IMPLEMENTAÇÃO EM TEMPO REAL DE UM SISTEMA DE RECONHECIMENTO DE FALA CONTÍNUA

Carlos A. Ynoguti[1] & J. Domingos Adriano[2]

Resumo *— A popularização dos computadores pessoais e a generalização de aplicações computacionais na vida das pessoas exigem o desenvolvimento de interfaces homem-máquina mais amigáveis. Das formas de comunicação do ser humano, a mais natural é a fala. Assim, interfaces baseadas na fala que permitam a comunicação em tempo real com uma máquina apresentam várias possibilidades de aplicação. Entretanto, a construção de sistemas que operam em tempo real é uma tarefa bastante complexa, pois envolve análises de problemas que não são encontrados em protótipos que não operam deste modo. Neste trabalho é apresentado um software para reconhecimento de fala contínua, independente do locutor, e que opera em tempo real, usando a técnica de Modelos Ocultos de Markov.*

Palavras Chave — reconhecimento de fala, processamento de voz, sistemas em tempo real.

INTRODUÇÃO

Poder comunicar-se com computadores e outras máquinas através da fala é um sonho antigo, que motivou vários pesquisadores ao longo de várias décadas. Graças a esse esforço de pesquisa, iniciado na década de 50, atualmente já podem ser encontrados os primeiros sistemas de reconhecimento de fala de aplicação comercial. Contudo ainda existe muito a ser aperfeiçoado até que seja criado um sistema que permita a comunicação entre máquinas e pessoas através da fala de uma forma natural.

Uma interface baseada na fala é mais amigável e, dependendo de como é implementada, faz com que não sejam necessários conhecimentos específicos, facilitando a comunicação dos sistemas informatizados com as pessoas, reduzindo assim a exclusão social causada pela falta de domínio das novas tecnologias.

Outras possíveis aplicações para sistemas de reconhecimento de fala são:

- controle de máquinas via voz, deixando as mãos do operador livres para executarem outras tarefas;
- controle de funções de um automóvel (por exemplo, faróis, vidros, limpadores de pára-brisas e rádio), o que ajudaria a evitar acidentes;
- telefone para surdos: uma pessoa que deseja telefonar para uma pessoa com deficiência auditiva poderia falar normalmente. Na casa do deficiente, um sistema de reconhecimento poderia, por exemplo, escrever em uma tela o que a outra pessoa falasse, ou controlar uma cabeça artificial, que mexeria a boca conforme o que estivesse sendo falado para que o deficiente fizesse uma leitura labial;
- auxílio a pessoas deficientes: pessoas paraplégicas ou tetraplégicas poderiam controlar cadeiras de rodas e computadores através de comandos de voz;

Este artigo descreve um sistema de reconhecimento de fala contínua baseado em Modelos Ocultos de Markov contínuos, independente de locutor, para vocabulários pequenos, e que opera em tempo real.

RECONHECIMENTO DE FALA

O processo de reconhecimento de fala consiste em mapear um sinal acústico, capturado por um transdutor (usualmente um microfone ou um telefone) em um conjunto de palavras.

Os sistemas de reconhecimento de fala podem ser caracterizados por vários parâmetros sendo que alguns dos mais importantes se encontram resumidos na Tabela 1 [3].

Tabela 1: Parâmetros típicos usados para caracterizar a capacidade de sistemas de reconhecimento de fala.

Parâmetros	Faixa
Modo de Pronúncia	palavras isoladas a fala contínua
Estilo de pronúncia	leitura a fala espontânea
Treinamento	dependente de locutor a independente de locutor
Vocabulário	pequeno (< 20 palavras) a grande (> 20000 palavras)
Modelo de linguagem	estados finitos a sensível a contexto
Perplexidade	pequena (< 10) a grande (> 100)
SNR	alta (> 30 dB) a baixa (< 10 dB)
Transdutor	microfone com cancelamento de ruído a telefone

Um sistema de reconhecimento de palavras isoladas requer que o locutor efetue uma pequena pausa entre as palavras, enquanto que um sistema de reconhecimento de fala contínua não apresenta este inconveniente.

A fala quando gerada de modo espontâneo é mais relaxada, contém mais coarticulações, e portanto é muito mais difícil de reconhecer do que quando gerada através de leitura.

[1] Carlos Alberto Ynoguti, INATEL – Instituto Nacional de Telecomunicações, Av. João de Camargo, 510, 37.540-000 Santa Rita do Sapucaí, MG, Brasil, yoguti@inatel.br
[2] José Domingos Adriano, INATEL – Instituto Nacional de Telecomunicações, Av. João de Camargo, 510, 37.540-000 Santa Rita do Sapucaí, MG, Brasil, jose-adriano@inatel.br

Os sistemas dependentes de locutor necessitam de uma fase de treinamento para cada usuário antes de serem utilizados, o que não acontece com sistemas independentes do locutor, desde que estes já foram previamente treinados com vários locutores.

O reconhecimento torna-se mais difícil à medida em que o vocabulário cresce, ou apresenta palavras parecidas.

Quando a fala é produzida em seqüências de palavras, são usados modelos de linguagem para restringir as possibilidades de seqüências de palavras. O modelo mais simples pode ser definido como uma máquina de estados finita, onde são explicitadas as palavras que podem seguir uma dada palavra. Os modelos de linguagem mais gerais, que aproximam-se da linguagem natural, são definidos em termos de gramáticas sensíveis a contexto.

Uma medida popular da dificuldade da tarefa, que combina o tamanho do vocabulário e o modelo de linguagem, é a *perplexidade*, grosseiramente definida como a média do número de palavras que pode seguir uma palavra depois que o modelo de linguagem foi aplicado.

Existem também parâmetros externos que podem afetar o desempenho de um sistema de reconhecimento de fala, incluindo as características do ruído ambiente e o tipo e posição do microfone.

O reconhecimento de fala é um problema difícil devido às várias fontes de variabilidade associadas ao sinal de voz [3]:

- *variabilidades fonéticas* : as realizações acústicas dos fonemas, a menor unidade sonora das quais as palavras são compostas, são altamente dependentes do contexto em que aparecem [1]. Por exemplo o fonema /*t*/ em *tatu* tem uma articulação puramente oclusiva, e em *tia*, dependendo do locutor, pode ter uma articulação africada, onde à oclusão se segue um ruído fricativo semelhante ao do início da palavra "chuva". Além disso, nas fronteiras entre palavras, as variações contextuais podem tornar-se bem mais acentuadas fazendo, por exemplo, com que a frase '*a justiça é ...*' seja pronunciada como '*ajusticé...*'
- *variabilidades acústicas*: podem resultar de mudanças no ambiente assim como da posição e características do transdutor.
- *variabilidades intra-locutor*: podem resultar de mudanças do estado físico/emocional dos locutores, velocidade de pronúncia ou qualidade de voz.
- *variabilidades entre-locutores*: originam-se das diferenças na condição sócio - cultural, dialeto, tamanho e forma do trato vocal para cada uma das pessoas.

Os sistemas de reconhecimento tentam modelar as fontes de variabilidade descritas acima de várias maneiras:

- Em termos fonético acústicos, a variabilidade dos locutores é tipicamente modelada usando técnicas estatísticas aplicadas a grandes quantidades de dados de treinamento. Também têm sido desenvolvidos algoritmos de adaptação ao locutor que adaptam modelos acústicos independentes do locutor para os do locutor corrente durante o uso [14][17].
- As variações acústicas são tratadas com o uso de adaptação dinâmica de parâmetros [14], uso de múltiplos microfones [15] e processamento de sinal [3].
- Na parametrização dos sinais, os pesquisadores desenvolveram representações que enfatizam características independentes do locutor, e desprezam características dependentes do locutor [4][5].
- Os efeitos do contexto lingüistico em termos fonético-acústicos são tipicamente resolvidos treinando modelos fonéticos separados para fonemas em diferentes contextos; isto é chamado de modelamento acústico dependente de contexto [8].
- O problema da diferença de pronúncias das palavras pode ser tratado permitindo pronúncias alternativas de palavras em representações conhecidas como redes de pronúncia. As pronúncias alternativas mais comuns de cada palavra, assim como os efeitos de dialeto e sotaque são tratados ao se permitir aos algoritmos de busca encontrarem caminhos alternativos de fonemas através destas redes. Modelos estatísticos de linguagem, baseados na estimativa de ocorrência de seqüências de palavras, são geralmente utilizados para guiar a busca através da seqüência de palavras mais provável [6].

Atualmente, os algoritmos mais populares na área de reconhecimento de fala baseiam-se em métodos estatísticos. Dentre estes, dois métodos têm se destacado: as redes neurais artificiais (*Artificial Neural Networks*, ANN) [16] e os modelos ocultos de Markov (*Hidden Markov Models*, HMM) [7]. Mais recentemente, implementações híbridas que tentam utilizar as características mais favoráveis de cada um destes métodos também têm obtido bons resultados [13].

SISTEMA IMPLEMENTADO

A rigor, um sistema que opera em tempo real deve ser capaz de efetuar todas as operações a uma taxa maior ou igual à taxa com que o sinal de entrada chega. Devido às verificações efetuadas pelo modelo de linguagem, não é possível atender a este critério. Entretanto, do ponto de vista do usuário, para que se tenha a impressão de operação em tempo real, basta que o tempo de resposta seja tal que o atraso devido ao processamento não seja perceptível (ou demasiadamente longo), e foi baseado nesta premissa que este sistema foi implementado.

Como tecnologia básica, foi escolhido o método baseado em modelos ocultos de Markov contínuos com modelamento por subunidades fonéticas. Para estas foram escolhidos os fones independentes de contexto[6]. Na Figura 1 tem-se um diagrama em blocos do sistema implementado.

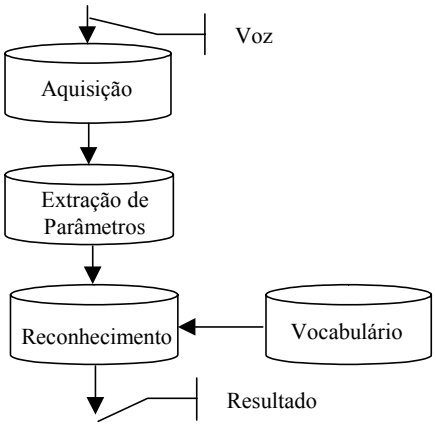

FIGURA 1
DIAGRAMA EM BLOCOS DE UM SISTEMA DE RECONHECIMENTO DE FALA.

Neste diagrama, podemos identificar os seguintes módulos:

1. **Módulo de aquisição de dados**: responsável por converter um sinal analógico, proveniente de um microfone, para um sinal digital em formato PCM.
2. **Módulo de extração de parâmetros**: tem por função converter o sinal de voz em uma seqüência de vetores de parâmetros. Isto é feito com o objetivo de representar os eventos acústicos relevantes do sinal de fala em termos de um conjunto compacto e eficiente de parâmetros.
3. **Módulo de reconhecimento**: por fim, a locução parametrizada é inserida no módulo de reconhecimento, que tem por função comparar a seqüência de parâmetros da locução a ser reconhecida com os modelos nele armazenados. A palavra cujo modelo que apresentar a maior similaridade com a locução de entrada vai ser a palavra reconhecida.

A seguir, cada um destes itens será mostrado em maiores detalhes:

Módulo de aquisição

Como dito anteriormente, este módulo é a interface do sistema com o usuário: coleta os sinais acústicos provenientes de um microfone e os apresenta para serem processados. Entretanto, como os algoritmos de reconhecimento apresentam um alto custo computacional, seria interessante que este só entrasse em ação quando houvesse um sinal de voz presente, deixando os recursos do sistema livres para outros aplicativos enquanto o locutor não estiver falando.

Analisando um sinal típico de a fala, nota-se que grande parte do tempo de uma conversação é ocupado por silêncio, como pode ser observado na Figura 2. Desta forma, para otimizar a utilização do sistema de reconhecimento, os intervalos de silêncio devem ser descartados, sendo processada somente a fala válida. Neste trabalho foi desenvolvido um algoritmo baseado em níveis de energia, uma adaptação do método proposto no artigo clássico de Rabiner & Sambur [11].

A identificação da presença de fala válida no sinal adquirido pode ser feita de várias formas, e a dificuldade em fazê-la depende do ambiente onde se encontra o locutor, pois a ausência de fala não significa silêncio absoluto. De fato, na maioria das aplicações práticas sempre haverá um ruído de fundo. Em ambientes demasiadamente ruidosos serão encontradas dificuldades para distinguir a fala do ruído não só na aquisição como também na etapa de reconhecimento.

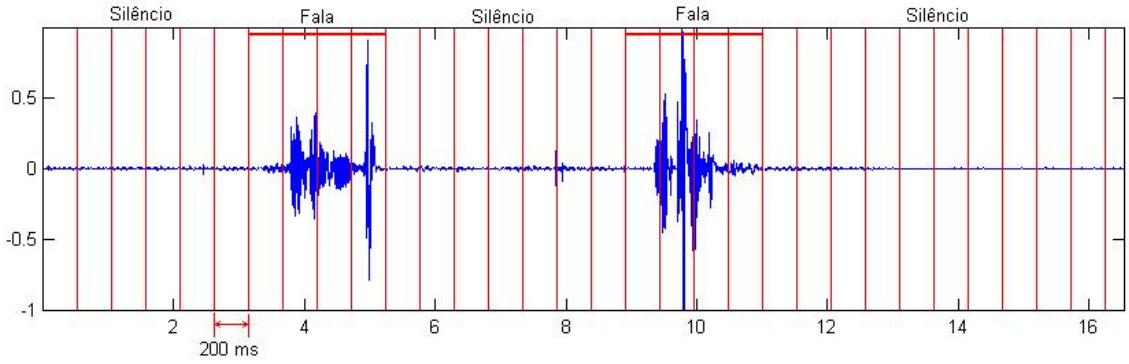

FIGURA 2
DIAGRAMA DO SISTEMA DE RECONHECIMENRO DE FALA

Para ambientes com uma relação sinal-ruído razoável uma técnica que apresenta bons resultados é a análise do nível de energia do sinal. Considerando que o sinal de voz do locutor possui um nível de energia notavelmente superior ao nível do sinal de energia do ruído, é possível identificar os intervalos que possuem fala.

Para isto, inicialmente é necessário medir o nível de sinal quando não se tem sinal de voz presente, isto é, o nível de energia do ruído ambiente. Depois, quando em operação, o sistema verifica se a entrada ultrapassa um determinado limiar acima do nível do ruído, e quando isto acontece, o sinal é enviado ao sistema de reconhecimento.

A aplicação do algoritmo de detecção de início e fim pode ser exemplificado pela

Figura 3, que apresenta a forma de onda de uma palavra, seus níveis de energia e as marcações de inicio e fim.

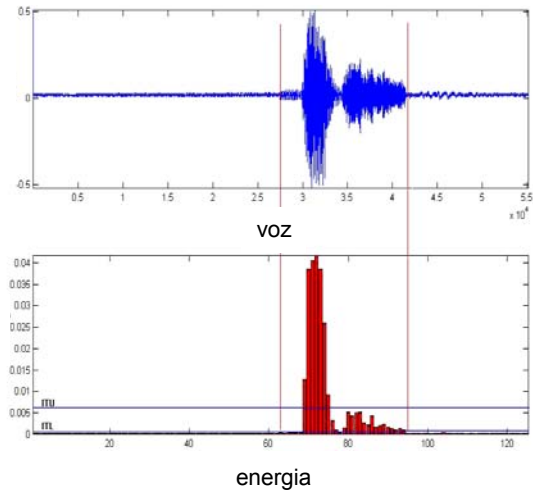

FIGURA 3
APLICAÇÃO DO ALGORITMO DE DETECÇÃO DE INICIO E FIM NA PALAVRA 'DOIS'.

Módulo de extração de parâmetros

A próxima etapa consiste na extração de parâmetros do sinal de voz. Para este trabalho foram escolhidos os parâmetros mel-cepstrais [4], amplamente utilizados na maioria dos sistemas de reconhecimento de fala atuais.

O processo de obtenção dos parâmetros acústicos da fala envolve 3 etapas: a pré-ênfase, o janelamento e a análise espectral, representadas conforme a Figura 4.

FIGURA 4
DIAGRAMA DE BLOCOS DO PROCESSO DE EXTRAÇÃO DOS PARÂMETROS MEL-CEPSTRAIS

A pré-ênfase, realizada por um filtro passa altas ($1 - 0{,}95z^{-1}$), tem como função compensar a atenuação de 6dB/oitava nas altas freqüências. Esta atenuação é ocasionada pelo efeito combinado do espectro decrescente dos pulsos glotais (-12dB/oitava) e pelo efeito de radiação dos lábios (+6dB/oitava) [10].

Os parâmetros do sinal de fala são atualizados a cada 10 ms, sendo o janelamento do sinal calculado através da janela de Hamming de 20 ms. Este janelamento tem como função produzir suavização da amplitude do sinal amostrado, nos extremos do segmento de análise, dando maior ênfase às amostras localizadas no centro da janela. Desta forma, tem-se uma superposição entre os dados de análise de 2 janelas adjacentes. Este processo pode ser observado na Figura 5.

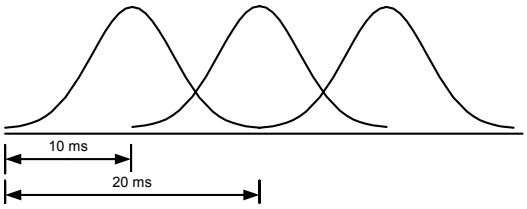

FIGURA 5
PROCESSO DE SUPERPOSIÇÃO DE JANELAS PARA O CÁLCULO DOS PARÂMETROS MEL-CEPSTRAIS

Como finalização do processo, têm-se a análise espectral, onde é realizada a conversão da representação temporal do sinal analisado, para alguma forma de representação espectral. Basicamente dois métodos de análise espectral predominam nos sistemas de reconhecimento de fala: o método de análise espectral LPC (Linear Predictive Coding) e o método de análise espectral por banco de filtros, obtido a partir da Transformada Rápida de Fourier (FFT). Para este trabalho, optou-se pelo segundo método, por ser mais eficiente na obtenção dos parâmetros mel-cepstrais.

Tanto o espectro resultante da FFT quanto o espectro resultante da predição linear são representações bem mais relacionadas ao processo de audição e percepção humana do que os métodos de representação temporal (taxa de cruzamentos por zero, perfil de energia, log-energia, entre outros), utilizados na caracterização da fala. Isto justifica a ampla utilização de parâmetros extraídos a partir da representação espectral do sinal acústico, em relação aos parâmetros de representação temporal.

Além dos parâmetros cepstrais (mel-cepstrais), foram utilizados também os parâmetros diferenciais (delta-mel-cepstrais, delta-delta-mel-cepstrais), com o intuito de uma melhor caracterização das variações temporais do sinal de fala.

Módulo de reconhecimento

O módulo de reconhecimento é o responsável pelo mapeamento dos parâmetros acústicos correspondentes à locução de entrada em sua transcrição ortográfica. Foram implementados três algoritmos de busca para o reconhecimento de fala contínua: o *Level Building* [11], o *One Step* [9] e o Herrman-Ney [2]. Para melhorar o desempenho do sistema em termos de taxa de acertos foram incluídos o modelo de duração de palavras [12] e o modelo de linguagem bigrama [6]. Também foi implementada a estratégia Viterbi *Beam Search* [12] de poda de caminhos. para diminuição do custo computacional e consequentemente do tempo de processamento. Um diagrama de blocos para este sistema é mostrado na Figura 6.

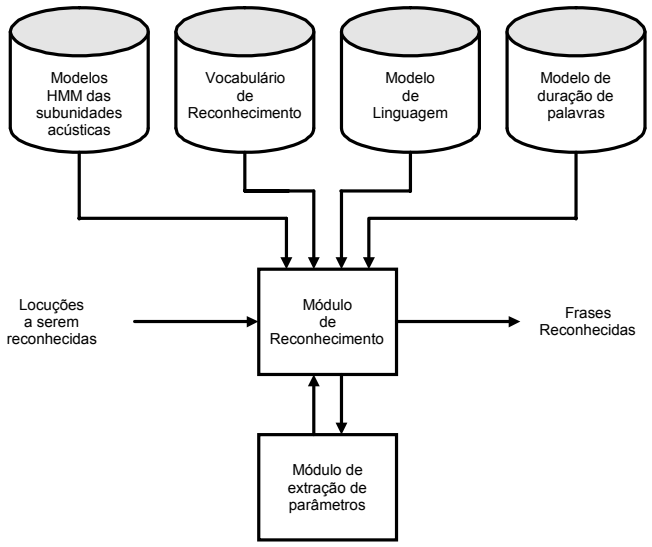

FIGURA 6
DIAGRAMA DE BLOCOS DO MÓDULO DE RECONHECIMENTO.

Estes algoritmos são extremamente complexos, e a descrição detalhada dos mesmos foge do escopo deste trabalho. Em [12] tem-se uma boa referência para o estudo dos mesmos.

Uma consideração importante a ser feita é sobre o tamanho do vocabulário, ou seja, o número de palavras que o sistema pode reconhecer. Quanto maior o vocabulário, maior será o tempo necessário para que uma palavra seja reconhecida. Com estes algoritmos, temos conseguido, para este sistema, operação em tempo real para vocabulários de algumas dezenas de palavras. Outros algoritmos de reconhecimento, mais eficientes estão sendo estudados para permitir operação em tempo real para vocabulários maiores.

CONCLUSÃO

Neste artigo foi apresentado um breve resumo da tecnologia de reconhecimento de fala e suas aplicações, bem como a descrição da implementação em tempo real de um sistema de reconhecimento de fala contínua independente de locutor baseado em modelos ocultos de Markov, para um vocabulário pequeno.

A implementação de um sistema para operação em tempo real apresenta dificuldades que não são encontradas em protótipos de laboratório que não operam em tempo real, levando a considerações e a soluções próprias deste tipo de sistema. Ainda, do ponto de vista da pesquisa, é possível testar novas idéias e soluções em um sistema real, verificando o ganho obtido e mesmo a viabilidade das idéias concebidas.

AGRADECIMENTOS

À FAPEMIG pelo financiamento parcial desta pesquisa. Processo TEC 85015/01.

REFERÊNCIAS

[1] Alcaim, A., Solewicz, J. A., Moraes, J. A., "Freqüência de ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro". *Revista da Sociedade Brasileira de Telecomunicações*. Vol 7, No 1, Dezembro, 1992, pp. 23-41.

[2] Aubert, X. Ney, H., "Large vocabulary continuous speeech recognition using word graphs", *Proceedings of ICASSP*, Detroit, MI, May, 1995.

[3] Cole, R. A., ed., "Survey of the State of the Art in Human Language Technology". http://cslu.cse.ogi.edu/publications/index.htm, (26/10/98).

[4] Davis, S. & Melmertstein, P. "Comparison of parametric representations for monossylabic word recognition in continously spoken sentences", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol ASP-28, No 4, August, 1980, pp. 357-366.

[5] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", *Journal of the Acoustical Society of America*, Vol 87 No 4, 1990, pp. 1738-1752.

[6] Jelinek, F., "Statistical Methods for speech recognition", *MIT Press*, Cambridge, Massachussets, 2001.

[7] Lee, K. F., Hon, H. W., Reddy, R., "An overview of the SPHINX speech recognition system". *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol 38, No 1, April, 1990, pp. 35-45.

[8] Lee, K. F., "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol 38, No 4, April, 1990, pp. 599-609.

[9] Ney, H., "The use of a one-stage dynamic programming algorithm for connected word recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol ASSP-32, No. 2, April 1984.

[10] Picone, J. W., "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, Vol 81, No 9, September 1993, pp. 1215-1247.

[11] Rabiner, L. R. and Sambur, M. R., "An Algorithm for Determining the Endpoints of Isolated Utterances", *The Bell System Technical Journal*, February, 1975, pp. 297-315.

[12] Rabiner, L., "Fundamentals of speech recognition", *Prentice Hall Press*, 1993.

[13] Schalwyk, J. et al., "Embedded implementation of a hybrid neural-network telephone speech recognition system", *IEEE International Conference on Neural Networks and Signal Processing*, Nanjing, China, December 10-13, 1995, pp. 800-803.

[14] Siegler, M. A. and Stern, R., "On the effects of speech rate in large vocabulary speech recognition systems", *Proceedings of the ICASSP* , Detroit, MI, May 1995, pp. 612-615.

[15] Sullivan, T. M. & Stern, R., "Multi-microphone correlation-based processing of robust speech recognition", *Proceedings of the ICASSP*, Minneapolis, Minnesota, 1993.

[16] Tebelskis, J., "Speech recognition using neural networks", *Ph.D. Thesis*, School of Computer Science, Carnegie Mellon University, 1995.

[17] Zhan, P. et al., "Speaker normalization and sepaker adaptation – a combination for conversational speech recognition", *Proceedings of EUROSPEECH*, 1997.