# Um Estudo Empírico da Contribuição Melódica e Rítmica da Fala para a Percepção da Identidade do Orador

José Roberto Sousa Leal Junior, Elker Anderson Nunes Carvalho, Jugurta Montalvão

***Abstract*— O resultado de experimentos com voluntários sobre a percepção do ritmo da fala e do contorno da entonação em uma sentença curta (3s de duração) é apresentado e confrontado a resultados obtidos através de sistemas automáticos de reconhecimento de orador, relatados na literatura. Os resultados encontrados com voluntários humanos são apresentados com a finalidade de servir como referência de desempenho para sistemas automáticos, em duas modalidades: (a) usando o contorno da entonação ao longo do tempo, e (b) usando apenas a informação de ritmo (a entonação e o conteúdo espectral são mascarados).**

***Index Terms*— Biometria comportamental, reconhecimento de orador, prosódia.**

## I. Introdução

A identidade de um orador pode ser percebida por um observador humano em várias níveis cognitivos. Sistemas automáticos de reconhecimento de orador (*Automatic Speaker Verification* - ASV), baseados em computador, tiram proveito disso através de métodos baseados tanto em conteúdos espectrais de curta duração (dezenas de milisegundos), quanto em prosódia de registros com alguns segundos de sinais de voz [1] .

Prosódia é comumente usado para referenciar aquelas propriedades da voz que não podem ser obtidos da seqüência de fonemas que compõem uma sentença, tais como o ”pitch”, a duração relativa das sílabas e a flutuação intencional da intensidade da voz. Em suma, a prosódia da fala pode ser vista como uma combinação intencional de aspectos melódicos e rítmicos da voz [2].

Sabe-se que inúmeras características da voz podem ser utilizadas na verificação de locutor, tais como as formantes da voz[1] que, quando corretamente extraídas, refletem o formato instantâneo do trato vocal, através de suas ressonâncias. Claramente, o formato do trato vocal é um atributo morfológico do indivíduo, o que justifica o uso da extração de formantes como principal estratégia para identificação de pessoas através da voz.

Porém, a aquisição de sinais de voz em ambientes não controlados torna os sistemas automáticos vulneráveis a ruídos e distorções. Em contrapartida, características como o ritmo da fala e o contorno da sua freqüência fundamental (F0) – equivalente acústico do *pitch* – são praticamente invariantes a distorções provocadas por canais de transmissão lineares (distorções convolutivas) e/ou por ruídos aditivos (ruído ambiente). Isto é, mesmo quando o sinal é distorcido/degradado ao ponto das frases não poderem mais ser reconhecidas, ainda é possível algum nível de reconhecimento do orador através do ritmo (balanço entre durações de sons vocálicos, fricativos e pausas) e do contorno da F0 ao longo de frases, que caracteriza a ’melodia’ da voz de cada indivíduo.

Universidade Federal de Sergipe (UFS), Núcleo de Engenharia Elétrica (NEL) E-mail: jmontalvao@ufs.br

Há uma grade quantidade de trabalhos publicados recentemente onde a prosódia é usada ora como fonte de informação complementar, ora como única fonte nas tarefas de identificação biométrica (ver [3], [2], [4] e referências apontadas por esses autores). Através dos resultados obtidos e publicados na literatura, fica evidente que a informação vinda da prosódia é relevante para sistemas baseados em computador. Por outro lado, também aceitamos a hipótese de que a forma como humanos realizam essa tarefa é muito mais complexa e robusta (e.g. poucos segundos de sinal através de um canal telefônico deteriorado são geralmente suficientes para que um locutor seja identificado com sucesso, mesmo com baixa relação sinal/ruído e distorções severas de canal).

Assim, assumindo, por hipótese, que a cognição humana é uma referência natural para o desempenho de sistemas de reconhecimento automático de orador, neste trabalho, foi montado e executado um experimento com voluntários humanos para estabelecer uma possível referência de desempenho na verificação de oradores. Além disso, como este trabalho focaliza apenas a identificação através da prosódia, os experimentos foram realizados em dois tempos, de forma a enfatizar apenas o ritmo da fala, ou o contorno da F0 (contorno melódico, que inclui parte do ritmo) de um sinal de voz.

Na seção II, a base de registros de vozes usadas no experimento é descrita. Nas seções III e IV, são detalhados os processamentos usados para extração do ritmo e da melodia de cada frase registrada, o que permite, na seção V, a descrição dos experimentos com base nos sinais processados. Na seção VI, são apresentados os resultados consolidados dos experimentos, juntamente com alguns resultados coletados na literatura, com sistemas automáticos baseados em computador. Finalmente, na seção VII, são apresentadas as conclusões parciais deste trabalho.

## II. Base de dados usados nos experimentos

Neste trabalho, foram utilizadas amostras de 22 locutores, sendo metade de cada sexo, com idades distribuídas entre 20 e 50 anos. Cada locutor voluntariamente registrou 10 repetições da frase: ”Chocolate, Zebra, Banana, Táxi”, com duração aproximada de 3 segundos. Cada voluntário foi solicitado em

duas sessões de gravação: na primeira, apenas 5 registros foram tomados, enquanto que os demais registros foram obtidos pelo menos um mês após a primeira sessão.

Cada registro foi codificado em um arquivo individual de computador, com amostras tomadas a 22050 Hz, numa resolução de 16 bits por amostra (escala linear).

## III. Extração do Ritmo

O ritmo[1] é um parâmetro lingüístico cognitivo que pode ser definido pela seqüência de durações de intervalos de fala quase-estacionários – vozeados ou fricativos –, pela marcação de sons explosivos, e pelos intervalos de supressão intencional da fala. Considera-se um intervalo de fala quase-estacionário aquele intervalo em que o ouvinte pode identificar um som emitido que pode ser categorizado como uma vogal, ou um som fricativo, como aqueles representados por 'sh', 'f', 'ss', etc.

Visando isolar permitir que um ouvinte voluntário (no experimento) pudesse perceber apenas o ritmo das vozes gravadas, foi desenvolvido um método de processamento dos sinais que preserva o ritmo da fala, ao mesmo tempo que que degrada completamente todas as demais informações de prosódia. Este novo método consiste simplesmente na utilização de quatro filtros, digitais, do tipo passa-faixa, cujos parâmetros – freqüência média, freqüência de corte inferior e superior – estão explicitados na tabela I. A associação desses filtros é feita em paralelo, como ilustrado na figura 1.

TABLE I
VALORES DOS PARÂMETROS DOS FILTROS FFT PASSA-FAIXAS.

| Filtro | Corte inferior | Corte superior | Centro |
|---|---|---|---|
| 1 | 983.44 Hz | 1086.8 Hz | 1035.12 Hz |
| 2 | 1476.74 Hz | 1634.72 Hz | 1555.73 Hz |
| 3 | 2488.2 Hz | 2749.72 Hz | 2618.96 Hz |
| 4 | 3957.81 Hz | 4373.79 Hz | 4165.8 Hz |

Como resultado da passagem do sinal de voz pelo banco de filtros, é obtido um som simbilante, semelhante ao som de sopros em uma garrafa. Assim, as vogais (as vogais distorcidas são percebidas como sibilos estacionários distintos), sons fricativos e pausas ainda são claramente percebidas (ritmo), caracterizando claramente as transições e durações do conteúdo espectral, enquanto outras características vocais são indistinguíveis. De fato, é fácil intuir que, dispondo de apenas 'quatro finas fatias' do espectro do sinal, o observados não possui informação suficiente para estimar nem o *pitch*, nem as formantes da voz.

## IV. Extração do contorno Melódico

O contorno da freqüência fundamental (F0) – equivalente acústico do *pitch* – de intervalos de fala vozeada pode ser entendido como um sinal 'melódico', ou um 'canto'.O pitch, é uma característica vocal que muitas vezes é provisoriamente

[1]Segundo Platão [5], o ritmo vem da mente humana, não do corpo. E Robert Jourdain complementa: "(...) o ritmo tem a ver com agrupamento, com reunião do conteúdo do mundo em conjuntos discerníveis. É inerente a todos os tipos de cognição (...)".

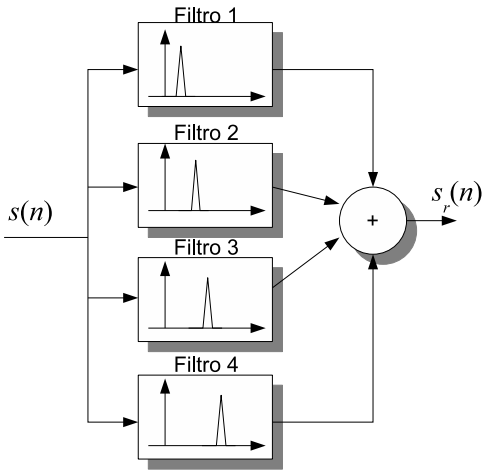


Fig. 1. Banco de filtros usado para extrair ritmo.

definida como a freqüência fundamental da oscilação glotal (vibração das dobras vocais). No entanto, essa definição não é consistente, visto que o pitch é uma característica psicoacústica (subjetiva) [6], [7]. Assim, o pitch é mais bem definido como um atributo auditivo do som, através do qual os sons podem ser ordenados numa escala de baixo à alto. Para extrair o contorno melódico de cada frase, foi estimado o seu F0 máximo, usando o algoritmo robusto descrito em [7], e então ajustada a freqüência de corte de um filtro digital passa-baixas, de modo que esse filtro cortasse as freqüências superiores. O som resultante soa então como um 'murmúrio' no qual não são percebidas sílabas. Apenas o contorno melódico da frase é preservado. A figura 2 ilustra o filtro passa-baixas adaptado a cada nova frase.

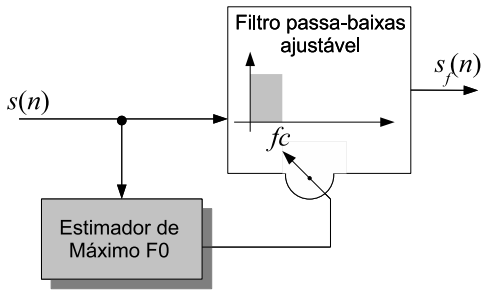


Fig. 2. Filtro usado para extrair contorno melódico.

## V. Descrição do Experimento

A partir das amostras de voz previamente gravadas, em duas sessões de coleta por voluntário, foram criados três grupos: Grupo de Contrelo (GC), Grupo pró-F0 (GF) e Grupo pró-ritmo (GR). Para criação do grupo GC, foram selecionadas amostras de locutores diferentes, de forma aleatória. Já para criação do segundo grupo, GF, foram selecionadas aleatoriamente amostras em igual número, sem intersecção com GC, que foram processadas pelo filtro de extração do contorno melódico. Por último, o grupo GR foi formado pelo terço

de amostras restantes, processadas pelo banco de filtro para extração do ritmo.

Em seguida, cinqüenta ouvintes voluntários foram convidados a realizar testes de verificação de locutor, segundo dois procedimentos distintos:

#1 No primeiro teste, cada voluntário ouviu dois registros: um correspondente à uma voz 'limpa', sorteada de GC, e outra distorcida, sorteada de GR. O algoritmo de sorteio das amostras foi feito de forma que a probabilidade dos sons limpo e distorcido serem do mesmo orador era 50%. Sem ter esta informação *a priori*, o voluntário era solicitado a decidir se a amostra do GR vinhas do mesmo orador que a amostra do GC, ou seja, se o ritmo e a voz apresentados pertenciam à mesma pessoa (decisão baseada no ritmo da frase). Para cada voluntário esse procedimento foi repetido cinco vezes.

#2 No segundo teste, o mesmo procedimento do primeiro era realizado, com a diferença de que o GF era utilizado no lugar de GR. Ou seja, o voluntário decidia se a amostra do GF era correspondente a amostra do GC (decisão baseada na melodia da frase, que inclui parcialmente o ritmo).

## VI. Resultados e Discussões

Os resultados consolidados obtidos nos dois testes podem ser vistos na tabela II, em termos de taxa de decisões erradas.

TABLE II
RESULTADOS OBTIDOS NO 1º E 2º TESTE.

| Informação utilizada | Decisões erradas(%) | Duração(s) |
|---|---|---|
| sinal puramente rítmico | 33.2 | 3 |
| sinal com contorno melódico | 21.5 | 3 |

Vale notar que, considerando-se as probabilidades *a priori* do experimento de 50% sinais limpos e distorcidos vindos da mesma fonte (mesmo orador), o resultado esperado, para o caso das informações contidas no ritmo e na melodia da voz serem desprezíveis para a verificação do orador, seria também de 50%.

Assim, os resultados obtidos são evidências de que essas informações são relevantes, sendo que a melodia da frase parece ter um poder discriminatório maior que o ritmo da fala.

Esses experimentos foram realizados com frases curtas, com 3 segundos apenas, e com um texto fixo, escolhido arbitrariamente. Em conseqüência, uma comparação desse resultado com outros semelhantes, publicados na literatura (onde sistemas automáticos assumem o lugar do homem no momento da decisão) deve ser feito com prudência. Primeiramente porque a duração do sinal fornecido ao sistema (ou ao humano) deve afetar a taxa de reconhecimentos corretos, assim como o uso de frases fixas, em princípio, deve favorecer a verificação correta.

No entanto, apesar desses riscos de comparações inapropriadas, listamos, na tabela III alguns resultados de outros trabalhos publicados, em termos de *Equal Error Ratio (EER)* , que corresponde ao ponto de trabalho do sistema em que as taxas de falsa aceitação igualam as taxas de falsa rejeição.

Esses desempenhos são tabelados aqui apenas para fornecer uma visão mais ampla de desempenho de sistemas de verificação de orador baseados em prosódia. No entanto, num sentido largo, podemos comparar o EER com as taxas de erros de verificação medidas em nossos experimentos.

Além disso, a coluna rotulada com a palavra 'duração' representa o tempo de gravação de cada amostra entregue ao sistema (no nosso caso, tanto a amostra de referência quanto a de teste duram aproximadamente 3 segundos). Em todos os casos tabelados, o treinamento do modelo é feito com amostras de referência bem mais longas que as de teste).

TABLE III
RESULTADOS OBTIDOS NO 1º E 2º TESTE.

| Fonte | Contorno de | EER(%) | Duração (s) |
|---|---|---|---|
| Farahani et al. [3] | $f_0$ | 4.8 | 180 |
| Farahani et al. | $f_0$ | 8.0 | 20 |
| Farahani et al. | $f_0$ | 13.0 | 10 |
| Farahani et al. | $f_0$ | 18.0 | 5 |
| Farahani et al. | $f_0$ | 28.0 | 2 |
| A. G. Adami [2] | $f_0$ | 20.5 | 2 |
| A. G. Adami | energia | 23.5 | 2 |
| A. G. Adami | energia e $f_0$ | 14.2 | 2 |
| A. G. Adami | energia e $f_0$ + duração | 11.4 | 2 |
| R. D. Zilca [4] | GMM | 13.3 | 15-45 |
| R. D. Zilca | SMR | 23.8 | 15-45 |
| R. D. Zilca | SG | 26.4 | 15-45 |
| R. D. Zilca | DSR | 27.6 | 15-45 |

Na tabela, a siglas SMR, SG e DSR são abreviações de, respectivamente, *Sphericity Measure Ratio*, *Single Gaussian full covariance Gaussian Mixture Model* e *Divergence Shape Ratio*. Vale notar que essas medidas não são ligadas a contornos estimados explicitamente, mas a parâmetros espectrais médios de uma sentença. Isto é, são usados esquemas de *utterance level scoring*, em contraposição aos usuais *frame level scoring*. Assim, não são estimados explicitamente o contorno de F0 ou de energia. São usados os *Mel frequency cepstral coefficients* (MFCC) [8], estimados em uma sentença longa.

Ainda com relação à tabela, no trabalho de A. G. Adami [2], o termo 'duração' representa o uso de estimação seqüencial de intervalos de tempo, dentro dos quais o sinal é considerado uniforme, em termos de prosódia.

## VII. Conclusões

O uso da prosódia no reconhecimento automático de orador continua sendo um foco relevante de pesquisa, como pode ser evidenciado pela grande quantidade de artigos sobre o tema, publicados em eventos e periódicos importantes. Alguns desses trabalhos foram citados aqui e, em particular, em [2] é apresentada uma extensa revisão bibliográfica sobre o tema.

Todos esses trabalhos compartilham uma motivação principal: a de que a prosódia carrega uma grande quantidade de informação complementar aos parâmetros acústicos usuais. Além disso, este trabalho, em particular, encontra sua motivação na hipótese de que os contornos de energia e

$F_0$, que naturalmente também exibem informações rítmicas discriminantes, são menos sensíveis a perturbações de canais, tais como os ruídos aditivos e distorções por multipercurso (distorções convolucionais).

Por outro lado, um observador humano é capaz de se adaptar a canais severos. Por ilustração, não é incomum testemunharmos situações em que, mesmo numa comunicação telefônica muito degradada, situação em que a maioria dos sistemas automáticos falhariam, um humano é capaz de identificar a identidade de um orador familiar, mesmo quando a própria mensagem não pode ser compreendida.

Assim, como parte preliminar de um projeto recém-iniciado, com foco nas identificação biométrica robusta baseada em ritmo, este trabalho apresenta o resultado de um teste de reconhecimento de identidade com voluntários humanos.

Esses resultados serão usados por nós como referência de desempenho para sistemas automáticos. Isto é, ao invés de compararmos desempenhos entre sistemas implementados, pretendemos comparar cada sistema à referência humana, que, por hipótese, deve superar a maioria dos sistemas automáticos, em situações adversas, com canais degradados.

Embora a base de sinais usada não seja grande, acreditamos que essa estratégia de usar uma referência baseada em decisões humanas pode ser igualmente útil a outros grupos de pesquisa. Por esta razão, além dos resultados consolidados aqui apresentados, também disponibilizamos livremente a base de amostras utilizadas, agora disponível para *download* em www.ufs.br/biochaves.

## ACKNOWLEDGMENTS

Este trabalho tem sido financiado pelo *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq). Nós também agradecemos muito os estudantes, colegas e demais voluntários que participaram das sessões de coleta de amostras de vozes usadas neste trabalho de pesquisa, e dos voluntários que, pacientemente, participaram dos experimentos de verificação de identidades pelos sinais de voz.

## REFERENCES

[1] R.W. Schafer L.R. Rabiner, *Digital Processing of Speech Signals*, Prentice Hall Sig. Proc. Series, New Jersey, 1978.

[2] André Gustavo Adami, "Modeling prosodic differences for speaker recognition," *Speech Communication (Elsevier)*, , no. 49, pp. 277—-291, 2007.

[3] F. Farahani, P. G. Georgiu, and S. S. Narayanan, "Speaker identificattion using supra-segmental pitch pattern dynamics," in *ICASSP 2004*, 2004, pp. I–89 – I92.

[4] Ran D. Zilca, "Text-independent speaker verification using utterance level scoring and covariance modeling," *IEEE TRANS. ON SPEECH AND AUDIO PROCESSING*, vol. 10, no. 6, pp. 363—-370, SEPTEMBER 2002.

[5] R. Jourdain, *Music, The Brain, and Ecstasy*, William Morrow Press., 1997.

[6] M. Kemal Sömez, Larry Heck, Mitchel Weintraub, and Elizabeth Shriberg, "A lognormal tied mixture model of pitch for prosody based speaker recognition," in *EUROSPEECH, 1997*, 1997.

[7] Elker Anderson Nunes Carvalho, Luciana Maria Fontes Maciel, José Roberto Sousa Leal Jr., and Jugurta Montalvão, "Simplified automatic detection of parkinson's disease evidences through pitch dynamics analysis," in *BIOMAT - International Symposium on Mathematical and Computational Biology*, Nov. 2007, pp. 13–20.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.