Synthesis of Sung Spanish Vowels in Lyrical Singing by Sopranos

L. Barrientos, and E. Cataldo

***Abstract*—The aim of this paper is perform the synthesis of sung Spanish vowels considering the soprano vocal category of lyrical singers, including variation of sustained pitches with vibrato and tremolo effects, considering sounds from Spanish language. The Fant source-filter theory is used to model the production of the sung vowels: the source is based on the Rosenberg glottal pulse model and the filter (the vocal tract) is composed by an all-pole filter model with formant frequencies and bandwidths from the vowels of the Spanish language, obtained through experimental voice signals from two soprano singers. All the sounds synthesized are available to be accessed and they were submitted to a group of listeners which gave a very good evaluation with respect to the intelligibility and naturalness of the sounds.**

***Index Terms*—Glottal signal, Rosenberg model, singing voice, vibrato effect, voice synthesis.**

I. Introdução

A síntese da voz cantada é um dos tópicos mais desafiadores estudados por pesquisadores em síntese da fala. Desde os primeiros trabalhos publicados nos anos 60 [1], a síntese da voz cantada atraiu a atenção de muitos pesquisadores que desenvolveram inúmeros sintetizadores [2], [3]. Esses sintetizadores foram aprimorados continuamente ao longo dos anos, em termos de inteligibilidade e naturalidade. Na literatura, encontram-se realizações notáveis que foram alcançadas na área de síntese da voz cantada como, por exemplo, o instrumento musical simulado computacionalmente que permite a síntese em tempo real de vozes cantadas desenvolvido por D'Alessandro e Woodruff [4], ou ainda o trabalho recente sobre um sintetizador de voz para canto baseado em Redes Neurais Artificiais apresentado por Chandna e Blaauw [5].

Hoje em dia, os sintetizadores de voz para canto têm sido utilizados não apenas como passatempo, mas também para produção musical profissional. Podem ser citados robôs cantores e, também, aulas de técnica vocal que se tornaram muito úteis na indústria do entretenimento [6]. Há ainda alguns problemas a serem superados, como a síntese da voz feminina cantada que precisa de um pouco mais de esforço para atingir um grau maior de naturalidade [7]. De acordo com Sundberg [8], uma cantora soprano geralmente canta notas musicais cuja frequência fundamental (F_0) é mais alta que a frequência do primeiro formante (F_1), nas vogais sustentadas. Ou seja, no registro agudo do canto feminino, os valores de F_0 são superiores ao valor de F_1 da fala, principalmente nas vogais $/e/$, $/i/$, $/o/$ e $/u/$. Nesse caso, F_1 não aumenta a amplitude de F_0 e, portanto, o som emitido é mais *fraco*. Para contornar esse problema, em geral, as cantoras sopranos aumentam a abertura da boca, quase involuntariamente, e, com isso, aumentam a frequência F_1, fazendo-a coincidir com os valores de F_0, permitindo que a frequência do primeiro formante reforce a amplitude de F_0. Esse aumento da frequência do primeiro formante geralmente torna os valores de F_1 mais próximos da frequência do primeiro formante da vogal $/a/$ e, perceptivelmente, essa vogal pode ser identificada. Porém, o efeito obtido é adverso da inteligibilidade das vogais e isso causa um problema para a síntese [9]. Esse comportamento é conhecido como ajuste de formantes e foi observado por Johan Sundberg através de uma técnica experimental para estimar as frequências formantes no canto feminino [8], [10].

Pesquisas relacionadas à síntese da voz cantada e à inteligibilidade vocálica foram realizadas por pesquisadores, em diversas línguas: inglesa americana, francesa, alemã e japonesa [2], [3], [6], [14]. No entanto, não foram encontrados estudos na língua espanhola americana.

O estudo presente nesse artigo combina conteúdos de análise e processamento de sinais, síntese de voz e psicofísica para tratar de detalhes sobre a inteligibilidade das vogais da língua espanhola (americana) que são afetadas na extensão vocal do soprano lírico. Para desenvolver este estudo, sintetizamos as vogais cantadas usando um sintetizador de formantes baseado no modelo fonte-filtro de Fant [15], [16].

A fonte é composta por pulsos glóticos baseados no modelo de Rosenberg e com algumas características acústicas dos cantores líricos, como vibrato (efeito onde a frequência da nota é levemente alterada com base em um tempo determinado pelo efeito) e tremolo (alteração no volume na nota reproduzida, sem alterar a frequência). A fonte glótica desenvolvida neste trabalho não modela fenômenos como *jitter* e *shimmer* [12], [13], e apenas uma forma de pulso glotal, considerando o comportamento glótico na escala de notas mais agudas das sopranos, como investigada por Garnier e Henrich [11].

A partir de vozes gravadas de duas cantoras sopranos profissionais, foram obtidas as frequências dos formantes e suas respectivas larguras de faixa, usadas no modelo de filtro.

Os dados de áudio digital obtidos das gravações das vozes das cantoras sopranos foram registradas em disco de estado sólido (SSD). O formato dos áudios é WAVE e os parâmetros de qualidade são: taxa de amostragem de 44100 Hz a 16 bits por amostra, um canal (Mono) e sem compressão de áudio.

Um grupo de ouvintes avaliou a síntese das vogais cantadas e os resultados dessa avaliação foram analisados e, posteriormente, discutidos. Foi realizada uma análise de naturalidade e inteligibilidade na síntese das vogais, bem como do possível

L. E. Barrientos is with the Graduate Program in Electrical and Telecommunications Engineering, Universidade Federal Fluminense, Niterói, RJ, Brazil (e-mail:eduardo_sandoval@id.uff.br).

E. Cataldo is with Electrical and Telecommunications Engineering Department, Universidade Federal Fluminense, Niterói, RJ, Brazil (e-mail:ecataldo@id.uff.br).

formante do cantor em sopranos, como apresentado por Johan Sundberg em 1988 [8], [10]. Todos os sons obtidos estão disponíveis e podem ser acessados facilmente através de um link que será indicado mais adiante nesse texto.

II. A Teoria Fonte-Filtro

O modelo completo apresentado aqui é baseado na teoria fonte-filtro Fant [15], ilustrada na Fig. 1.

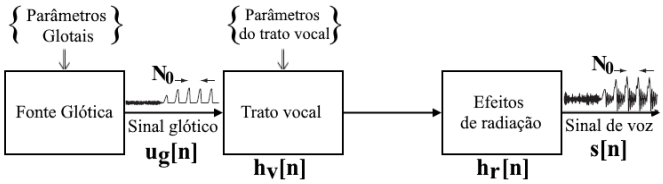


Fig. 1. Modelo fonte-filtro de Fant.

O sinal de voz gerado, $s[n]$, é dado pela convolução do sinal glótico $u_g[n]$, a função de resposta ao impulso correspondente $h_v[n]$ do filtro que modela o trato vocal, e a radiação pela boca para a qual a função de resposta ao impulso é $h_r[n]$. A equação (1) é então obtida:

$$s[n] = h_r[n] * h_v[n] * u_g[n]\,, \tag{1}$$

ou, no domínio da frequência,

$$S(k) = H_R(k)\, H_V(k)\, U_G(k)\,, \tag{2}$$

onde $S(k)$ significa a transformada de Fourier discreta (TFD) do sinal de voz gerado. Com essa formulação, não há acoplamento entre a fonte glotal e o trato vocal, simplificando o modelo. Neste artigo, $u_g[n]$ é construído usando o modelo de pulso glótico de Rosenberg com uma modulação específica para obter efeitos no som produzido como vibrato; $h_v[n]$ é tomada da literatura e modificada de acordo com as primeiras cinco frequências formantes obtidas da voz cantada de duas cantoras líricas profissionais Colombianas, e $h_r[n]$ é um filtro FIR (resposta de impulso finita) passa-alta de primeira ordem, como sugerido em [18].

A. O Sinal Glótico

O sinal glótico foi gerado a partir do modelo do pulso glótico de Rosenberg [19], dado pela Eq. 3:

$$g[n] = \begin{cases} 0.5A_v[1 - cos(\pi n/(N_1))] & ,\ 0 \le n \le N_1 \\ A_v cos(\pi(n - N_1)/(2N_2)) & ,\ N_1 < n \le N_1 + N_2 \\ 0 & ,\ N_1 + N_2 < n \le N_0\,. \end{cases} \tag{3}$$

Onde N_0 representa o período fundamental do sinal glótico; A_v é a constante relacionada à amplitude do pulso glótico; N_1 é o tempo de abertura, porção do pulso com inclinação positiva e N_2 é o tempo de fechamento, porção do pulso com inclinação negativa. Os instantes relativos de abertura e fechamento são dados por $\alpha_1 = N_1/N_0$ e $\alpha_2 = N_2/N_0$ respectivamente.

O quociente de abertura glótica OQ aumenta suavemente com F_0 durante a emissão de fonemas sustentados nas notas mais agudas das sopranos líricas. Os valores máximos de OQ, superiores a $0,8$ e, principalmente, em torno de $0,9$, são atingidos no limite superior da transição laríngea [11].

O OQ é definido como a razão entre a duração da fase de abertura glótica e o período fundamental. A partir da relação $OQ = (N_1 + N_2)/N_0$, os valores para α_1 e α_2 podem ser calculados [14], [20].

Neste artigo, considerou-se $OQ = 0,78$, seguindo o comportamento glótico na faixa de altas frequências das sopranos, como investigado por Garnier e Henrich [11]. Com essa consideração, os valores dos instantes relativos de abertura e fechamento glótico foram calculados com $\alpha_1 = 58\,\%$ e $\alpha_2 = 20\,\%$. A frequência fundamental escolhida F_0 equivale a uma nota da extensão vocal das cantoras sopranos. Com as informações de um pulso glótico completo, um sinal glótico foi gerado através da convolução do pulso com trem de impulsos dado pela Eq. 4.

$$u_g[n] = \sum_{k=0}^{N-1} x[n] \cdot g[n-k]\,. \tag{4}$$

Considerando que: $u_g[n]$ é o sinal glótico; $x[n]$ representa o trem de impulsos unitários e $g[n]$ é o pulso glótico do modelo de Rosenberg.

Na Fig. 2 apresenta-se o sinal glótico, $u_g[n]$, formado por uma sequência de 7 pulsos produzidos com o modelo do pulso glótico de Rosenberg.

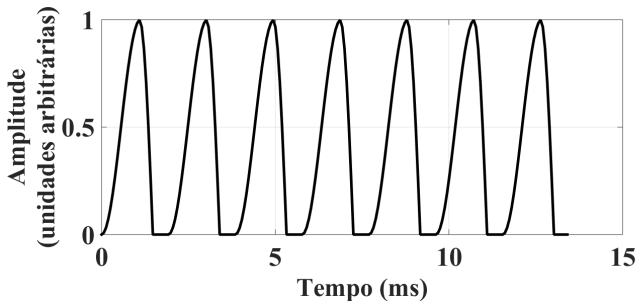


Fig. 2. Sinal glótico gerado através do modelo de Rosenberg e considerando os parâmetros: $F_0 = 520$ Hz, $\alpha_1 = 58\,\%$ e $\alpha_2 = 20\,\%$.

B. Efeito Vibrato

O vibrato é uma flutuação periódica na frequência de uma nota comumente utilizada na música. Segundo [21], o vibrato é caracterizado por quatro parâmetros mensuráveis: taxa, extensão, regularidade e forma de onda. A taxa de vibrato (V_{Rate}) determina o número de flutuações por segundo, normalmente 4 Hz - 14 Hz. A extensão do vibrato (V_{Ext}) está associada com a quantidade de variação do tom, com valores médios entre 5 ms a 10 ms. A regularidade tende a variar mais durante a fase negativa e a forma de onda é aproximadamente senoidal. Para gerar o efeito de vibrato na voz cantada deste trabalho, os valores dos parâmetros utilizados foram: $V_{Ext} = 0,4$ ms e $V_{Rate} = 4,5$ Hz, seguindo as características acústicas correspondentes ao vibrato para a identificação de cantores, como discutido por Nwe e Li [22].

Instrumentos que produzem vibrato, como a voz, e os instrumentos de corda e sopro, caracterizam-se por uma variação

periódica na intensidade do tom que é conhecida como tremolo [24]. O tremolo ocorre conforme o vibrato é produzido e pode ser descrito por dois parâmetros: nível de amplificação (A_{Level}) e taxa de tremolo (T_{Rate}). A taxa do tremolo segue a mesma faixa da taxa do vibrato: $T_{Rate} = 4,5$ Hz, assim como o nível de amplificação segue a faixa da extensão do vibrato: $A_{Level} = 0,4$ ms.

C. Modelo do Trato Vocal

Os efeitos do trato vocal foram desenvolvidos usando um modelo de filtro *all-pole filter*, conforme descrito em [18], considerando as ressonâncias (formantes) correspondentes aos polos da função de transferência do sistema discreto do trato vocal, $V_k(z)$, mostrada na Eq. 5:

$$V_k(z) = \frac{1 - 2|z_k|\cos(2\pi F_k T) + |z_k|^2}{1 - 2|z_k|\cos(2\pi F_k T)z^{-1} + |z_k|^2 z^{-2}}\,. \quad (5)$$

Onde z_k representa aos polos de ordem k do filtro digital; F_k é a frequência de ressonância de ordem k e T é o período de amostragem.

Como nas larguras de banda de tempo contínuo (analógico) as frequências ressonantes são cerca de $2\sigma_k$ e a frequência central é $2\pi F_k$, no plano complexo, temos que o raio da origem ao polo determina a largura de banda, como a Eq. 6:

$$|z_k| = e^{-\sigma_k T}\,, \quad (6)$$

onde $|z_k| < 1$, considerando que todos os polos correspondentes à função de transferência do sistema discreto do trato vocal, $V_k(z)$, devem estar dentro do círculo unitário, conforme necessário para a estabilidade.

Da mesma forma, o ângulo, $\theta_k = \angle z_k$ está relacionado com a frequência central dada pela Eq. 7:

$$\theta_k = 2\pi F_k T\,, \quad (7)$$

D. Modelo de Radiação

A pressão acústica na boca (incluindo os efeitos da radiação) foi modelada usando a transformada z, como na Eq. 8, tal como sugerido em [18].

$$P_L(z) = R(z)U_L(z)\,. \quad (8)$$

Sendo:

$P_L(z)$: pressão nos lábios;
$R(z)$: efeitos de radiação;
$U_L(z)$: fluxo acústico nos lábios.

Uma aproximação empregada neste trabalho, para modelar os efeitos da radiação, é dada pela Eq. 9.

$$R(z) = 1 - 0.95z^{-1}\,. \quad (9)$$

Através das equações apresentadas anteriormente, um algoritmo foi desenvolvido e implementado em MATLAB, com o objetivo de sintetizar as vogais cantadas, na língua espanhola, com características acústicas próprias de um cantor lírico da categoria soprano.

III. Formantes do Trato Vocal das Sopranos Líricas

A. Formantes das Vogais da Língua Espanhola

As frequências ressonantes do trato vocal são chamadas de frequências formantes e são fundamentais na construção dos sons sonoros. Os primeiros cinco formantes são os mais importantes: os dois formantes mais baixos determinam a maior parte do colorido da voz, enquanto o terceiro, o quarto e o quinto formantes contribuem para o timbre da voz [8].

A análise acústica das vogais espanholas ($/a/$, $/e/$, $/i/$, $/o/$, $/u/$) é realizada usando os dois primeiros formantes, de acordo com a seguinte distribuição de posicionamento, ilustrada na Fig. 3.

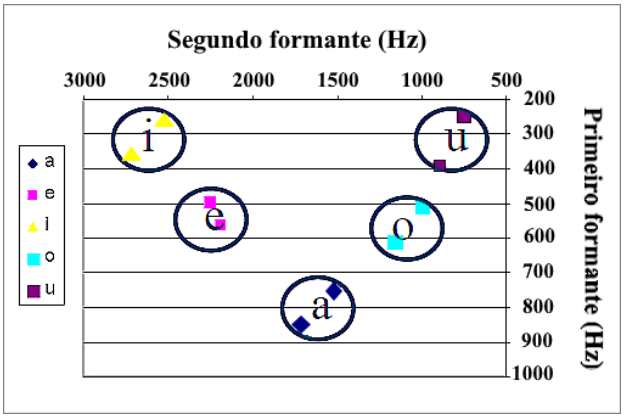


Fig. 3. Triângulo articulatório de fonemas vocálicos da língua espanhola [23].

As vogais $/i/$ e $/u/$ ocupam as áreas limítrofes (canto superior direito e esquerdo), $/a/$ ocupa a parte central inferior deste espaço entre $/i/ - /u/$ e as demais vogais espanholas estão em posições intermediárias entre $/i/ - /a/$ e $/u/ - /a/$, criando um triângulo articulatório.

As vogais $/a/$, $/i/$ e $/u/$ estão nas pontas do triângulo articulatório. Os valores possíveis da frequência do primeiro formante, F_1, estão no eixo vertical e os valores possíveis da frequência do segundo formante, F_2, estão no eixo horizontal.

B. Classificação Vocal dos Cantores Líricos

No canto, as vozes podem ser classificadas em diferentes categorias vocais. As vozes femininas são classificadas como: contralto, mezzo-soprano e soprano; enquanto que as vozes masculinas são classificadas como: baixo, barítono e tenor. [17], [25]. A Tabela. I mostra a classificação vocal.

TABELA I
CLASSIFICAÇÃO VOCAL DOS CANTORES LÍRICOS.

Masculino		Feminino	
Categoria Vocal	Extensão	Categoria Vocal	Extensão
Baixo	D2 - F4	Contralto	F3 - G5
Barítono	G2 - A4	Mezzo-Soprano	G3 - A5
Tenor	B2 - C5	Soprano	C4 - D6 / G6

As qualidades vocais podem ser classificadas de acordo com os registros de voz, isto é, cada registro corresponde a

uma região da extensão vocal composta por uma série de tons consecutivos produzidos com qualidade vocal semelhante [25]. Os registros de vozes mais conhecidos no canto masculino e feminino são: voz de peito/modal, voz média, de cabeça e apito.

Do ponto de vista fisiológico, a voz humana é produzida através de quatro mecanismos laríngeos (M0, M1, M2, M3), cada um associado a uma configuração diferente das pregas vocais, indo do mais grave, M0, ao mais agudo, M3. O mecanismo laríngeo M1 é usado para produzir voz modal, de tórax e de cabeça masculina. A voz feminina principal é produzida em M2, enquanto para atingir as notas mais altas na gama superior de sopranos leves (registro de apito) pode-se usar M3. Em particular, a primeira transição na voz de soprano pode ocorrer em torno de $E4 - F4$. Este ponto de transição corresponde à mudança M1-M2 no mecanismo laríngeo [11], [26].

O cantor pode alcançar os tons mais agudos de seu registro de voz através de dois mecanismos de técnicas de canto vocal:

1) Aumentando a pressão subglótica e consequentemente a tensão muscular nas pregas vocais. Este mecanismo é usado por cantores populares.
2) Abaixando a laringe e dilatando a faringe inferior, gerando alongamento das pregas vocais com menor variação da pressão subglótica. Esse mecanismo, chamado de cobertura de tons agudos, é usado por cantores líricos masculinos.

C. Formante do Cantor

A diferença mais considerável entre as características espectrais dos fonemas vocálicos cantados por cantores masculinos e os pronunciados por não cantores está no formante do cantor. É um pico proeminente do envelope do espectro que aparece próximo de 3 kHz (entre $2,2$ kHz e $3,8$ kHz) em todos os espectros de vogais cantadas usando a técnica de cobertura por cantores masculinos e pertencente ao típico característico de uma vogal cantada. O formante do cantor é gerado a partir do agrupamento dos formantes superiores (terceiro, quarto e quinto) com frequências próximas umas das outras.

A amplitude da potência espectral do formante do cantor (em dB) depende da classificação vocal. É mais baixo para a categoria de baixo e mais alto para a categoria de tenor. Em relação aos sopranos, a amplitude desse pico é bem menor do que nas demais categorias vocais, podendo até ser considerado um terceiro e ou quarto formante normal. Da mesma forma, o nível desse pico também varia em função da intensidade da fonação, efeito derivado da fonte glótica.

Considerando a frequência central, o pico varia dependendo da categoria de voz. Em cantores com categoria vocal de baixo, a frequência central é de cerca de $2,2$ kHz; em barítonos, cerca de $2,7$ kHz; em tenor, cerca de $3,2$ kHz; e em altos, em torno de $2,8$ kHz. Essas diferenças de frequência parecem contribuir significativamente para as diferenças de timbres entre essas categorias de voz [8], [27].

D. Estimação das Frequências Formantes

As frequências dos formantes das vogais espanholas cantadas foram obtidas a partir de sinais de vozes experimentais produzidos por 2 cantores líricos, em um total de 23 registros (uterações). Os sinais foram gravados em condições semelhantes: ambiente silencioso (estúdio de gravação musical) e vozes gravadas diretamente num computador. Os sujeitos eram 2 mulheres treinadas profissionalmente em canto lírico na categoria vocal de soprano, com idades de 22 e 24 anos. Ambas têm pelo menos 6 anos de treinamento regular de canto lírico.

Cada cantora sustentou por $9\,s$ cada uma das vogais espanholas nas diferentes notas de seu alcance vocal. Elas começaram a cantar a vogal $/a/$ no tom mais baixo de sua extensão vocal e, continuamente, alcançaram o tom mais alto que podiam produzir, seguindo uma escala diatônica. A primeira cantora, denominada soprano 1, produziu as vogais cantadas desde a nota musical $F4$ até $E5$ (349 Hz - 659 Hz), e a segunda cantora, denominada soprano 2, produziu vogais cantadas desde a nota musical $E3$ até $F5$ (167 Hz - 698 Hz).

Os valores das cinco primeiras frequências de ressonância e suas respectivas larguras de faixa foram extraídos por meio de um programa de análise e síntese de fala (Praat, versão 5.3). Da mesma forma, a frequência fundamental para cada nota musical entoada foi salva em uma planilha para ser utilizada posteriormente na síntese da voz cantada. Essas informações podem ser obtidas em https://cutt.ly/ayFS4qr.

IV. Avaliação Perceptual das Sínteses

As sínteses foram separadas por notas musicais e correspondem às que foram comuns entre as duas sopranos durante as gravações iniciais e submetidas a um grupo de pessoas para que pudessem fazer uma avaliação, chamado de *teste de escuta*.

A. Teste de Escuta

O teste de escuta foi realizado por dez brasileiros e dez colombianos, com idades entre 20 e 53 anos, sendo oito mulheres e doze homens.

O mesmo grupo classificou as vozes sintetizadas em relação à naturalidade e, também, à inteligibilidade, atribuindo notas, variando de 0, no pior caso, a 5, no caso em que a síntese apresentava-se excelente.

Os dados obtidos nos testes perceptuais de naturalidade e inteligibilidade foram discutidos através das técnicas de estatística descritiva: obtenção, análise e interpretação de dados através de tabelas, gráficos e indicadores numéricos.

A informação foi obtida através de um questionário desenvolvido para classificar a naturalidade e a inteligibilidade das sínteses das vogais cantadas, enviado a cada um dos ouvintes do grupo. Os dados obtidos são de natureza quantitativa, organizados em tabelas e apresentados em gráficos de barras.

A média aritmética foi utilizada como medida de tendência central e permitiu identificar o modo como os dados dos testes perceptuais foram distribuídos. O desvio padrão foi utilizado como medida de dispersão.

V. Resultados

A. Avaliação Perceptual de Naturalidade

O grupo escolhido para avaliação foi solicitado a classificar 14 sínteses da voz cantada por sopranos como natural ou

artificial, sem receber nenhuma informação adicional sobre os estímulos. A avaliação podia adquirir valores discretos desde 0 até 5, sendo 0 uma síntese artificial e 5 um excelente som natural. Esse teste foi realizado pelos ouvintes individualmente.

O teste perceptual realizado segue as técnicas aplicadas para estimar a naturalidade de sons sintetizados gerados por modelos de trato vocal como feito em [28], [29].

Os dados recolhidos na avaliação perceptual de naturalidade foram organizados numa tabela. Essa informação pode ser obtida em: https://cutt.ly/gh0cDAH. As linhas da tabela referem-se às avaliações dos ouvintes e as colunas referem-se ao tom sustentado por cada soprano.

O gráfico de barras da Fig. 4 apresenta os resultados da avaliação perceptual de naturalidade na síntese da voz cantada com os parâmetros de ressonância das sopranos 1 e 2, no intervalo $F4 - E5$.

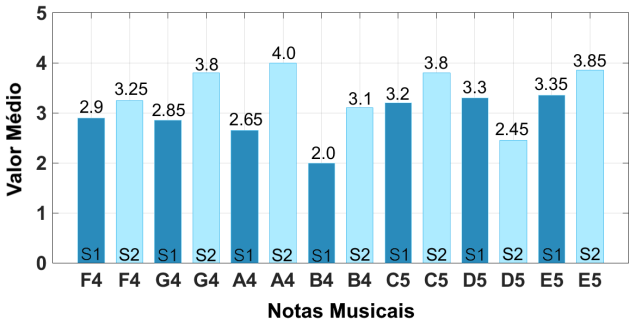


Fig. 4. Gráfico de barras do teste perceptual das sínteses de vogais cantadas por sopranos considerando a média de naturalidade.

A maioria das sínteses com os dados da soprano 2 obtiveram uma média de naturalidade superior em comparação às sínteses da voz cantada com os dados da soprano 1.

As notas musicais $G4$, $A4$, $C5$ e $E5$ produzidas com as frequências de ressonância da soprano 2 obtiveram uma média de naturalidade de $3,8$, $4,0$, $3,8$ e $3,85$ respectivamente. As sínteses das vogais cantadas apresentaram características robóticas devido às qualidades do sinal glótico: a forma de onda do pulso e ausência de *jitter* e *shimmer*. Por outro lado, a adição das caraterísticas acústicas de vibrato e tremolo contribuíram satisfatoriamente à naturalidade das sínteses.

A partir dos resultados do teste de escuta observou-se que a maioria das notas musicais que apresentaram a menor média de naturalidade correspondem às sínteses geradas com as frequências de ressonância da soprano 1, ainda que as qualidades do sinal glótico sejam as mesmas que as utilizadas nas sínteses das vogais produzidas com as frequências de ressonância da soprano 2.

Na Tabela. II apresenta-se o desvio padrão, σ, dos dados correspondentes às avaliações perceptuais de naturalidade.

TABELA II
DESVIO PADRÃO DOS DADOS DAS AVALIAÇÕES PERCEPTUAIS DE NATURALIDADE.

Soprano 1							
Tom	$F4$	$G4$	$A4$	$B4$	$C5$	$D5$	$E5$
σ	1,21	1,27	1,14	1,45	1,32	1,22	1,23
Soprano 2							
Tom	$F4$	$G4$	$A4$	$B4$	$C5$	$D5$	$E5$
σ	1,16	0,83	1,03	1,33	1,01	1,15	0,99

Os desvios calculados para as notas musicais $G4$ e $E5$ sustentadas pela soprano 2 permitiram identificar uma menor variabilidade dos dados e, portanto, uma maior confiabilidade nos resultados obtidos.

B. Avaliação Perceptual de Inteligibilidade

O mesmo grupo de ouvintes classificou as sínteses entre inteligíveis ou desconhecidas, sem receber nenhuma informação adicional sobre os estímulos. As sínteses foram separadas por notas musicais e contêm as vogais, da língua espanhola, na seguinte ordem: ($/a/$, $/e/$, $/i/$, $/o/$, $/u/$). A avaliação constava de valores discretos de 0 até 5, sendo 0 desconhecido e 5 completamente inteligível. Esse teste também foi realizado pelos ouvintes individualmente.

Os dados recolhidos na avaliação perceptual de inteligibilidade foram organizados em tabelas e classificados de acordo com a nota musical sustentada ($F4$ - $E5$). As colunas da tabela referem-se às cinco vogais espanholas e as linhas referem-se às avaliações dos ouvintes para as sínteses com parâmetros de ressonância das sopranos 1 e 2. Os dados podem ser obtidos acessando o seguinte link: https://cutt.ly/sh2v7Hp.

O gráfico de barras da Fig. 5 apresenta as cinco vogais da língua espanhola com maior média de inteligibilidade entre todas as sínteses submetidas no teste perceptual, e os áudios correspondentes a essas sínteses podem ser ouvidos em https://cutt.ly/4yVGUcr.

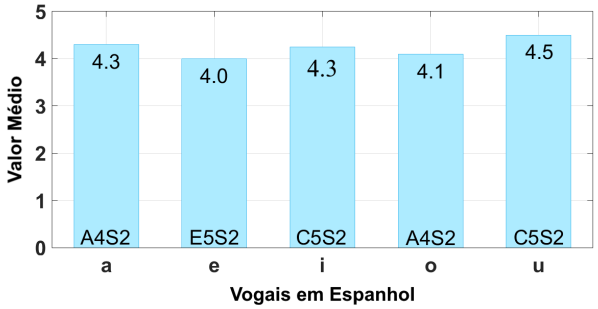


Fig. 5. Gráfico de barras do teste perceptual das síntese de vogais cantadas por sopranos considerando a média de inteligibilidade.

As sínteses das vogais espanholas apresentadas na Fig. 5 são de grande interesse para uma subsequente analise espectral pois não só alcançaram a maior média de inteligibilidade más também foram sustentadas nas notas musicais ($A4$, $C5$ e $E5$) com a maior média de naturalidade entre todas as sínteses submetidas aos testes perceptuais.

Na Tabela. III apresenta-se o desvio padrão, σ, dos dados correspondentes as avaliações perceptuais de inteligibilidade das cinco vogais espanholas.

TABELA III
DESVIO PADRÃO DOS DADOS DAS AVALIAÇÕES PERCEPTUAIS DE INTELIGIBILIDADE.

Vogal	Tom	Cantora	Média	σ
$/a/$	$A4$	$S2$	4,3	0,86
$/e/$	$E5$	$S2$	4,0	0,92
	$F4$	$S1$	4,0	1,72
$/i/$	$C5$	$S2$	4,3	1,02
$/o/$	$A4$	$S2$	4,1	0,85
$/u/$	$C5$	$S2$	4,5	0,83

Os desvios para as avaliações perceptuais da inteligibilidade das vogais espanholas apresentadas na Fig. 5 permitiram identificar uma menor variabilidade dos dados em torno da média ($\sigma \leq 1,02$).

Observa-se que a vogal $/e/$ produzida com os parâmetros de ressonância da soprano 2 apresentou um desvio padrão de $0,92$ sendo inferior ao calculado para a mesma vogal produzida com os parâmetros de ressonância da soprano 1. Devido a esse resultado, considerou-se uma maior confiabilidade na inteligibilidade da vogal $/e/$ produzida com os parâmetros de ressonância da soprano 2.

C. Formante do Cantor em Sopranos

Na Fig. 6 apresentam-se as respostas em frequência das cinco vogais espanholas que alcançaram a maior média de inteligibilidade no teste perceptual.

A resposta em frequência foi obtida em 1024 amostras abrangendo o círculo unitário completo e usando uma janela retangular sem sobreposição entre as amostras.

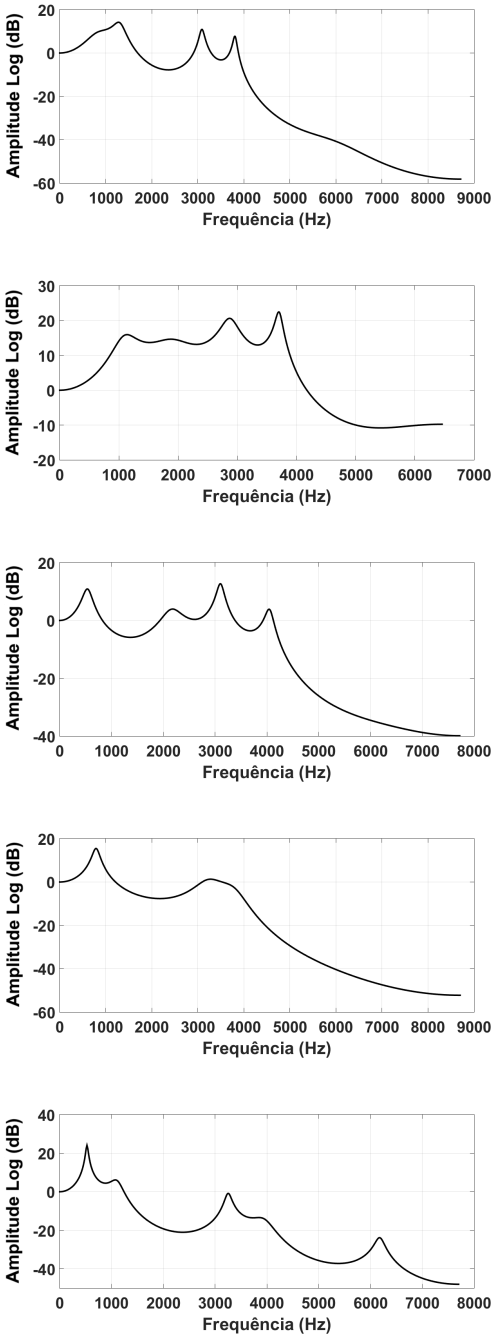


Fig. 6. Resposta em frequência para as cinco vogais espanholas cantadas que apresentaram a maior média de inteligibilidade (a) vogal $/a/$, (b) vogal $/e/$, (c) vogal $/i/$, (d) vogal $/o/$ e (e) vogal $/u/$.

Na resposta em frequência da vogal $/e/$ observa-se que, em torno de 3 kHz, a energia acústica eleva-se para acima de 20 dB, porém a média de inteligibilidade dessa vogal é a mais baixa de todas as vogais, evidenciando-se que ocorre um efeito adverso na inteligibilidade da vogal quando a energia acústica do som eleva-se.

É importante notar um pico do envelope espectral em torno de 3 kHz, causado pelo agrupamento das frequências formantes superiores (F_3, F_4, F_5). Em relação a esse agrupamento, F_5 está longe em frequência de F_4. Portanto, não participa

adequadamente da construção do formante do cantor. Esta é uma característica comum nas cinco respostas em frequência desenvolvidas.

As cinco primeiras frequências formantes (em Hz) correspondem às vogais espanholas que alcançaram a maior média de inteligibilidade e aparecem na Tabela. IV. Assim, pode-se observar que os valores de F_1 das vogais $/e/$, $/i/$, $/o/$ e $/u/$ são superiores aos valores de F_1 da voz falada que foram apresentados no triangulo articulatório das vogais espanholas da Fig. 3. Esse resultado deve-se ao fato que a cantora ajustou a F_1 das vogais aos valores de F_0 das notas musicais sustentadas.

No caso da vogal $/a/$, a F_1 da voz falada foi superior à F_0 da nota musical sustentada e, portanto, significa que a cantora não precisou ajustar a F_1. [30], [31].

TABELA IV
FREQUÊNCIAS FORMANTES (EM HZ) UTILIZADAS PARA PRODUZIR AS VOGAIS ESPANHOLAS CANTADAS.

Tom	F_0	Vogal	F_1	F_2	F_3	F_4	F_5
A4	435	a	808	1304	3088	3808	6025
E5	646	e	1091	1891	2873	3706	6869
C5	515	i	536	2159	3104	4051	6276
A4	435	o	790	1551	3226	3795	5928
C5	513	u	526	1108	3251	3947	6179

As larguras de faixa das cinco primeiras frequências formantes que foram utilizadas para produzir as vogais espanholas e que alcançaram a maior média de inteligibilidade nesta pesquisa são mostradas na Tabela. V e representam os efeitos da perda do trato vocal [18].

Em particular, as larguras de faixa das frequências formantes utilizadas para produzir a vogal $/u/$ são relativamente baixas (menos de $200\,Hz$), refletindo baixas perdas no trato vocal da soprano 2 e, portanto, a propagação do som através do trato vocal é mais eficiente que nas outras vogais (/a/, /e/, /i/, /o/). Essa característica influi nos resultados positivos de naturalidade e inteligibilidade da vogal $/u/$ sustentada pela soprano 2 na nota musical $C5$.

TABELA V
LARGURA DE FAIXA (EM HZ) DAS FREQUÊNCIAS DOS FORMANTES UTILIZADAS PARA PRODUZIR AS VOGAIS ESPANHOLAS CANTADAS.

Tom	Vogal	Bw_1	Bw_2	Bw_3	Bw_4	Bw_5
A4	a	282	126	53	40	808
E5	e	195	357	146	59	998
C5	i	90	202	68	76	1411
A4	o	77	1676	291	286	2153
C5	u	23	116	62	198	87

VI. CONCLUSÕES

O modelo de fonte-filtro de Fant foi usado para produzir a síntese de vogais cantadas, da língua espanhola, por cantores sopranos, incluindo os fenômenos de vibrato e tremolo, muito usados no canto lírico. Foi observado que não só as características do sinal glótico contribuíram para a naturalidade das sínteses mas, também, a relação entre as frequências formantes e a frequência fundamental, F_0.

As sínteses foram submetidas a testes perceptuais, realizadas por grupos de brasileiros e colombianos. As sínteses das vogais cantadas nas notas musicais $G4$, $A4$, $C5$ e $E5$ alcançaram a maior média de naturalidade das avaliações e as sínteses nas notas musicais $G4$ e $E5$ apresentaram a menor dispersão em torno da média. Em termos de inteligibilidade, as maiores médias obtidas foram para as mesmas notas musicais que obtiveram as maiores médias de naturalidade. Particularmente, a vogal $/u/$ sustentada pela soprano 2 na nota musical $E5$ alcançou a média máxima de inteligibilidade (Média=4,5) e uma baixa medida de dispersão em torno da média (S=0,83).

Nas análises espectrais realizadas para as sínteses das vogais espanholas que alcançaram as maiores médias de inteligibilidade observou-se um pico no espectro em torno dos 3 kHz, conhecido como formante do cantor, que foi construído pelo agrupamento das frequências formantes F_3 e F_4.

De acordo com as análises espectrais de todas as sínteses, observou-se que o ajuste de formantes depende da vogal produzida e, quando a energia do som em torno de 3 kHz é elevada, a média de inteligibilidade das vogais decresce, mostrando o efeito adverso do ajuste de formantes, neste caso, para a inteligibilidade das vogais.

A apropriada relação entre as frequências formantes e a frequência fundamental teve como consequência o fato de que a maioria das sínteses com os parâmetros de ressonância da soprano 2 obtiveram uma média de naturalidade e inteligibilidade superior em comparação às sínteses da voz cantada com os parâmetros de ressonância da soprano 1.

Novas perspectivas emergiram deste trabalho. Modelos mais complexos do pulso glótico podem ser usados para desenvolver sínteses das vogais espanholas cantadas com as características de ressonância do soprano 2. Consequentemente, essas sínteses podem ser submetidas a testes perceptuais com os mesmos ouvintes, desenvolvendo-se assim uma comparação dos modelos glóticos quanto à média de naturalidade e inteligibilidade. Por outro lado, diferentes efeitos acústicos também podem ser explorados para melhorar a naturalidade e expressividade das sínteses, principalmente da soprano 1 que não obteve boa média nos testes perceptuais desta pesquisa.

ACKNOWLEDGMENT

Este trabalho foi financiado pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Grant 303234 / 2017-2) - Brasil.

REFERÊNCIAS

[1] C. C. Lochbaum and J. L. Jr. Kelly, "Speech synthesis," *Proceedings of the Speech Communication Seminar*, pp. 583-596, 1962.

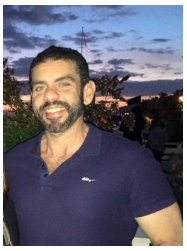
[2] M. Mellody, F. Herseth, G. H. Wakefield, "Modal distribution analysis, synthesis, and perception of a soprano's sung vowels," *Journal of Voice*, vol. 15, n. 4, pp. 469-482, 2001.

[3] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, k. Tokuda, "Singing Voice Synthesis Based on Deep Neural Networks," *Interspeech*, pp. 2478-2482, September 2016.

[4] N. D'Alessandro, P. Woodruff, Y. Fabre, T. Dutoit, S. Le Beux, B. Doval, C. d'Alessandro, "Realtime and accurate musical control of expression in singing synthesis," *Journal on Multimodal User Interfaces*, vol. 1, n. 1, pp. 31-39, 2007, doi: 10.1007/BF02884430.

[5] P. Chandna, M. Blaauw, J. Bonada, and E. Gomez, "Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan," *in 2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1-5, 2019, doi: 10.23919/EUSIPCO.2019.8903099..
[6] T. Nose, M. Kanemoto, T. Koriyama, T. Kobayashi, "Hmm-based expressive singing voice synthesis with singing style control and robust pitch modeling," *Computer Speech & Language*, vol. 34, n. 1, pp. 308–322, 2015.
[7] L. Ardaillon, *Synthesis and expressive transformation of singing voice*. Ph.D. thesis, Paris 6, 2017.
[8] J. Sundberg, "Vocal tract resonance in singing," *The NATS Journal*, vol. 44, n. 4, pp. 11-20, 1988.
[9] P. P. de Julián, "Modificación o aggiustamento de las vocales españolas en el canto lírico," *Estudios de fonética experimental*, pp. 263-293, 2016.
[10] J. Sundberg, "Research on the singing voice in retrospect," *TMH-QPSR*, vol. 45, n. 1, pp. 11-22, 2003.
[11] M. Garnier, N. Henrich, L. Crevier-Buchman, C. Vincent, J. Smith, J. Wolfe, "Glottal behavior in the high soprano range and the transition to the whistle register," *The Journal of the Acoustical Society of America*, vol. 131, n. 1, pp. 951-962, 2012.
[12] I. R. Titze, *Principles of voice production*. Prentice Hall, 1994.
[13] E. Cataldo and C. Soize, "Stochastic mechanical model of vocal folds for producing jitter and for identifying pathologies through real voices," *Journal of biomechanics*, vol. 74, pp. 126–133, June 2018.
[14] N. H. Bernardoni, *Etude de la source glottique en voix parlée et chantée: modélisation et estimation, mesures acoustiques et électroglottographiques, perception*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI (UPMC), 2001.
[15] G. Fant, "Acoustic theory of speech production.'sgravenhage: Mouton," *The Netherlands*, 1960.
[16] G. Fant, "The source filter concept in voice production," *STL-QPSR*, vol. 22, n. 1, pp. 21–37, 1981.
[17] M. L. Facal, *La voz del cantante: estúdio comparativo del análisis objetivo y subjetivo de la voz hablada y cantada*. Librería Akadia Editorial, 2005.
[18] L. R. Rabiner, and R. W. Schafer, *Theory and applications of digital speech processing*. vol. 64, Pearson Upper Saddle River, NJ, 2011.
[19] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, n. 2B, pp. 583-590, 1971.
[20] B. Doval, C. d'Alessandro, N. Henrich, "The spectrum of glottal flow models," *Acta acustica united with acustica*, vol. 92, n. 6, pp. 1026–1046, 2006.
[21] M. Hirano, S. Hibi, S. Hagino, "Physiological aspects of vibrato," *Vibrato*, pp. 9-33, 1995.
[22] T. L. Nwe, H. Li, "Exploring Vibrato-Motivated Acoustic Features for Singer Identification," *IEEE Trans. Audio Speech Lang. Process*, vol. 15, n. 2, pp. 519–530, 2007, doi: 10.1109/TASL.2006.876756.
[23] M. Salas, "Aplicaciones del análisis acústico en los estudios de la voz humana," *Seminario Internacional de Acústica*, 2003.
[24] L. Regnier, G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1685–1688, 2009.
[25] S. Hertegard, J. Gauffin, J. Sundberg, "Open and covered fiberoptics, inverse singing as studied by means of filtering, and spectral analysis," *J Voice*, vol. 4, pp. 220-230, 1990.
[26] M. Kob, N. Henrich, H. Herzel, D. Howard, I. Tokuda, J. Wolfe, "Analysing and understanding the singing voice: recent progress and open questions," *Current bioinformatics*, vol. 6, n. 3, pp. 362–374, 2011.
[27] T. J. Millhouse, F. Clermont, "Perceptual characterisation of the singerâ€™s formant region: a preliminary study," *Proceedings of the Eleventh Australian International Conference on Speech Science and Technology*, pp. 253–258, 2006.
[28] D. R. Allen, W. J. Strong, "A model for the synthesis of natural sounding vowels," *The Journal of the Acoustical Society of America*, vol. 78, n. 1, pp. 58-69, 1985.
[29] E. Cataldo, F. R. Leta, J. Lucero, L. Nicolato, "Synthesis of voiced sounds using low-dimensional models of the vocal cords and time-varying subglottal pressure," *Mechanics Research Communications*, vol. 33, n. 2, pp. 250–260, 2006.
[30] G. Berndtsson, J. Sundberg, "Perceptual significance of the center frequency of singer's formant," *Scandinavian Journal of Logopedics and Phoniatrics*, vol. 20, n. 1, pp. 35–41, 1995.
[31] R. Weiss, W. Brown Jr, J. Moris, "Singer's formant in sopranos: fact or fiction?," *Journal of Voice*, vol. 15, n. 4, pp. 457–468, 2001.

Luis E. Barrientos was born in Lourdes, N. de S, Colombia in 1987. He received the B.S. degree in electronic engineering from the Universidad Francisco de Paula Santander, Cúcuta, in 2011 and received the M.S. degree in electrical and telecommunications engineering from the Universidade Federal Fluminense, Niterói, Rio de Janeiro, in 2018. He is currently pursuing the D.S. degree in electrical and telecommunications engineering at Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brazil. He has developed the synthesis of singing voice since 2017. He has been a researcher in the Department of Electrical and Telecommunications Engineering, Universidade Federal Fluminense, since 2016. His research interest includes the development of the synthesis of voice, digital signal processing, and biomedical applications.



Edson Cataldo was born in Rio de Janeiro, Brazil, in 1967. He received the degree in Telecommunications Engineering and also the M.S. degree in Mathematics from the Universidade Federal Fluminense, Brazil, in 1988 and in 1993, respectively. In 2000, he received the D.S. degree in Mechanical Engineering and he has completed a posdoc stage in 2006, in Université Gustave-Eiffel, Paris, France. He is professor in the Telecommunications Engineering in the Universidade Federal Fluminense. His research interest includes digital signal processing, voice synthesis, probabilistic modeling and biomedical applications.