology, Braunschweig, Germany
e-mail: Fodor@ifn.ing.tu-bs.de; scheler@ifn.ing.tu-bs.de; fingscheidt@ifn.ing.tu-bs.de

J.H.L. Hansen et al. (eds.), *Digital Signal Processing for In-Vehicle Systems and Safety*,
DOI 10.1007/978-1-4419-9607-7_7,

the PTS button has been pressed. Similarly, even experienced users may not always conform to the required sequence simply for impatience or because they are concentrating on the driving task. As a consequence, the portion of speech uttered prematurely will not be processed by the system, resulting in recognition errors.

Another source of degradation is acoustic leaking of music or speech being presented via the car audio system into the hands-free microphone. Since the automatic speech recognition (ASR) engine generally cannot distinguish such signal components from the user's voice commands, the result will be recognition errors. In many commercial systems, this problem is approached by muting the loudspeakers upon PTS button actuation. However, muting cannot be performed instantaneously, thus leaving some disturbances in the microphone signal. Moreover, it is not always advisable to mute the loudspeaker signal. For example, the car computer may need to deliver urgent voice notifications at any time, regardless of whether the system is engaged in a speech dialog.

Instead of muting, some state-of-the-art systems employ acoustic echo cancelation (AEC) methods [1, 2], which strive to estimate and remove the acoustic signal component captured by the hands-free microphone originating from the car loudspeakers. While AEC makes muting unnecessary, this method alone still does not provide for intuitive dialog initiation. An extended and more flexible solution, the so-called talk-and-push (TAP) system, has been proposed in [3]. It allows the user to start speaking within a certain time frame before or after PTS button actuation. This is achieved by employing a look-back speech buffer in conjunction with an AEC unit and a robust SOU detection. The experiments in [3] were conducted at a sampling frequency of 8 kHz and using the normalized least-mean-square (NLMS) algorithm for AEC.

In this chapter, we investigate the performance of a TAP system operating at 16 kHz sampling frequency and employing the frequency-domain adaptive filter (FDAF) as proposed in [4] for AEC. While the higher sampling rate was chosen to open the prospect of more complex ASR tasks, the FDAF offers lower computational complexity than a 16 kHz NLMS algorithm, as well as a built-in postfilter for residual echo suppression.

The remainder of this chapter is organized as follows: Section 7.2 outlines the TAP system architecture. The implementation of the system components—AEC, noise reduction, and SOU detection—is described in Sects. 7.3 and 7.4. Section 7.5 then summarizes the experimental setup, followed by a discussion of the simulation results in Sect. 7.6.

7.2 The Talk-and-Push System

We assume the typical setup of an in-car speech dialog system: It consists of a speaker (e. g., the driver) seated in a vehicle, a hands-free microphone for voice control, and an in-car loudspeaker system reproducing voice prompts or music from the FM radio. In the microphone, the speaker's speech signal s is disturbed by additive background noise n and the reverberated loudspeaker signal d. In the

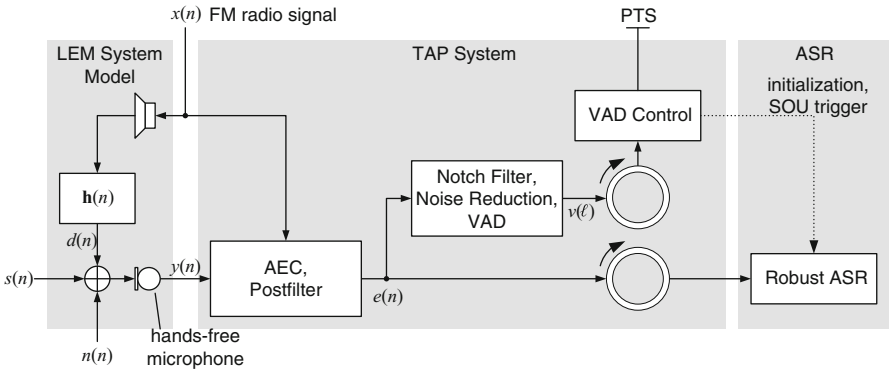

Fig. 7.1 Block diagram of the talk-and-push (TAP) system

discrete-time domain, using n as discrete-time index at sampling frequency $f_s = 16$ kHz, the microphone signal can thus be expressed as the sum:

$$y(n) = s(n) + d(n) + n(n) \tag{7.1}$$

This relation is depicted on the bottom left of Fig. 7.1.

To model the acoustic leaking from the loudspeaker into the microphone, we assume that the echo signal $d(n)$ results from the loudspeaker source signal $x(n)$ by convolution with a discrete-time, time-variant impulse response

$$\mathbf{h}(n) = [h_0(n), h_1(n), \ldots, h_{N-1}(n)]^T, \tag{7.2}$$

where N denotes the finite impulse response length and $(\cdot)^T$ is the transpose.

For simplicity, a mono source signal $x(n)$ is assumed. The impulse response $\mathbf{h}(n)$ models the entire loudspeaker–enclosure–microphone (LEM) system—i. e., the path from the digital–to–analog converter before the loudspeaker via the acoustic enclosure to the analog–to–digital converter after the microphone.

Hence, the reverberated loudspeaker signal can be written as

$$d(n) = \mathbf{h}^T(n) \cdot \mathbf{x}(n), \tag{7.3}$$

where $\cdot$ denotes the scalar product and $x(n) = [x(n), x(n–1), \ldots, x(n\text{-}N + 1)]^T$ is a time-inverted segment of the loudspeaker signal of length N.

As shown in Fig. 7.1, the first stage of the TAP system is an acoustic echo cancelation (AEC) unit. It computes an estimate $\hat{d}(n)$ of the echo component according to [4] and subtracts it from the microphone signal.

For this purpose, the LEM system transfer function is estimated using the FDAF described in Sect. 7.3. The FDAF furthermore contains a postfilter, which reduces residual echo components as well as some background noise $n(n)$ present in the microphone signal.

The resulting error signal $e(n)$ is processed in two different branches: As shown at the bottom of Fig. 7.1, it is stored in a circular buffer to be fed into the ASR engine without further processing. In the upper branch of the TAP system, it is analyzed by an integrated additional noise reduction and voice activity detection (VAD) as described in Sect. 7.4. The latter's output is a voice activity signal which is buffered and evaluated by a control unit. Upon receiving a PTS event, the control unit locates the speech onset using buffered voice activity signal both from the past and present. The control unit also initializes and triggers the ASR engine, which is then supplied with a correct portion of the error signal from the lower buffer, depending on the detected SOU.

7.3 Acoustic Echo Cancelation and Postfilter

The AEC stage of our system employs the FDAF as described in [4], which unifies AEC and a postfilter for residual echo and noise suppression in the frequency domain. While most echo cancellers model the impulse response $\mathbf{h}(n)$ of the LEM system—or its transfer function—deterministically, the FDAF is based on a statistical model.

As proposed in [4], the impulse response $\mathbf{h}(n)$ is modeled as a random process with the expectation $\mathbf{h}_0(n)$ and covariance vector $\Phi_{hh}(n)$.

Actual estimation is performed in the frequency domain. Assuming that variations of the LEM path over time are gradual, the LEM system transfer function estimate $\widehat{H}_\ell(k)$ is updated recursively according to

$$\widehat{H}_{\ell+1}(k) = A\widehat{H}_\ell(k) + \Delta H_\ell(k), \tag{7.4}$$

where ℓ is the time frame index, k is the frequency bin index, $A = 0.9995$ is the transmission factor, and $\Delta H_\ell(k)$ is the echo path update as computed according to [4].

Multiplying the estimated LEM transfer function $\widehat{H}_\ell(k)$ with a short-time Fourier transform (STFT) $X_\ell(k)$ of the loudspeaker source signal yields the estimated echo component $\hat{D}_\ell(k)$ in the short-time spectral domain. This estimate is then subtracted from the STFT $Y_\ell(k)$ of the microphone signal, resulting in an error signal $\tilde{E}_\ell(k)$. Note that before applying the STFT to the signals $x(n)$ and $y(n)$, they are subject to a high-pass filter with a cutoff frequency $f_c = 200$ Hz to remove low-frequency noise.

To reduce the noise component and to suppress the residual echo that is still present in the error signal $\tilde{E}_\ell(k)$, the FDAF includes an additional frequency-domain postfilter. Its application to the error signal yields an improved estimate of the desired speech signal as

$$E_\ell(k) = \tilde{E}_\ell(k) \times W_\ell(k), \tag{7.5}$$

where the postfilter is given by the generalized Wiener filter

$$W_\ell(k) = \frac{\Phi_{ss,\ell}(k)}{\Phi_{ss,\ell}(k) + |X_\ell(k)|^2 \times \Phi_{hh,\ell}(k) + \Phi_{nn,\ell}(k)}, \tag{7.6}$$

with $\Phi_{ss,\ell}(k)$, $\Phi_{hh,\ell}(k)$, and $\Phi_{nn,\ell}(k)$ denoting the power spectral density (PSD) of the desired speech signal $s(n)$, the echo path covariance in the frequency domain, and the PSD of the background noise $n(n)$, respectively. Since the covariance $\Phi_{hh,\ell}(k)$ can be taken as an uncertainty measure of the LEM system identification, the product $|X_\ell(k)|^2 \times \Phi_{hh,\ell}(k)$ represents the PSD of the residual echo. The PSDs $\Phi_{ss,\ell}(k)$ and $\Phi_{nn,\ell}(k)$ are estimated according to [4]. Finally, the postfilter gain $W_\ell(k)$ is floored to $W_{\min} = -12.6$ dB.

7.4 Integrated Noise Reduction and Voice Activity Detection

Subsequent to echo cancelation, residual vehicle noise $n(n)$ as well as some remains of the beep may still be contained in the error signal $e(n)$. In the upper path of the TAP system, robust detection of the speech onset therefore requires these disturbances to be distinguished from the desired speech component $s(n)$. This problem is here approached with a combined additional noise reduction and VAD operating on the short-time spectrum $E_\ell(k)$ of the error signal.

For the removal of the beep, all frequency bins corresponding to the frequency range from about 1.83 to 2.45 kHz are set to zero. For each frame ℓ and frequency bin k, the estimated clean speech spectrum $\hat{S}_\ell(k)$ is obtained from the error signal $E_\ell(k)$ by applying a Wiener filter based on the a priori signal-to-noise ratio (SNR) as described in [5] and [6]. For the computation of this SNR, the power spectral density of the noise is estimated by employing a 3-state time- and frequency-dependent VAD [3].

The output of the VAD is transformed into a per-frame voice activity signal $v \in [0, 1]$ by averaging over relevant frequency bins (see [3]) and then stored in the upper circular buffer as shown in Fig. 7.1. The final decision about the time of the speech onset is made by the VAD control unit: The hypothesized speech onset frame ℓ_{SOU} is the latest nonspeech frame (i. e., $v(\ell_{\text{SOU}}) \approx 0$) before $v(\ell)$ exceeds an empirical threshold [3].

7.5 Experimental Setup

For experimental evaluation, we performed an offline batch simulation of the TAP system using the Cambridge Hidden Markov Model Toolkit (HTK) for ASR. Instead of a physical LEM system, we used a digital LEM impulse response measured inside a vehicle. In the next two subsections, the near-end speech files as well as the noise and echo signals are described.

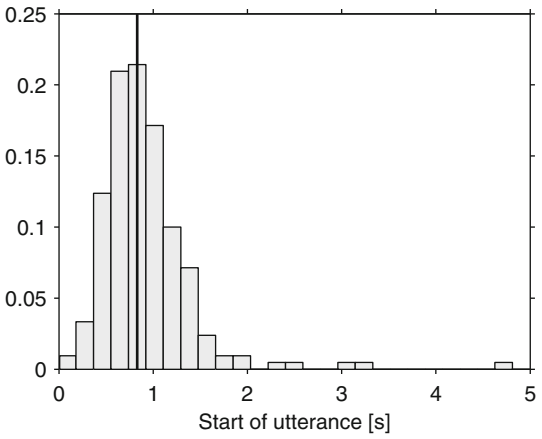

Fig. 7.2 Normalized histogram of the start of utterance (SOU) with respect to the beginning of the speech file; the thick black line marks the median of 0.83 s [3]

For reference, we performed a similar experiment where the TAP system is replaced by the state of the art: Upon PTS button actuation, the in-car audio system is muted—i. e., no echo component is added at the microphone—and the unprocessed microphone signal is passed to the ASR engine. Any speech parts preceding the PTS event are discarded because there are no look-back buffers.

7.5.1 Test Speech Data

The test speech data consisted of a subset of the US-English SpeechDat-Car connected-digit corpus [7]. The set comprised 210 utterances spoken by 35 speakers, each utterance containing four to sixteen digits. Since the test files were artificially degraded with background noise (see next section), we used close-talk recordings only, which approximately represent clean speech.

As described in [3], PTS actuation was assumed to occur 0.83 s relative to the beginning of each test speech file. Since the actual time of the speech onset varied from file to file, a probabilistic displacement of the SOU with respect to the PTS event was achieved. The histogram in Fig. 7.2, which was generated by forced Viterbi alignment, visualizes the distribution of the speech onset we found in the test speech files. By assuming the PTS event at the median of the SOUs, both premature and delayed speech were simulated.

7.5.2 Artificial Degradation with Echo and Noise

We used different loudspeaker source signals to excite the LEM system as well as a set of vehicle noise files to simulate the disturbance of the desired speech on the microphone. Two different simulations were performed: In one case, the loudspeaker

signal $x(n)$ contained only music, which was randomly chosen from six files of varying musical styles. In the other case, $x(n)$ consisted of speech files, which were randomly chosen from 96 speech files taken from the English subset of the NTT-AT Multilingual database; the files were spoken by four female and four male speakers. In addition, a beep signal at 2.1–2.4 kHz was added to all loudspeaker source signals 0.25 s after the virtual PTS event. In the baseline reference case, however, no beep was added because we assumed strict muting of the loudspeakers. To obtain the simulated echo signals $d(n)$, the loudspeaker source signals were convolved with a time-invariant LEM system impulse response measured in a Volkswagen Passat car type.

For simulating the background noise component $n(n)$, four different vehicle noise files recorded in two different cars at two different velocities were used randomly.

Noise and echo components were added to the test speech signals at different signal-to-noise ratios (SNRs) and signal-to-echo ratios (SERs), respectively. By this means, we were able to investigate the system behavior under varying disturbance conditions. As in [3], we performed the SNR and SER adjustment based on the active speech level (ASL) according to ITU-T recommendation P.56 [8]. However, all signals were subject to a 50–7,000 Hz band-pass filter prior to the P.56 level measurement to eliminate speech-irrelevant frequency components.

7.5.3 Automatic Speech Recognition Setup

The ASR experiments were conducted using a feature extraction frontend for mel-frequency cepstral coefficients (MFCCs) and a set of hidden Markov models (HMMs) trained on American English connected-digit strings.

The frontend settings were as follows: A pre-emphasis value of 0.9, a frame shift of 10 ms, a frame length of 25.6 ms, a Hamming window, and a 512-point FFT. No noise reduction was applied in the frontend, but the HMMs were trained on recordings containing slight vehicle noise. For each frame, twelve MFCCs (without the zeroth coefficient) were computed using 26 uniform, triangular filterbank channels on the mel scale and ignoring frequencies below 50 Hz and above 7 kHz. A log energy coefficient as well as first and second order time derivatives were appended. Cepstral mean normalization was performed separately for each utterance.

For acoustic modeling, we employed 42 tied-state HMMs representing acoustic–phonetic units, differentiating also by the immediate left and right context via triphone modeling within words. Each HMM consisted of one to three emitting states, each of which was assigned a continuous output probability density function modeled by a Gaussian mixture model with 32 components each. Diagonal covariance matrices were assumed. The training material consisted of 3,325 utterances spoken by 245 speakers and was taken from the connected-digit corpus of the US-English SpeechDat-Car database [7]; to ensure speaker independence, two disjunct sets of speakers were used for training and testing.

Recognizing the undegraded set of test utterances with the trained HMM set yielded a word error rate (WER) of 0.59%, which posed a lower bound to the remaining recognition experiments.

7.6 Results

Our experimental results are summarized in Table 7.1, which lists the obtained WERs in % for different disturbance conditions. In case (a), the echo signal was music, whereas in case (b), the echo signal was speech. For reference, the lines labeled "Muting" contain the results obtained with the baseline system. Since this system was assumed to mute the car loudspeakers instantly upon receiving a PTS event, its performance is independent of echo type and SER. Note that the baseline results must be interpreted with care as they strongly depend on the timing of the PTS event relative with the SOU. If, in practice, more speakers than the assumed 50% start speaking *after* PTS actuation, better baseline performance will result. Nevertheless, an actual state-of-the-art system may suffer from additional impediments not considered here: For example, the muting of the loudspeakers will occur with additional delay; moreover, the beep would not be omitted in practice.

The results in Table 7.1 show that the TAP system outperforms the reference system under all test conditions. In the absence of noise SNR $\rightarrow \infty$, the TAP system yields WERs of 0.73–2.29%, which is much closer to the limit of 0.59% than the 4.20% WER obtained in the reference case. Moreover, the dependence on the SER is negligible for SER $< \infty$, indicating that the AEC works reliably even when there is noise. This seems to be a major advantage over the NLMS algorithm when considering the results obtained in [3] and might be attributed to the residual echo

Table 7.1 WER in % achieved with the TAP system under different SNR and SER conditions. For comparison, the performance of a state-of-the-art system employing muting is included

		SNR [dB]						
		−5	0	5	10	15	20	∞
	Muting	73.41	37.90	14.93	7.17	5.02	4.54	4.20
(a) Echo signal is music								
	0	43.22	22.83	10.29	5.02	2.88	2.24	1.90
SER [dB]	5	42.83	22.83	10.44	4.83	2.98	2.34	1.95
	10	42.73	22.49	10.59	4.88	2.88	2.29	1.95
	∞	43.85	24.63	11.71	6.10	3.27	2.68	0.73
(b) Echo signal is speech								
	0	43.02	22.39	10.63	5.32	3.17	2.39	2.29
SER [dB]	5	43.46	22.39	10.68	4.88	3.02	2.34	2.10
	10	42.98	22.54	10.78	5.12	2.93	2.20	2.49
	∞	43.85	24.63	11.71	6.10	3.27	2.68	0.73

suppression of the postfilter. However, the TAP system exhibits decreased performance when there is background noise but no echo signal (SNR $< \infty$, SER $\rightarrow \infty$); this may indicate that in the absence of LEM excitation, the operation of the postfilter is suboptimal.

When judging the SNR dependence of the TAP system, note the following: Since the test speech files were close-talk recordings made in a vehicle environment, they are not entirely clean with respect to background noise. As a consequence, the SNR values shown in Table 7.1 are biased towards higher values as they only reflect the amount of noise added artificially.

7.7 Conclusion

We have investigated the performance of a so-called TAP system, which tolerates imperfect user behavior when initiating a speech dialog. As in [3], we have demonstrated that the TAP system significantly improves recognition performance assuming that half of the users actuate the push-to-speak button shortly after they start speaking. This is achieved by means of two synchronized circular buffers providing a look-back capability and a robust speech onset detection. We have included an AEC and noise reduction unit operating in the frequency domain to eliminate loudspeaker signal as well as background noise leaking into the microphone. Further investigations will include AEC for multichannel source signals as well as improved methods to measure the SNR and SER. In addition, more complex ASR tasks will be evaluated using the TAP system.

References

1. Shozakai M, Nakamura S, Shikano K (1998) Robust speech recognition in car environments. In: Proceedings of ICASSP'98, Seattle, pp 269–272
2. Matassoni M, Omologo M, Zieger C (2003). Experiments of in-car audio compensation for hands-free speech recognition. In: 2003 IEEE workshop on automatic speech recognition and understanding. pp 369–374
3. Fodor B, Scheler D, Suhadi S, Fingscheidt T (2009) Talk-and-Push (TAP) – towards more natural speech dialog initiation. In: AES 36th international conference, Dearborn
4. Enzner G, Vary P (2006) Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones. Signal Process 86(6):1140–1156, Elsevier
5. Scalart P, Filho J (1996) Speech enhancement based on a priori signal to noise estimation. In: Proceedings of ICASSP 1996, Atlanta, pp 629–632
6. Ephraim Y, Malah D (1984) Speech enhancement using a inimum Mean-square Error Short-time Spectral Amplitude Estimator. IEEE Trans Acoustics Speech Signal Process 32(6):1109–1121
7. Moreno A, Lindberg B, Draxler C, Richard G, Choukri K, Euler S, Allen J (2000) SpeechDat-Car: a large database for automotive environments. In: Proceedings of LREC 2000, Athens
8. International Telecommunication Union (1993) ITU-T recommendation P.56