Re-síntese Concatenativa de Voz Cantada

Nuno Fonseca[1], Aníbal Ferreira[2], Ana Paula Rocha[3]

[1] DEI/CIIC/ESTG, Instituto Politécnico de Leiria
[2] Faculdade de Engenharia, Universidade do Porto
[3] LIACC, Faculdade de Engenharia, Universidade do Porto
nuno.fonseca@ipleiria.pt; ajf@fe.up.pt; arocha@fe.up.pt

Abstract. Este *paper* apresenta um resumo do trabalho realizado no âmbito da re-síntese da voz cantada, isto é, permitir que o utilizador possa controlar directamente um sintetizador de voz cantada, usando a sua própria voz. Desta forma, o sintetizador deverá ser capaz de replicar, da melhor forma possível, a mesma melodia, a mesma sequência de fonemas e a mesma performance musical. Com base numa abordagem concatenativa, extraindo a componente dinâmica, altura e informação fonética, e usando um sistema de selecção de unidades e semelhança fonética, uma sequência de *frames* da biblioteca interna de sons é escolhida para replicar a performance do áudio original. Embora ainda existam artefactos áudio, impedindo o seu uso em aplicações profissionais, foi criado o conceito de uma nova ferramenta áudio, que apresenta grande potencial para trabalhos futuros, não apenas para voz cantada, mas para fala ou outros domínios musicais.

Keywords: re-síntese, voz, canto, síntese, análise, música.

1 Introdução

A voz cantada sempre teve um papel importante nas nossas vidas, e embora os sintetizadores tentem replicar todos os instrumentos musicais existentes, apenas nos últimos nove anos as primeiras soluções comerciais de síntese de voz cantada apareceram (baseadas no trabalho de [1] e [2]), permitindo a combinação de música e texto, ou seja, "cantar". Essas soluções podem criar resultados realistas em determinadas situações, mas requerem processos muito morosos e utilizadores experientes.

Posteriormente surgiram alguns trabalhos em que a voz do utilizador é usada ajudar a controlar a síntese de voz cantada [3][4], no entanto o processo não é totalmente automático, sendo sempre necessária a introdução da informação fonética.

Este *paper* apresenta um resumo do trabalho realizado no âmbito da re-síntese da voz cantada, isto é, permitir que o utilizador possa controlar directamente um sintetizador de voz cantada, usando a sua própria voz. O sintetizador deverá ser capaz de replicar, da melhor forma possível, a mesma melodia, a mesma sequência de fonemas e a mesma performance musical.

Para além da sua aplicação como controlo de instrumentos virtuais de voz cantada, o mesmo princípio pode ser aplicado como efeito áudio, transformando uma gravação de voz cantada, substituindo a voz original por amostras sonoras de uma outra voz existente na biblioteca interna do sistema. Em vez da utilização de técnicas de processamento de sinal que permitam efectuar alterações ao sinal original (várias técnicas são descritas em [5]), através de uma abordagem ao nível semântico consegue-se reconstruir um novo *stream* áudio com um novo timbre.

A secção 2 apresenta o sistema proposto, incluindo os seus módulos internos e a biblioteca de som usada. A secção 3 apresenta alguns resultados experimentais. As conclusões são apresentadas na secção 4.

2 Sistema

O sistema proposto baseia-se na análise e futura síntese concatenativa de voz cantada, usando uma biblioteca de sons interna com gravações de voz, estando dividido em três módulos: extracção de informação, selecção de unidades e síntese (figura 1).

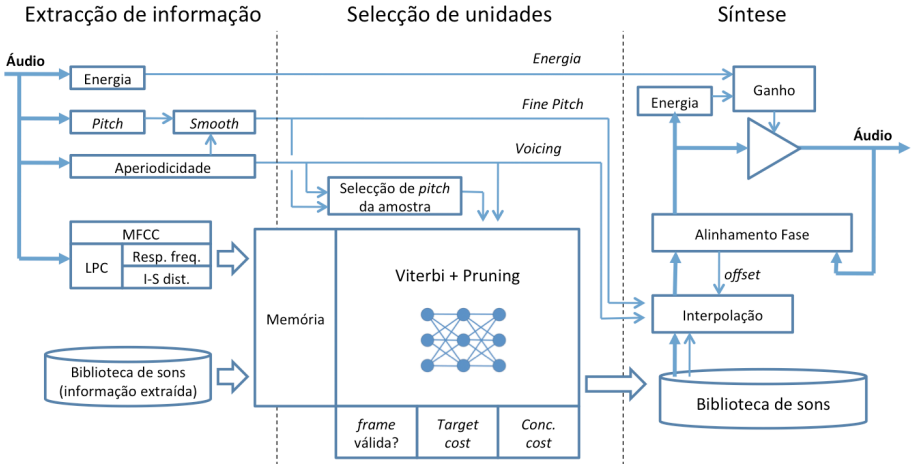

Fig. 1. Sistema de re-síntese de voz cantada.

2.1 Extracção de informação

No primeiro módulo, o *stream* de áudio original é analisado, e a informação relevante é extraída, nomeadamente energia, altura (*pitch*), aperiodicidade e um conjunto de parâmetros relativos à componente fonética (coeficientes MFCC e LPC). A altura e aperiodicidade são extraídas através do método YIN [6].

2.2 Selecção de unidades

O segundo módulo é responsável pela escolha dos segmentos da biblioteca interna de sons que irão ser usados posteriormente durante a síntese. Para cada *frame* original, o modulo de selecção de unidades irá escolher a *frame* interna (da biblioteca de sons) que oferece a menor distância fonética relativamente à *frame* original.
O sistema utiliza o conceito de *target cost*/*concatenation cost* [7]. O *target cost* preocupa-se com a semelhança entre a *frame* original e a *frame* escolhida, enquanto o *concatenation cost* se preocupa com a diferença entre a *frame* escolhida e a *frame* escolhida no instante anterior, de forma a que a futura concatenação não apresente mudanças demasiado abruptas.
A fórmula para o *target cost* é baseada num distância Euclidiana (eq.1) com quatro domínios: MFCC, resposta em frequência do LPC, distância Itakura-Saito do LPC e aperiodicidade. O *concatenation cost* apenas considera a resposta em frequência do LPC (eq. 3).

$$D_{i,j} = \frac{\sqrt{D_{MFCC}(i,j)^2 + D_{LPC\ resp}(i,j)^2 + D_{LPC\ dist}(i,j)^2 + D_{Ap}(i,j)^2}}{2} \tag{1}$$

$$D_{MFCC}(i,j) = \frac{\sqrt{\sum_1^{12}(c_n^i - c_n^j)^2}}{norm_1} \tag{2}$$

$$D_{LPC\ resp}(i,j) = \frac{\sqrt{\sum_1^{128}(X_n^i - X_n^j)^2}}{norm_2} \tag{3}$$

$$D_{LPC\ dist}(i,j) = \frac{\left[D_{IS}\big(LPC(i), LPC(j)\big) + D_{IS}\big(LPC(j), LPC(i)\big)\right]}{norm_3} \tag{4}$$

$$D_{Ap}(i,j) = \frac{|Ap(i) - Ap(j)|}{norm_4} \tag{5}$$

A procura da sequência de *frames* internas com o menor custo total é efectuada com base numa procura Viterbi [8].

2.3 Síntese

O ultimo módulo, responsável pela síntese final, efectua pequenas alterações às *frames* escolhidas, concatenando-as de forma a obter o *stream* áudio final. Para tal, a altura (*pitch*) da *frame* é acertada para o valor da *frame* original, seguindo-se o alinhamento da fase com as *frames* de saída anteriores. Por fim, o ganho da *frame* é alterado de forma a ter a mesma energia da *frame* original.

2.4 Biblioteca de sons

A biblioteca de sons interna é constituída por áudio pré-gravado existente em [9], correspondendo a uma voz solista feminina, cantando 46 palavras, em 2 oitavas (ficheiro independente para cada par nota/palavra).

3 Resultados Experimentais

O sistema foi testado com fragmentos áudio das seguintes canções, sendo a duração do fragmento especificada entre parêntesis:

- Amazing Grace - LeAnn Rimes (0:16)
- Bohemian Rhapsody - Lauryn Hill (0:11)
- Frozen - Madonna (0:15)
- I Will Survive - Diana Ross (0:10)
- Tom's Diner - Susanne Vega (0:04)
- Whenever - Shakira (0:06)

Cada fragmento original contem apenas a voz da vocalista.

Os ficheiros originais, assim como os resultados re-sintetizados podem ser consultados em:

http://www.estg.ipleiria.pt/~nuno.fonseca/papers/dsp2011/SingingVoiceResynthesisExamples.zip

Embora genericamente a componente dinâmica e de pitch apresentem uma boa reconstrução, a componente fonética ainda contém artefactos significativos, criando transições abruptas em determinadas situações, o que ainda impede a sua utilização em aplicações profissionais.

4 Conclusão

Um sistema de re-síntese de voz cantada foi apresentado. Embora ainda existam artefactos áudio, impedindo o seu uso em aplicações profissionais, foi criado o conceito de uma nova ferramenta áudio, que apresenta muito potencial para trabalhos futuros, não apenas para voz cantada, mas para fala ou outros domínios musicais.

References

1. Bonada, J., Serra, X.: Synthesis of the Singing Voice by Performance Sampling and Spectral Models. IEEE Signal Processing Magazine, 24, pp. 67-79, 2007.

2. Fonseca, N.: VOTA Utility: Making the computer sing. Proc. of 114th AES Convention; Amsterdam, Netherlands, March 2003.
3. Janer, J., Bonada, J., Blaauw, M.: Performance-Driven Control For Sample-Based Singing Voice Synthesis. Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06), Montreal, Canada, September 18-20, 2006J.
4. Nakano, T., Goto, M.: VocaListener: a singing-to-singing synthesis system based on iterative parameter estimation. SMC 2009, Porto, Portugal, July 2009.
5. Zolzer, U.: DAFX – Digital Audio Effects. John Wiley & Sons, Ltd, 2002.
6. Cheveigné, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, Vol. 111, No. 4. (2002), pp. 1917-1930.
7. Hunt, A. J., Black, A. W.: Unit selection in a concatenative speech synthesis system using a large speech database. ICASSP96, volume 1, 7-10 May 1996, pp. 373 – 376.
8. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, p. 257-286.
9. EASTWEST: EASTWEST/Quantum Leap Voices of Passion. information available at http://www.soundsonline. com/Quantum-Leap-Voices-Of-Passion-Virtual-Instrument -pr-EW-174.html, Accessed in Feb. 2010.