Reconhecimento de fala em sistemas veiculares

Speech recognition in the car environment

Tatiane M. Vital, Carlos A. Ynoguti.

Como citar esse artigo. Vital TM, Ynoguti CA. Reconhecimento de fala em sistemas veiculares. Revista Teccen. 2015 Jul./Dez.; 08 (2): 45-52.

Resumo

Os sistemas de reconhecimento automático de fala funcionam bem em ambientes pouco ruidosos, mas seu desempenho cai drasticamente na presença de ruído. Neste contexto, propõe-se o emprego da técnica de adaptação baseada no critério de Máximo a Posteriori combinado ao treinamento multi-estilo com o intuito de minimizar os efeitos e as variabilidades indesejadas causadas pelo ruído de fundo.

Palavras-chave: Reconhecimento Automático de Fala. Máximo a Posteriori.

Abstract

Automatic Speech Recognition Systems work well in environments with few noise, but its performance drops dramatically when there are noisy interferences. In this context, we propose the use of Maximum a Posteriori Adaptation combined with multi training condition in order to minimize effects and unwanted behavior from background noise.

Keywords: Automatic Speech Recognition. Maximum a Posteriori.

Introdução

Hoje em dia, diversos dispositivos têm sido desenvolvidos para auxiliar e trazer comodidade ao motorista. Dentre estes, pode-se citar: computador de bordo que fornece informações de temperatura interna ou externa, temperatura do motor, nível de óleo, prazos de manutenção do veículo, avarias no motor, estado de sensores e lâmpadas, sistemas de localização (GPS - *Global Positioning System*), aparelho de som, conexão bluetooh, entre outros. O maior inconveniente é o número de botões que têm sido introduzidos no painel e volante para acionamento e controle desses equipamentos. Dada a importância da concentração do motorista no ato de dirigir, verifica-se que a utilização de comandos de voz pode contribuir para a segurança do mesmo, visto que poderá minimizar as distrações durante a operação destes dispositivos. Além de evitar a distração, a voz é uma alternativa relevante no que diz a respeito de controle de equipamentos veiculares para deficientes físicos.

A aplicação da voz para acionamento de dispositivos veiculares é objeto de estudo de diversas pesquisas e tem se mostrada promissora (Schless & Class 1997; Saruwatari *et al.*, 2003; Buera, Lleida, Miguel & Ortega, 2004). O grande desafio para que esta tecnologia possa ser implementada com sucesso é garantir a sua robustez ao ruído interno ou externo do próprio carro, bem como, fontes sonoras externas provenientes de outros veículos (Saitoh *et al.*, 2005; Faubel *et al.*, 2011; Li, Seltzer & Gong, 2012).

Diversas pesquisas e estudos têm proposto melhoria nas abordagens existentes, bem como, o desenvolvimento de novos métodos e algoritmos com o objetivo de diminuir a sensibilidade destes sistemas ao ruído. Neste trabalho, avaliamos a eficácia da combinação de treinamento multi-estilo e adaptação baseada no critério do Máximo a Posteriori (MAP –

Instituto Nacional de Telecomunicações.

Maximum a Posteriori), técnicas que têm sido propostas para agregar robustez e imunidade aos Sistemas de Reconhecimento Automático de Fala (ASR - Automatic *Speech Recognition*) em condições ambientais adversas (Ali, Haider & Pathan, 2012; Valério, 2009; Bippus, Fischer & Stahl, 1999; Gelin & Junqua, 1999).

Este trabalho está estruturado em 4 seções. Na próxima seção são apresentados importantes características do sistema de reconhecimento e o modelamento da técnica de adaptação proposta. A terceira seção apresenta os resultados dos testes realizados, bem como, a contribuição que diferentes técnicas podem proporcionar. A seção final tece comentários baseados nos resultados obtidos.

Técnicas de Robustez ao Ruído

O desempenho de sistemas ASRs degrada quando estes operam em condições ruidosas, sendo a principal causa o descasamento acústico entre as condições de treinamento e teste (Furui, 2007).

Há diversas técnicas propostas na literatura para combater este efeito, e neste trabalho foram empregadas as técnicas de treinamento multi-estilo e adaptação baseada no MAP.

Treinamento Multi-estilo

Com o intuito de agregar robustez aos Modelos Ocultos de Markov (HMM – *Hidden Markov Models*) às variabilidades provenientes do meio, distorções do canal, reverberação, entre outros efeitos indesejados, esta abordagem baseia-se na disponibilidade de uma base de dados de fala corrompida por variados modelos de ruído cuja recorrência seja mais observada nas diversas atividades humanas. Porém, a construção de uma base que retorne um ganho para todas as situações torna-se inviável dado à variabilidade de condições adversas encontradas no mundo real.

A técnica multi-estilo, também conhecida como multi-condição, emprega versões de locuções corrompidas artificialmente por diferentes tipos e níveis de ruído na etapa de treinamento com o intuito de minimizar a queda de desempenho dos sistemas ASRs operando em ambientes ruidosos (Lippmann, Martin & Paul, 1987). Os ruídos a serem utilizados para gerar a base de treinamento podem ser os mais comumente encontrados no dia a dia, como por exemplo, os encontrados dentro de meios de transporte (carro, trem e metrô), locais públicos (rua, aeroporto e restaurante) entre outros.

Adaptação Baseada no MAP

A técnica de adaptação baseada no critério de Maximo a Posteriori (MAP), também conhecida como adaptação Bayesiana, mapeia os modelos acústicos do ambiente de treinamento para o modelo acústico do ambiente de reconhecimento. Em geral, os métodos de adaptação dos modelos acústicos fornecem melhores resultados que as técnicas que mapeiam o espaço característico do vetor de reconhecimento para o espaço característico de treinamento, pois possibilitam o modelamento da incerteza causada pelas estatísticas ruidosas do meio (Buera *et al.,* 2007).

A principal motivação do emprego da adaptação Bayesiana do modelo canônico gerado a partir do treinamento multi-estilo é alcançar um desempenho superior através da adaptação do modelo para determinado tipo e nível de ruído na etapa de reconhecimento.

Um modelo canônico é o HMM gerado a partir do treinamento utilizando locuções limpas ou corrompidas de vários locutores. Características acústicas do ruído inerente do meio ao qual o sistema ASR está operando são utilizadas para realizar a adaptação destes modelos. O modelo de fala hipotético é derivado da adaptação dos parâmetros do modelo canônico a partir dos dados de treinamento e adaptação Bayesiana (Reynolds, Quatieri & Dunn, 2000).

As equações inerentes da adaptação são descritas a seguir. Dado um modelo canônico e os vetores de amostras do ruído, $X = \{x_1, x_2,..., x_T\}$, primeiro é necessário determinar o alinhamento probabilístico dos vetores do ruído para as componentes do modelo, ou seja, para a mistura *i* no modelo canônico, tem-se:

$$\Pr(i|x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j x_t} \quad (1)$$

onde *M* é o número de observações, ω e *p* correspondem ao peso e à função densidade de probabilidade associados a uma determinada gaussiana.

A partir da Equação 1, é possível calcular os novos parâmetros estatísticos que serão utilizados posteriormente para atualizar o modelo canônico velho, obtendo-se o modelo adaptado.

$$n_i = \sum_{t=1}^{T} \Pr(i|x_t) \quad (2)$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t) x_t \quad (3)$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t) x_t^2 \quad (4)$$

onde n_i é o peso, $E_i(x)$ é a média e $E_i(x^2)$ é a variância.

Portanto, os modelos adaptados para peso ($\hat{\omega}_i$), média ($\hat{\mu}_i$) e variância ($\hat{\sigma}_i$) serão obtidos pela Equações 5, 6 e 7, respectivamente.

$$\hat{\omega}_i = \left[\alpha_i^{\omega} n_i / T + \left(1 - \alpha_i^{\omega}\right)\omega_i\right]\gamma \quad (5)$$

$$\hat{\mu}_i = \alpha_i^{m} E_i(x) + \left(1 - \alpha_i^{m}\right)\mu_i \quad (6)$$

$$\hat{\sigma}_i = \alpha_i^{v} E_i(x^2) + \left(1 - \alpha_i^{v}\right)\left(\sigma_i^2 + \mu_i^2\right) - \hat{\mu}_i^2 \quad (7)$$

O coeficiente de adaptação (α) controla a quantidade de informação do ruído que será introduzida no modelo. O intuito é introduzir informações do ruído de forma a aproximar o modelo da fala que será utilizado como referência no classificador ao modelo da fala capturada no ambiente minimizando o descasamento entre as condições de treinamento e testes. A utilização de diferentes valores de alfa permite um ajuste de adaptação diferente para os pesos, médias e variâncias. Porém, o ganho proporcionado é pequeno, dessa forma optou-se por utilizar um coeficiente de adaptação único para todos os parâmetros ($\alpha_i^{\omega} = \alpha_i^{m} = \alpha_i^{v}$) (Reynolds, Quatieri e Dunn, 2000). A Seção 3 descreve como se utiliza os conhecimentos abordados na implementação de um sistema real.

Material e métodos

Nesta seção é descrito o sistema de reconhecimento de fala, bem como, as bases de dados utilizados neste trabalho.

Mecanismo de Reconhecimento

O software baseado em Modelos Ocultos de Markov contínuos proposto em (Ynoguti & Violaro, 2000; 2001) foi utilizado para realização dos testes. Cada subunidade fonética foi modelada por um HMM de três estados, como demonstrado na arquitetura apresentada na Figura 1. Esses HMMs foram inicializados a partir do algoritmo Segmental K-Means.

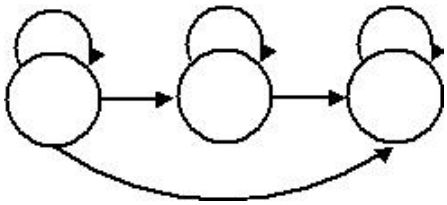

Figura 1. Modelo HMM utilizado no sistema

O sistema de reconhecimento, cujo diagrama em blocos é representado na Figura 2, emprega fones independentes de contexto e usa como algoritmo de busca o "One Step". Foram utilizados os parâmetros acústicos mel-cepstrais de ordem 12 com suas respectivas primeira e segunda derivadas (parâmetros delta e delta-delta), portanto vetores característicos de dimensão 36. Foram usadas 10 gaussianas em cada estado do HMM. Empregou-se uma gramática do tipo bigrama como modelo de linguagem. A transcrição fonética da base de dados foi realizada para cada uma das locuções, utilizando 36 sub-unidades fonéticas.

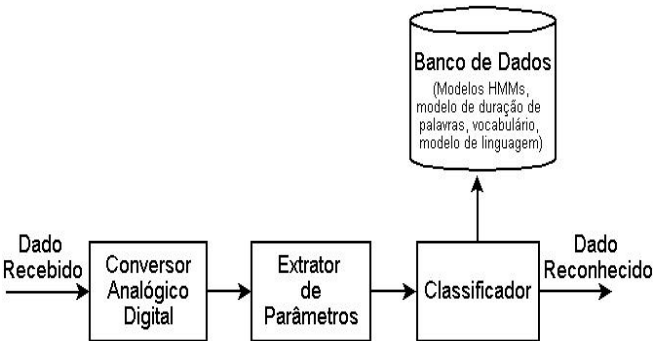

Figura 2. Diagrama em blocos do sistema ASR

Base de Dados

Para avaliação e comparação da técnica MAP com o sistema base foi utilizada a base de dados desenvolvida por (Ynoguti, 1999) baseada nas listas de locuções no idioma português foneticamente balanceadas propostas por (Alcaim, Solewicz & Moraes, 1992). A base foi construída a partir da colaboração de 40 locutores, sendo 20 do sexo feminino e 20 do sexo masculino. Esta base foi subdividida em dois grupos de forma a compor uma base para treinamento e uma para o reconhecimento contendo cada uma delas 1200 e 400 locuções, respectivamente. O conjunto conta com 694 palavras, portanto uma base de dados de médio porte.

Para avaliação da eficácia do MAP no controle de dispositivos disponíveis num veículo, foi utilizado o modelo de ruído disponível na base AURORA (Pearce & Hirsch, 2000), gerando novas versões da base de treinamento com locuções corrompidas mantendo uma relação sinal-ruído (SNR) de 15 e 20 dB e teste com locuções corrompidas de SNR 0, 5, 10, 15 e 20 dB. Os resultados dos testes que definem o desempenho do sistema são apresentados na seção a seguir.

Resultados Experimentais

Para avaliação do comportamento isolado das técnicas citadas na Seção 2, bem como, a combinação

Revista Teccen. 2015 Jul./Dez.; 08 (2): 45-52

das mesmas, realizou-se as seguintes simulações:

- na primeira etapa foi realizado o treinamento com locuções limpas seguido do reconhecimento de locuções limpas, para verificar o ganho máximo possível numa situação ideal, ou seja, avaliar o desempenho do sistema ASR na ausência de ruído;
- na segunda etapa, o HMM treinado com locuções limpas foi utilizado para os testes de reconhecimento de locuções ruidosas com o intuito de verificar a redução da taxa de acerto de palavras;
- na terceira etapa, o HMM treinado com locuções corrompidas por ruído de carro com SNR 15 dB e 20 dB foi utilizado nos testes de reconhecimento de locuções ruidosas, com a finalidade de identificar a influência do treinamento multi-estilo na resposta do sistema ASR;
- na quarta etapa, o HMM treinado com locuções limpas e adaptado para um determinado nível de ruído é empregado nos testes com locuções corrompidas, a fim de mensurar o ganho proporcionado pelo MAP;
- e, por fim, o HMM treinado com locuções corrompidas por ruído de carro com SNR 15 dB e 20 dB adaptado para um determinado nível de ruído foi utilizado nos testes com locuções corrompidas, cujo principal alvo, foi determinar o ganho proporcionado pela combinação das técnicas propostas.

Existem diversas métricas para medir a precisão no processo de reconhecimento de um sistema ASR. Neste trabalho, empregou-se a WER (Word Error Rate) para avaliar o desempenho do sistema.

$$WER = \frac{(D+S+I)}{N} * 100\% \quad (8)$$

onde D é o número de palavras excluídas, S indica o número de substituições de palavras, I representa palavras que foram inseridas durante o reconhecimento e N é o número de palavras da transcrição de referência.

A taxa de acerto de palavras (WA - Word Accuracy) em função da WER é mostrada na Equação 9.

$$WA = 100 - WER \quad (9)$$

ASR com Diferentes Formas de Treinamento

Esta seção apresenta os resultados obtidos nos testes realizados com o sistema treinado tanto para a base de dados limpa quanto para a base corrompida pelo modelo acústico de ruído de carro. A partir do treinamento com dados limpos e multi-estilo, realizou-se o teste dos dados da base de reconhecimento corrompida.

ASR Treinado com Locuções Limpas e Testado com Locuções Limpas

Para o ASR caracterizado na Seção 3, adotou-se o sistema base como referência o treinado e testado com locuções limpas. A taxa de palavras corretas, obtida pela utilização da Equação 9, foi de 75,6%.

ASR Treinado com Locuções Limpas e Testado com Locuções Corrompidas

A partir do treinamento com dados limpos, realizou-se o teste dos dados da base de reconhecimento corrompida artificialmente. Os resultados obtidos foram compilados utilizando a WER como métrica e estão apresentados na Tabela 1.

Nesta abordagem verifica-se que houve uma queda no desempenho do sistema ASR para as diferentes relações sinal-ruído, comprovando a sensibilidade do sistema à presença de ruído conforme resultados demonstrados.

ASR Treinado com Locuções Corrompidas e Testado com Locuções Corrompidas

Nesta etapa foi empregada a técnica multi-estilo no treinamento do sistema. Para tal fim, empregou-se a base de dados de treino e teste corrompida artificialmente com amostras do ruído de carro.

A Tabela 1 demonstra um ganho médio de aproximadamente 6,06% no desempenho da resposta do sistema utilizando a técnica do treinamento multi-estilo, comparado ao desempenho do reconhecimento de dados corrompidos a partir do treinamento com dados limpos.

Tabela 1. Comparação da taxa de acerto de palavras para o ASR treinado com locuções limpas e corrompidas

Relação Sinal-Ruído (dB)	WA para treinamento com locuções limpas (%)	WA para treinamento multi-estilo (%)	Ganho proporcionado pelo treinamento multi-estilo (%)
0	4,8	5,3	0,5
5	23,3	31,7	8,4
10	56,4	65,8	9,4
15	70,5	76,5	6,0
20	70,5	76,5	6,0

ASR com Adaptação

Uma outra proposta para agregar robustez ao ASR analisado foi o emprego da técnica MAP. Os modelos obtidos no treinamento com dados limpos e multi-estilo foram adaptados com o modelo de ruído de

carro e, posteriormente, foram realizados os testes de reconhecimento para dados corrompidos.

Os testes de reconhecimento foram realizados variando a SNR entre 0 e 20 dB com fator de adaptação, utilizado nas Equações 5, 6 e 7, variando entre 0,05 e 0,95 com intervalos de 0,05. Vale a pena ressaltar que pequenos valores do coeficiente enfatizam o sinal original e valores maiores dão mais ênfase ao ruído.

ASR Adaptado e Treinado com Locuções Limpas

Nesta abordagem, os modelos obtidos no treinamento de dados limpos são adaptados com estimativas estatísticas do ruído de carro formando um novo modelo adaptado com base no critério MAP.

As Figuras 3-7 mostram que a adaptação gerou um aumento na taxa de acertos de palavras em uma determinada faixa de valores de alfa tomando-se como referência para cada SNR os resultados obtidos para os testes dos dados corrompidos realizados nesta etapa.

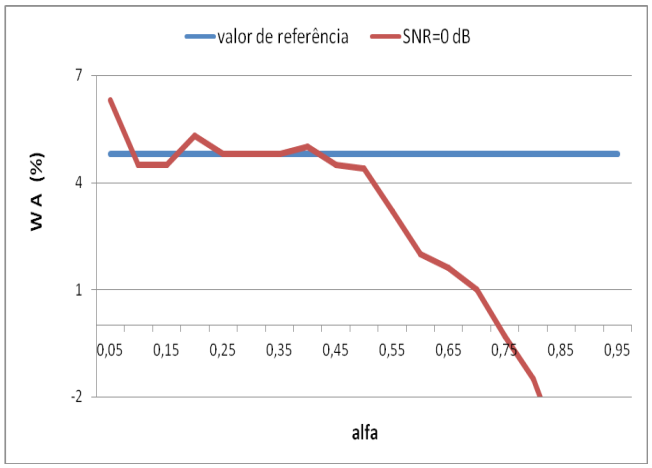

Figura 3. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 0 dB baseado na WER para ASR adaptado e treinado com dados limpos

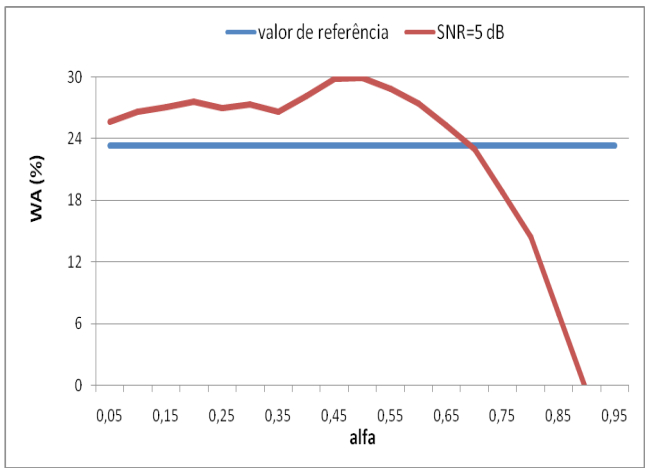

Figura 4. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 5 dB baseado na WER para ASR adaptado e treinado com dados limpos

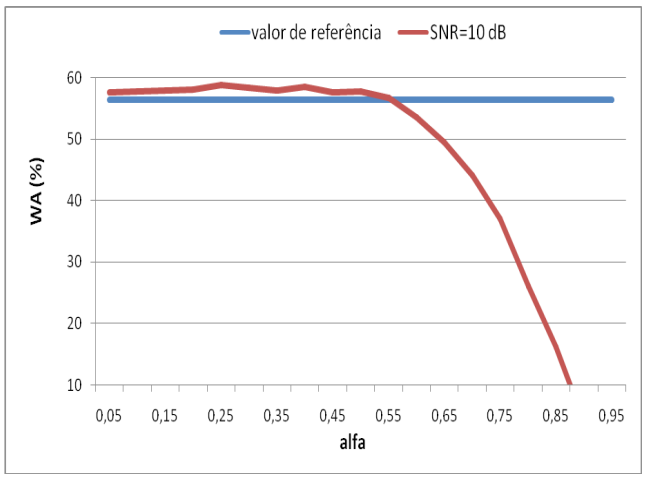

Figura 5. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 10 dB baseado na WER para ASR adaptado e treinado com dados limpos

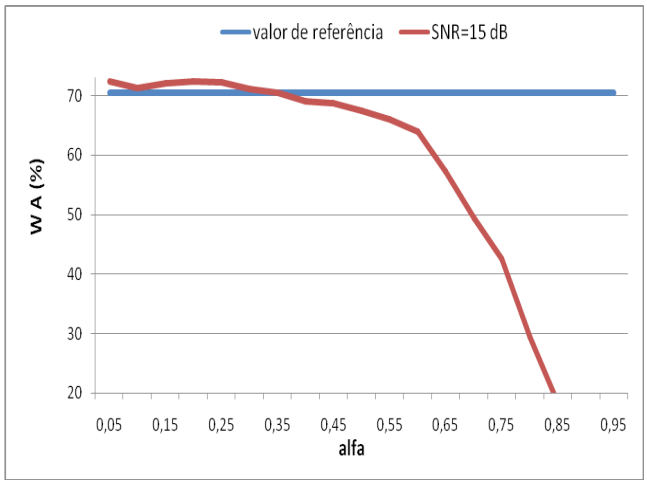

Figura 6. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 15 dB baseado na WER para ASR adaptado e treinado com dados limpos

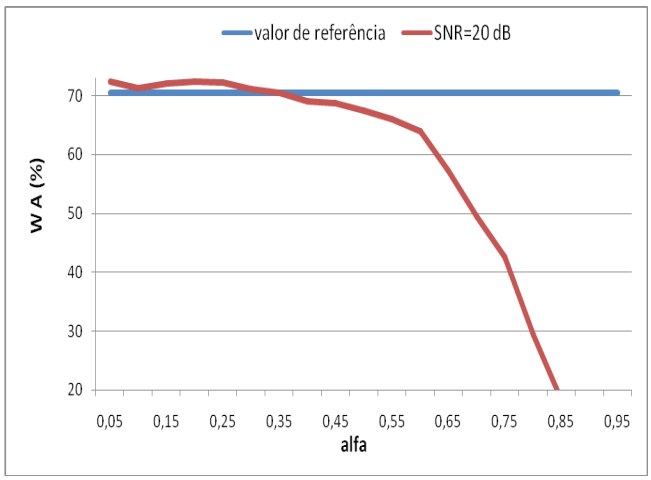

Figura 7. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 20 dB baseado na WER para ASR adaptado e treinado com dados limpos

Para análise dos resultados dos testes, considerou-se que valores da taxa de acertos acima do valor de referência são valores aceitáveis, uma vez que introduziram ganho no processo de reconhecimento. Desta forma, a partir da análise das Figuras 3-7 foi

Revista Teccen. 2015 Jul./Dez.; 08 (2): 45-52

determinada a faixa ótima de valores do fator de adaptação para cada SNR, conforme mostrado na Tabela 2.

Analisando a interseção dos valores apresentados na Tabela 2, conclui-se que a faixa de valores dos coeficientes de adaptação compreendidos entre 0,2 e 0,3 é a mais adequada, uma vez que proporciona um desempenho superior ao alcançado pelo sistema sem adaptação independentemente da SNR.

Tabela 2. Faixa ótima de valores do coeficiente de adaptação para cada SNR do sistema treinado com dados limpos

Relação Sinal-Ruído (dB)	Coeficiente de Adaptação
0	0,05 e 0,2-0,4
5	0,05-0,65
10	0,05-0,55
15	0,05-0,3
20	0,05-0,3

Tabela 3. Comparação da taxa de acerto de palavras para o ASR treinado com locuções limpas e adaptado

Relação Sinal-Ruído (dB)	WA para treinamento com locuções limpas (%)	WA máximo para sistema adaptado e treinado com locuções limpas (%)	Ganho máximo proporcionado pela adaptação (%)
0	4,8	6,3	1,5
5	23,3	29,9	6,6
10	56,4	58,8	2,4
15	70,5	72,4	1,9
20	70,5	72,4	1,9

Ao se analisar os dados da Tabela 3, observa-se que a adaptação retorna um ganho médio de 2,86%.

ASR com Adaptação e Treinamento Multi-estilo

Ao se verificar o ganho nos reconhecimentos realizados com o sistema ASR a partir da adaptação utilizando treinamento com dados limpos e com o sistema treinado com a técnica multi-estilo, propôs-se a combinação das duas técnicas. As Figuras 8-12 mostram os resultados obtidos nos testes de reconhecimento para SNR variando de 0 a 20 dB com fator de adaptação de 0,05 a 0,95 com intervalos de 0,05. Nestas figuras verifica-se que a adaptação proporcionou um aumento na taxa de acertos de palavras dentro de uma determinada faixa de valores de alfa onde os valores de referência usados para cada SNR são os resultados obtidos para os testes dos dados corrompidos realizados com sistema treinado com dados limpos. Portanto, foi definida uma faixa ótima de valores do fator de adaptação para cada SNR, conforme mostrado na Tabela 4.

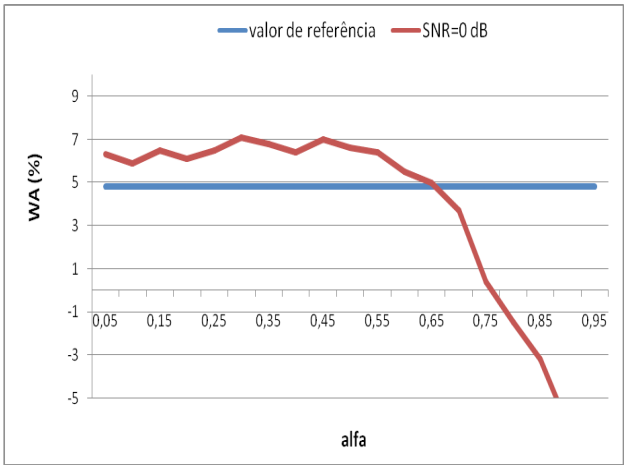

Figura 8. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 0 dB baseado na WER para ASR adaptado e treinado com dados corrompidos

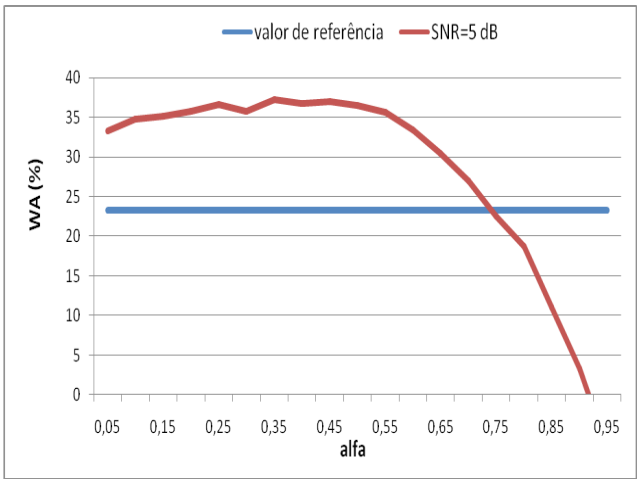

Figura 9. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 5 dB baseado na WER para ASR adaptado e treinado com dados corrompidos

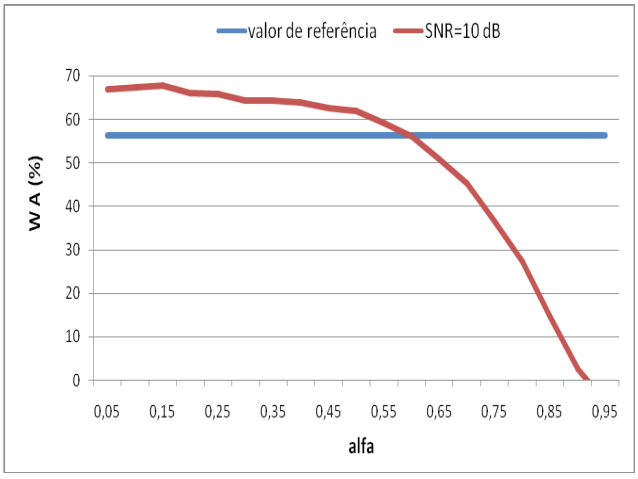

Figura 10. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 10 dB baseado na WER para ASR adaptado e treinado com dados corrompidos

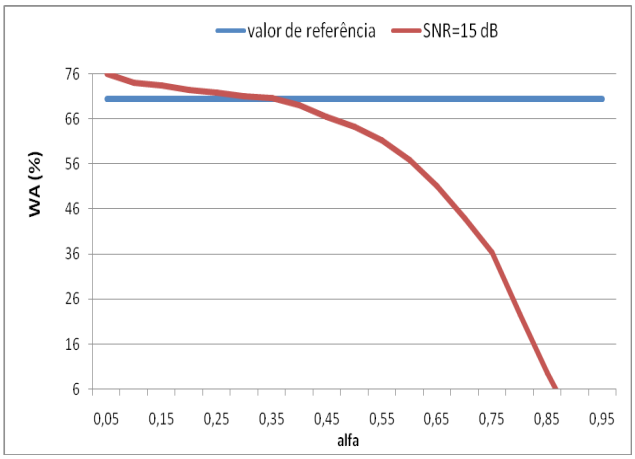

Figura 11. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 15 dB baseado na WER para ASR adaptado e treinado com dados corrompidos

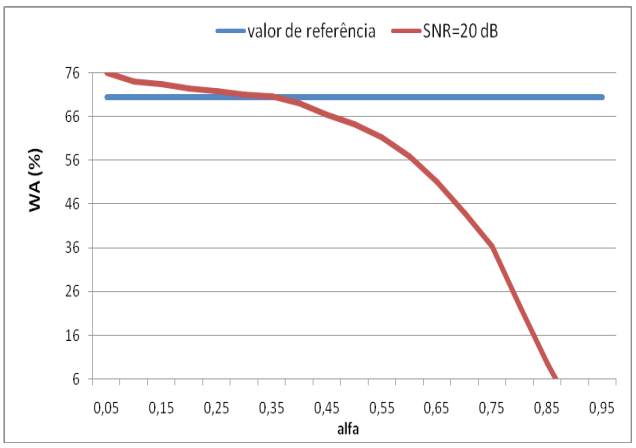

Figura 12. Taxa de acerto de palavras de testes de dados corrompidos com SNR de 20 dB baseado na WER para ASR adaptado e treinado com dados corrompidos

Tabela 4. Faixa ótima de valores do coeficiente de adaptação para cada SNR do sistema treinado com dados corrompidos

Relação Sinal-Ruído (dB)	Coeficiente de Adaptação
0	0,05-0,65
5	0,05-0,7
10	0,05-0,55
15	0,05-0,35
20	0,05-0,35

Analisando a interseção dos valores apresentados na Tabela 4, verifica-se que a faixa de valores dos coeficientes de adaptação compreendidos entre 0,05 e 0,35 é a mais adequada, uma vez que proporciona um desempenho superior ao alcançado pelo sistema sem adaptação independentemente da SNR.

Tabela 5. Comparação da taxa de acerto de palavras para o ASR treinado com locuções corrompidas e adaptado

Relação Sinal-Ruído (dB)	WA para treinamento multi-estilo (%)	WA máximo para sistema adaptado e treinado com locuções corrompidas (%)	Ganho máximo proporcionado pela adaptação (%)
0	5,3	7,1	1,8
5	31,7	37,3	5,6
10	65,8	67,7	1,9
15	76,5	76	-0,5
20	76,5	76	-0,5

Nas Figuras 8-12 mostra-se que a combinação das técnicas de adaptação com a de treinamento multi-estilo proporcionou uma melhora na eficiência da resposta do sistema ASR quando comparando ao sistema adaptado e treinado com dados limpos. Ao se analisar os dados da Tabela 5, pode-se observar que a adaptação retorna um ganho médio de 1,66%.

Tabela 6. Comparação da taxa de acerto de palavras para o ASR adaptado utilizando treinamento com locuções limpas e corrompidas

Relação Sinal-Ruído (dB)	WA para ASR adaptado e treinado com dados limpos (%)	WA para ASR adaptado e treinado com locuções corrompidas (%)	Ganho máximo proporcionado pela combinação das técnicas (%)
0	6,3	7,1	0,8
5	29,9	37,3	7,4
10	58,8	67,7	8,9
15	72,4	76	3,6
20	72,4	76	3,6

A partir da análise dos dados da Tabela 6, pode-se observar que a combinação das técnicas multi-estilo e adaptação Bayesiana proporciona um ganho médio de 4,86% comparado ao sistema treinado com dados limpos e adaptado.

Tabela 7. Comparação da taxa de acerto de palavras para o ASR treinado com dados limpos, treinamento multi-estilo, adaptado e treinado com locuções limpas e adaptado e treinado com locuções corrompidas

Relação Sinal-Ruído (dB)	WA para ASR treinado com dados limpos (%)	WA para ASR treinado com locuções corrompidas (%)	WA para ASR adaptado e treinado com dados limpos (%)	WA para ASR adaptado e treinado com locuções corrompidas (%)
0	4,8	5,3	6,3	7,1
5	23,3	31,7	29,9	37,3
10	56,4	65,8	58,8	67,7
15	70,5	76,5	72,4	76
20	70,5	76,5	72,4	76

Revista Teccen. 2015 Jul./Dez.; 08 (2): 45-52

A Tabela 7 possibilita uma comparação entre todos os testes realizados. É possível observar que o treinamento multi-estilo somado à adaptação proporciona uma melhora significativa na performance do sistema para SNR variando de 0 dB a 10 dB. Observa-se também que para valores altos de SNR (15 dB e 20 dB), a adaptação não proporciona ganho tendo em vista que por menor que seja o valor do coeficiente de adaptação, uma pequena parcela das estatísticas do modelo de ruído é introduzida no modelo canônico afastando-o do modelo hipotético.

Conclusão

De acordo com os resultados apresentados na seção anterior, verifica-se que a combinação das técnicas de treinamento multi-estilo e adaptação baseada no critério Máximo a Posteriori proporcionam ganho no desempenho do ASR na presença da distorção encontrada em ambientes veiculares. Os resultados demonstram ainda que a combinação destes métodos proporciona um aumento da taxa de acerto de palavras do sistema.

Ao se analisar os resultados obtidos nos testes realizados, uma faixa para o coeficiente de adaptação que resulta em uma melhor relação custo benefício é estabelecida. No caso do sistema treinado com dados limpos essa faixa está compreendida entre 0,2 e 0,3. Já no caso do treinamento multi-estilo, esta faixa é ampliada, estendendo-se de 0,05 a 0,35.

Referências

Alcaim, A., Solewicz, J. A., & Moraes, J. A. (1992). "Frequência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro", *Revista da Sociedade Brasileira de Telecomunicações, 7(1), 23-41*.

Ali, S., A., Haider, N., G., & Pathan, M., K. (2012). "A taxonomy-oriented overview of noise compensation techniques for speech recognition", In ARPN (Asia Research Publishing Network), *ARPN Journal of Engineering and Applied Sciences*, 7(7), 825-833.

Bippus, R., Fischer, A., & Stahl, V. (1999). "Domain adaptation for robust automatic speech recognition in car environments". In: *EUROSPEECH, 6th European Conference on Speech Communication and Technology*, Budapest, Hungria, 1943-1946.

Buera, L., Lleida, E., Miguel, A., & Ortega, A. (2004). "Multi-environment models based linear normalization for speech recognition in car conditions"; In *ICASSP, International Conference on Acoustics, Speech, and Signal Processing*. Montreal, Canadá, 1, 1013-1016.

Buera, L., Lleida, E., Miguel, A., Ortega, A., & Saz, O. (2007). "Cepstral Vector Normalization Based on Stereo Data for Robust Speech Recognition", *IEEE Transactions on Audio, Speech, and Language Processing*, 15, 1098-1113.

Faubel, F., Georges, M., Kumatani, K., Bruhn, A., & Klakow, D. (2011). "Improving hands-free speech recognition in a car trough audio-visual voice activity detection", In: *IEEE (Institute of Electrical and Electronic Engineering), Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HCMSA)*, 2, 70-75.

Furui, S. (2007). *"50 years of progress in speech recognition technology: Where we are, and where we should go? From a poor dog to a super cat"*. Keynote Presentation, ICASSP.

Gelin, P., & Junqua, J. (1999). "Techniques for robust speech recognition in the car environment", In *EUROSPEECH, 6th European Conference on Speech Communication and Technology*, Budapest, Hungria, 2483-2486.

Li, J., Seltzer, M. L., & Gong, Y. (2012). "Improvements to VTS feature enhancement", In *ICASSP, International Conference on Acoustics, Speech, and Signal Processing*. Kyoto, Japão, 4677-4680.

Lippmann, R., Martin, E., & Paul, D. (1987). "Multi-style training for robust isolated-word speech recognition", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 12, 705-708.

Pearce, D., & Hirsch, H. (2000). "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", In *ISCA (International Speech Conference Association), 6th International Conference on Spoken Language Processing*. Beijing, China.

Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000). "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, 10, 19-41.

Saitoh, D., Kaminuma, A., Sruwatari, H., Nishikawa, T., & Lee, A. (2005). "Speech extraction in car interior using frequency-domain ICA with rapid filter adaptations", In *INTERSPEECH*, Lisboa, Portugal.

Saruwatari, H., Sawai, K., Lee, A., Shikano, K., Kaminima, A., & Sakata, M.,(2003). "Speech enhancement and recognition in car environment using blind source separation and subband elimination processing", In *ICA, 4th International Symposium on Independent Component Analysis and Blind Signal Separation*. Nara, Japão, 367-372.

Schless, V., & Class, F. (1997). "Adaptive model combination for robust speech recognition in car environments", In *EUROSPEECH, 5th European Conference on Speech Communication and Technology*, 3, 1091-1094, Rhodes, Grécia.

Valério, T. A. F. (2009). *"Treinamento multi-estilo e adaptação de modelos via MAP para reconhecimento de fala em ambientes ruidosos"*. Dissertação de Mestrado. Inatel.

Ynoguti, C. A. (1999). *"Reconhecimento de fala contínua usando modelos ocultos de Markov"*. Ph. D. Universidade Estadual de Campinas.

Ynoguti, C. A., & Violaro, F. (2000). "Um sistema de reconhecimento de fala contínua baseado em modelos de Markov contínuos", In *SBrT (Sociedade Brasileira de Telecomunicações), XVIII Simpósio Brasileiro de Telecomunicações.* Gramado, Brasil.

Ynoguti, C. A., Violaro, F., (2001). "Desenvolvimento de um conjunto de ferramentas para pesquisa em reconhecimento de fala", *Revista* Telecomunicações, 4(2), 36-43.