SÍNTESE DE VOZ: ESTADO DA CUESTIÓN E PERSPECTIVAS

*Eduardo Rodríguez Banga*Dpto. Teoría do Sinal e Comunicacóns
Universidade de Vigo

1. INTRODUCCIÓN

Un conversor texto-voz é un sistema que a partir dun texto xera un sinal de voz sintético que trata de emular a forma en que unha persoa lería ese mesmo texto. A estrutura xenérica dun conversor texto-voz ilústrase na figura 1. Basicamente consta de dúas etapas: a primeira, que denominamos procesamento lingüístico-prosódico, é a encargada de facer unha análise lingüística do texto de entrada encamiñada a obter a secuencia de sons a xerar (transcrición fonética) e a súa correspondente entoación, duración e enerxía (información prosódica); a segunda, denominada procesamento acústico, é a encargada de sintetizar un sinal de voz coas características dadas pola etapa anterior.

Figura 1: *Etapas básicas dun sistema de conversión texto-voz*

2. PROCESAMENTO LINGÜÍSTICO E PROSÓDICO

Na figura 1 ilústranse as etapas básicas do procesamento lingüístico e prosódico nun conversor texto-voz. A práctica totalidade deste tipo de sistemas realizan un procesamento frase a frase do texto de entrada. Porén, a delimitación das distintas frases que compoñen un texto non é sempre doada, xa que os mesmos signos de puntuación que se soen empregar para marcar o inicio e final das frases poden aparecer nalgúns tipos de construcións

(abreviaturas, por exemplo) ou ben estar ausentes –como ocorre comunemente nos titulares dos periódicos.

Ademais das abreviaturas, nun texto escrito atopamos frecuentemente outras construción especiais (cifras numéricas, datas, siglas, etc.) que é necesario tratar axeitadamente. Polo tanto, é necesaria unha etapa de pre-procesamento do texto na que este tipo de estruturas sexa transformada nunha secuencia de palabras equivalente, o que non sempre é doado. Por exemplo, a frase “Carlos I viviu no século XVI” amosa como os números romanos ás veces son transformados en numerais cardinais e outras en ordinais. Noutras ocasións a expansión dos numerais precisa de información de xénero como por exemplo en “Carlos II e Isabel II” ou “22 libros e 22 mesas”.

Despois de procesar o texto de entrada, débese determinar a función de cada unha das palabras dentro da frase. Para iso, en primeiro lugar, realízase unha análise morfolóxica na que se especifican todas as posibles categorías gramaticais por cada palabra, considerada de forma illada e, posteriormente, realízase unha desambiguación gramatical tendo en conta o contexto en que se encontra (análise morfosintáctica). Nalgún sistemas realízase, ademais, unha sinxela análise sintáctica que divide a frase de entrada en distintos sintagmas.

Con toda a información lingüística recollida nas etapas anteriores, procédese a realizar a transcrición fonética do texto de entrada e a xerar a súa información prosódica asociada. O modelado prosódico será obxecto dunha das seccións deste traballo.

3. PROCESAMIENTO ACÚSTICO

Esta etapa é a encargada de xerar o sinal de voz sintético a partir da transcrición fonética e información prosódica de entrada. Existen diversos tipos de técnicas de xeración que podemos clasificar como:

Métodos baseados en regras: a xeración do sinal de voz faise segundo uns parámetros con valores determinados mediante un conxunto de regras deseñadas por expertos. Dentro deste tipo de métodos podemos considerar os primeiros sintetizadores de formantes e os modelos articulatorios. Trátase de modelos moi flexibles, xa que poden modificar calquera característica do sinal de voz, pero tamén moi complexos.

Métodos de concatenación de unidades: tratan de paliar o descoñecemento do proceso de fonación humano empregando unidades de voz

pregravadas. Son capaces de producir fala altamente intelixible sen demasiado esforzo, aínda que son pouco flexibles.

Métodos mixtos: empregan parametrizacións (traxectorias dos formantes, movementos articulatorios, etc.) obtidas de forma máis ou menos automática a partir dun amplo conxunto de unidades pregravadas.

3.1. *Métodos baseados en regras*

3.1.1. Síntese por formantes

Os sintetizadores de formantes modelan as distintas resonancias (formantes) do tracto vocal mediante unha serie de filtros dispostos en cascada ou paralelo. Cada un destes filtros modela un formante do sinal de voz, dependendo o número de formantes do ancho de banda considerado. Por exemplo, nun ancho de banda de 5 KHz é suficiente considerar uns cinco ou seis formantes. O sinal de entrada á configuración de filtros en cascada ou paralelo consiste, basicamente, nun sinal periódico no caso de sons sonoros e nun sinal aleatorio (ruído) se se trata dun son xordo.

Aínda que durante moitos anos houbo unha gran controversia sobre cal das configuracións –cascada ou paralelo– era máis axeitada, a discusión quedou practicamente finalizada cando se propuxo (Klatt, 1980) un modelo que combina as dúas configuracións, empregando unha ou outra dependendo do tipo de son. Klatt demostrou que axustando adecuadamente os parámetros deste modelo podía obterse fala indistinguible da natural, e todo iso cunha carga computacional moi reducida. O problema reside en que hai moitos parámetros que axustar (uns corenta) e a análise automatizada do sinal de voz non proporciona estimacións suficientemente fiables deses parámetros. Non obstante, os sintetizadores de formantes foron a técnica de síntese empregada nos primeiros sistemas de conversión texto-voz comercial e aínda hoxe continúan a ser empregados en múltiples aplicacións informáticas.

O Voder, presentado por Peter Dudley na Feira Mundial de Nova York en 1939, pode considerarse o precursor dos sintetizadores de formantes, aínda que neste caso o control dos parámetros do sintetizador facíase como se fora un instrumento musical: un teclado controlaba os formantes do tracto vocal, un pedal controlaba a frecuencia fundamental, etc). Precisábase, por tanto, dun experto ‘organista’ para xerar o sinal sintético.

O tamaño dos sintetizadores de formantes evolucionou ao longo dos anos. Nos anos 50 estes sintetizadores tiñan o tamaño dun armario, nos 80 o

dun pequeno periférico que se conectaba a un ordenador persoal e, na actualidade, son simplemente un programa informático.

3.1.2. Síntese articulatoria

Este tipo de sintetizador trata de modelar os fluxos de aire e movementos dos órganos articulatorios que interveñen no proceso de fonación. Non se trata, porén, de dispositivos mecánicos, senón de sistemas electrónicos ou programas informáticos equivalentes obtidos a partir da análise de datos articulatorios. Basicamente modelan o tracto vocal como unha serie de tubos de distinta sección, concatenados, que provocan distintas reflexións no sinal acústico que se propaga ó seu través (figura 2). A sección destes tubos obtense da medida da abertura de múltiples seccións transversais dun tracto vocal real.

Figura 2: *Modelo articulatorio*

Os principais inconvenientes dos modelos articulatorios residen na súa complexidade, aínda considerando simplemente modelos bidimensionais, e na dificultade de obter medidas articulatorias reais. Esta dificultade débese non só ó equipamento específico (palatógrafo, laringógrafo, resonancia magnética, raios X, etc.) para obter estas medidas, senón tamén a que algunhas destas técnicas son invasivas. Por este motivo, durante os últimos anos tomou forza a estimación da posición dos órganos articulatorios a partir da análise do sinal de voz. Aínda que esta correspondencia non é unívoca (hai distintas configuracións do tracto vocal que poden resultar nun mesmo sinal), a introdución de restricións de continuidade no movemento dos órganos articulatorios pode reducir notablemente a complexidade do problema.

Polo momento, a calidade dos sintetizadores articulatorios aínda dista moito da obtida cos sintetizadores de formantes e, sobre todo, da obtida coa técnica de moda: a concatenación de unidades.

Figura 3: Exemplo de difonema [os]. As frechas indican o inicio e o final da unidade, así como a fronteira entre os dous sons. Pódese observar a estacionariedade das características temporais e espectrais nos límites do difonema.

3.1.3. Síntese por concatenación de unidades

Estes sistemas xeran o sinal de voz sintético mediante a concatenación dun conxunto de unidades pregravadas. Á hora de considerar o tipo de unidade hai que ter en conta o número total de unidades requiridas e a calidade da voz sintética obtida. Aínda que nun principio o fonema poida parecer o tipo de unidade máis apropiada, debido o seu reducido número (arredor de 30 ou 40, dependendo do idioma), para que a fala sintética resulte intelixible é preciso modelar adecuadamente o fenómeno da coarticulación. Por este motivo o tipo de unidade máis empregado ó comezo deste tipo de sistemas foi o difonema, que contén a metade de dous sons consecutivos incluíndo a transición entre eles. O número teórico de difonemas – tendo en conta que existen combinacións de sons imposibles- é o cadrado do número de fonemas (arredor de 900 ou 1600). Tal e como se amosa na figura 3 os instantes de inicio e final dun difonema establécense na parte estacionaria do sinal, onde o efecto da coarticulación é menor. Implicitamente, estase asumindo que o efecto da coarticulación non se estende máis alá da metade do fonema seguinte, o que resulta inapropiado para certas secuencias de sons. Nestes casos pódese recorrer a trifenemas, unidade que consta de tres sons consecutivos, tendo só a metade do primeiro e do último. Dado que o

número de trifonemas posibles poden ser decenas de miles de unidades, estas unidades acostuman empregarse de forma combinada cos difonemas.

O feito de utilizar unidades pregravadas, que implicitamente xa teñen unha información prosódica asociada, fai necesario o emprego de algoritmos de modificación prosódica que axusten a información prosódica orixinal (frecuencia fundamental, duracións segmentais e enerxía) ós valores determinados polo modelado prosódico do conversor texto-voz. Existen distintas técnicas de modificación prosódica, cada unha coas súas vantaxes e inconvenientes. Non obstante, a técnica denominada TD-PSOLA (Time-Domain Pitch-Synchronous Overlap-Add) está entre as que proporciona mellor calidade e, ademais, cun custo computacional moi reducido.

Por último, a calidade dos sistemas de síntese por concatenación depende obviamente das representativas que sexan as unidades pregravadas. Conforme foi aumentando a capacidade de almacenamento e cálculo dos ordenadores, así como a redución do seu custo, foi posible considerar distintas realizacións de cada unidade co conseguinte aumento da calidade final. Chégase así ós sistemas de moda na actualidade, baseados en técnicas de selección de unidades, onde poden existir miles de instancias dunha única unidade, e onde, como veremos posteriormente, en lugar de falar da lonxitude da unidade empregada é máis apropiado falar da súa lonxitude mínima.

4. MODELADO PROSÓDICO

4.1. *Modelado da entoación*

Unha posible clasificación dos modelos de entoación é:

Modelos fonolóxicos: empregan una representación abstracta da entoación sen preocuparse dos valores concretos da frecuencia fundamental. Modelos fonéticos: céntranse desde un principio nos valores concretos do contorno de entoación.

Todos eles precisan, porén, dunha asociación entre eventos entoativos e a información lingüística extraída a partir da análise do texto. No noso sistema, Cotovía¹, esta asociación está baseada en traballos doutros autores que se fixeron para o castelán (López, 1993; Garrido, 1996). Estas investigacións establecen unha estrutura xerárquica dentro da frase. Segundo este esquema unha frase está composta por distintos grupos fónicos (delimitados por pau-

1. <http://www.gts.tsc.uvigo.es/cotovia/>

sas no discurso); asúmese normalmente que os grupos fónicos coinciden cos grupos de entoación. Cada un deste grupos de entoación esta composto á súa vez por unha secuencia de grupos acentuais, modelando cada un deles o contorno de entoación na proximidade dunha sílaba tónica. Frecuentemente o grupo acentual considérase formado por unha serie de palabras átonas que acaba nunha palabra tónica.

Atendendo á súa posición no grupo de entoación, os grupos acentuais poden clasificarse como: inicial, intermedio, final e inicial-final. Este último caso dáse cando o grupo de entoación ten un único grupo acentual. Se ademais temos en conta a posición da sílaba acentuada, falaremos de grupos acentuais oxítonos (acento na última sílaba), paroxítonos (na penúltima) e proparoxítonos (na antepenúltima). Os poucos casos nos que o acento recae nunha sílaba anterior á antepenúltima adoitan considerarse dentro dos proparoxítonos.

Figura 4: Exemplo de etiquetado co modelo ToBI no programa Praat (Boersma & Weenink 2005) (datos de <http://www1.cs.columbia.edu/~agus/tobi/tobi-praat>).

4.1.1. Modelos fonolóxicos

Dentro deste tipo de esquemas debemos citar o modelo ToBI (Tone and Break Indices) (Silverman et al., 1992), ilustrado na figura 4, que considera que o contorno de entoación está composto por unha serie de tons altos (etiquetados como H) e baixos (L). Estes tons (H e L) poden verse afectados polas sílabas tónicas ou *pitch accents* (H* ou L*), pola entoación base da frase ou phrase accent (H- ou L-) ou por estar situados na parte final dun gru-

po de entoación, *boundary tone*, (H% ou L%). Esta representación do contorno de entoación está collida directamente de Pierrehumbert (1980), onde se chega a presentar unha gramática da entoación para o inglés.

Ademais do etiquetado do contorno de entoación, o modelo ToBI emprega unha secuencia de índices de ruptura que indican o grao de ligazón entre palabras consecutivas. Estes índices toman un valor enteiro entre 0 e 4, correspondendo o catro á ruptura máis notoria.

Para a implementación dun modelo como o ToBI nun conversor texto-voz é necesario predicir a secuencia de tons e de índices de ruptura a partir do texto de entrada, ademais de transformar esa secuencia abstracta de tons en valores concretos da frecuencia fundamental. Por iso, neste tipo de sistemas é preciso dispoñer dun corpus de frases debidamente etiquetado que, a partir da súa análise, permita inferir o coñecemento necesario para levar a cabo esta tarefa. Xa mesmo a fase de etiquetado deste corpus non é doada e require a participación de persoal especificamente adestrado.

4.1.2. Modelos fonéticos

Céntranse dende un primeiro momento nos valores concretos de frecuencia fundamental. Estes valores pódense obter a partir dunha representación paramétrica intermedia ou dunha serie de datos da entoación directamente almacenados. Cabe falar entón de dous tipos de modelos: paramétricos e non paramétricos.

Dentro dos modelos paramétricos, o modelo de Fujisaki (Fujisaki & Hirose, 1984) é sen dúbida o máis coñecido. Trátase dun modelo superposicional onde se xera unha compoñente de frase e unha compoñente acentual. A compoñente de frase xunto cunha frecuencia fundamental base representa a tendencia global do contorno de entoación, mentres que a compoñente acentual modela as variacións locais (debidadas, por exemplo, á presenza de sílabas tónicas). Ambas compoñentes xéranse a partir dunha serie de parámetros obtidos da análise dun corpus para a entoación mediante un procedemento de análise e resíntese. Por suposto, para a aplicación deste modelo nun conversor texto-voz é preciso relacionar os distintos parámetros coa información lingüística, como se fai en Möbius, Pätzold, P. & Hess (1993) para o alemán.

4.2. Modelado das duracións segmentais

Nos primeiros sistemas de síntese de voz a duración dos sons era modelada normalmente mediante unha serie de regras que tiñan en conta o tipo

de son e a información do seu contexto. Porén, hoxe en día, a dispoñibilidade de corpora debidamente etiquetados permite o emprego de técnicas de predición estatística: regresión liñal, árbores de regresión, redes neuronais, etc. Non obstante, o tipo de preditores empregados son similares ás variables consideradas nos sistemas baseados en regras (tipo de son, sons que o rodean, posición na sílaba, tonicidade, etc.). En Santen (1998) atópase unha boa introdución ó tema do modelado das duracións segmentais mediante técnicas estatísticas.

5. SÍNTESE MEDIANTE SELECCIÓN DE UNIDADES

As técnicas de síntese baseadas na selección de unidades, tamén denominadas técnicas de síntese baseadas en corpora, xorden dun intento de paliar as limitacións dos primeiros sistemas de concatenación de unidades. A diferenza destes últimos onde normalmente só se consideraba unha realización por cada unidade acústica (tendo normalmente entre mil e dúas mil unidades), os sistemas baseados na selección de unidades propoñen o almacenamento de múltiples realizacións para cada unidade. Un algoritmo de programación dinámica será finalmente o encargado de seleccionar a realización máis axeitada de acordo cunhas funcións de custo (Hunt & A. Black, 1996) que teñen en conta distintos aspectos: tonicidade, duración da unidade, frecuencia fundamental, continuidade espectral entre unidades consecutivas, etc. Evidentemente, este tipo de sistemas precisa dun amplo corpus de fala pregravado e axeitadamente etiquetado, e o seu uso foi posible grazas ó rápido aumento da capacidade de cálculo dos ordenadores e do abarataemento dos elementos de memoria.

5.1. Selección de unidades acústicas

Os sintetizadores mediante selección de unidades proporcionan unha voz sintética moito máis intelixible e natural que os primeiros sistemas de concatenación. Neles, ademais, a determinación da lonxitude da unidade acústica que se debe considerar (fonemas, difonemas, trifonemas...) perde protagonismo, debido a que moitas veces se seleccionan secuencias de unidades que son consecutivas na gravación orixinal.

No caso do noso conversor texto-voz (Cotovía) a unidade básica empregada é o semifonema, aínda que normalmente o mecanismo de selección logra escoller secuencias de unidades acústicas consecutivas compostas por varios semifonemas, ademais de penalizar a concatenación de semifonemas

non consecutivos en zonas de transición entre sons. Isto faise de acordo coas funcións de custo de obxectivo e de concatenación, similares ás propostas en Hunt & A. Black (1996). O obxectivo da primeira destas funcións é preseleccionar unha serie de unidades acústicas candidatas que cumpran unha serie de requirimentos (tonicidade, duración, frecuencia fundamental, etc.), mentres que a segunda cualifica a concatenación de tódalas posibles secuencias de unidades candidatas de acordo con criterios de continuidade nas características do sinal de voz, tratando de evitar, por exemplo, discontinuidades espectrais. A busca da secuencia de unidades que minimiza a función de custo total (suma ponderada das dos dous anteriores) determínase mediante o algoritmo de Viterbi, empregándose tamén técnicas de poda que descartan, por exemplo, aquelas secuencias parciais de unidades que xa levan acumulado un custo demasiado alto.

Coas técnicas de poda a carga computacional dos métodos de selección de unidades redúcese significativamente, pero aínda así o número de combinacións de unidades pode resultar excesivo. Por iso é frecuente formar grupos de unidades acústicas que comparten unha serie de características (tipo de fonema, contexto do fonema, tonicidade, tipo de proposición, ...) e, de acordo coa análise lingüística do texto de entrada, considerar só as unidades candidatas dos grupos apropiados.

Unha vantaxe adicional dos métodos de selección de unidades é que, debido ó gran número de realizacións dunha mesma unidade que consideran, moitas veces atopan unha unidade coa frecuencia fundamental e duración moi similares ós valores pretendidos, facendo innecesaria a manipulación prosódica destas unidades (que como calquera manipulación introduce certa distorsión).

Como datos orientativos do número de realizacións empregados por un sintetizador mediante selección de unidades, en Cotovía, que emprega un corpus de fala duns noventa minutos, temos arredor de 120.000 realizacións de semifonemas, existindo para os máis frecuentes centenaes de realizacións (máis de mil nalgún caso).

5.2. Selección de unidades de entoación

O modelado da entoación inicial de Cotovía estaba baseado na concatenación dun conxunto reducido de patróns de entoación característicos dos distintos grupos acentuais. Non obstante, hai uns anos evolucionou a un modelo de entoación baseado tamén en técnicas de selección de unidades, que propuxemos en Campillo & Banga (2002) e que con sucesivas melloras deu lugar ó modelo actual (Campillo 2005; Campillo & Banga 2006). A idea

inicial é sinxela: a consideración de múltiples realización dos patróns de entoación asignados a cada tipo de grupo acentual; posteriormente o algoritmo de selección decantase por aquela secuencia de patróns de entoación cun menor valor dunha función de custo. Con este obxecto foi necesario definir patróns de entoación tendo en conta a clasificación dos grupos acentuais.

A novidade máis importante do noso modelo de entoación reside en que considera a posibilidade de xerar varios contornos de entoación e condiciona a selección do contorno de entoación final segundo a existencia dunha secuencia de unidades apropiada. Desta forma, evítanse modificacións prosódicas innecesarias e ó mesmo tempo introdúcese unha maior variabilidade nos contornos de entoación, xa que o contorno final vai depender tamén do mecanismo de selección das unidades acústicas.

6. OUTRAS LIÑAS DE INVESTIGACIÓN

Evidentemente, calquera dos temas de traballo descritos neste artigo é aínda unha liña de investigación aberta. Non obstante, neste parágrafo comentarei brevemente algunhas liñas de investigación no ámbito da síntese de voz que hoxe en día están de actualidade e que non son tan evidentes. Concretamente:

Medidas de distancia obxectivas altamente correlacionadas coa percepción. No campo da síntese de voz é moi habitual empregar medidas de distancia que indiquen, por exemplo, a similitude entre dúas realizacións de unidades de voz entre dúas realizacións de unidades de voz, tal e como se fai nas funcións de custo das técnicas de selección de unidades. Por outra banda, pódense empregar técnicas de medida obxectivas para medir a distorsión introducida polos algoritmos de modificación prosódica. O problema reside en que non sabemos *a priori* o impacto que vai ter na percepción unha mellora nestas medidas obxectivas, polo que para avaliar calquera mellora será necesario recorrer a probas subxectivas, custosas en tempo e diñeiro. Aínda que se están realizando traballos neste camiño (Wouters & Macon 1998; Donovan 2001), polo momento non hai resultados realmente satisfactorios.

Transformación de voces. Na actualidade empezan a xurdir os primeiros sistemas experimentais de tradución de voz a voz que permiten a comunicación entre dúas persoas que falan linguas distintas. Basicamente o proceso é o seguinte: un recoñecedor de fala transforma o que di un locutor nunha lingua A nun texto que é traducido mediante un tradutor

de texto e que despois é sintetizado na lingua B do seu interlocutor mediante un conversor texto-voz. En principio, a mensaxe na lingua B terá as características da voz sintética do conversor, no entanto sería interesante modificar as súas características de forma que soara coma se o locutor orixinal falase na lingua B. Nesta aplicación reside boa parte da importancia da transformación de voces, que é unha liña de investigación que nos últimos anos está comezando a producir resultados realmente interesantes (Duxans 2006; Ye & Young 2006).

Síntesis de voz mediante HMMs. Foi proposto recentemente por Donovan (1996) e Tokuda, Zen, & A. W. Black (2002), que empregan unha parametrización dun corpus de fala para adestrar unha serie de modelos ocultos de Markov (HMMs) dependentes do contexto. Unha vez adestrados, estes modelos xeran, a partir da información obtida da análise do texto, unha secuencia de parámetros de síntese cos que se xera o sinal sintético.

AGRADECEMENTOS

Os traballos que en síntese de voz se están a desenvolver no grupo de investigación ó que pertence o autor están parcialmente financiados pola Xunta de Galicia, o goberno español e fondos FEDER no marco dos proxectos PGI0IT05TIC32202PR e TEC2006-13694-C03-03. No desenvolvemento do conversor texto-voz *Cotovía*, participan investigadores das universidades de Vigo e Santiago e o Centro Ramón Piñeiro para a Investigación en Humanidades.

REFERENCIAS BIBLIOGRÁFICAS

- Campillo, F. & E. R. Banga (2002): "Combined prosody and unit selections for corpus-based text-to-speech systems", in *Proceedings of ICSLP* 1, 141-144.
- Boersma, P., & D. Weenink, D. (2005): *Praat: doing phonetics by computer* (Version 4.3.14) [Computer program]. <http://www.fon.hum.uva.nl/praat/>
- Campillo, F., & E. R. Banga, E. R. (2006): "A method for combining intonation modelling and speech unit-selection in corpus-based speech synthesis systems", *Speech Communication* 48, 941-956.
- Campillo, F., & E. R. Banga, E. R. (2002): "Combined prosody and unit selections for Corpus-based text-to-speech systems", in *Proceedings of ICSLP* 1, 141-144.
-

- Campillo, F. (2005): *Síntesis de voz basada en selección de unidades acústicas y prosódicas*. Tese doutoral de doutoramento. ETSI Telecomunicación, Universidade de Vigo.
- Donovan, R. (1996): *Trainable sSpeech sSynthesis*. PhD thesis Tese de doutoramento. Cambridge University.
- Donovan, R. (2001): "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers", in *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, nº 123.
- Duxans, H. (2006): *Voice conversion applied to text-to-speech systems*. Tese de doutoramento. Universitat Politècnica de Catalunya.
- Fujisaki, H., & K. Hirose (1984): "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *Journal of Acoustic Society of Japan* 54, 233-242.
- Garrido, J. M. (1996): *Modelling Spanish iIntonation for tText-to-sSpeech aApplications*. Tese de doutoramento. Universitat Autònoma de Barcelona.
- Hunt, A., & A. Black (1996): "Unit selection in a concatenative speech synthesis system using a large speech database", in *Proceedings of ICASSP'96*, vol. 1 1, 373-376.
- Klatt, D. (1980): "Software for a cascade/parallel formant synthesizer", *Journal of the Acoustical Society of America* 67, 971-995.
- López, E. (1993): *Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto-voz en español basados en concatenación de unidades*. Tese de doutoramento. Universidad Politécnica de Madrid.
- Möbius, B., M. Pätzold & W. Hess (1993): "Analysis and synthesis of German F0 contours by means of Fujisaki's model", *Speech Communication* 13, 53-61.
- Pierrehumbert, J. (1980): *The phonology and phonetics of English intonation*. PhD thesis Tese de doutoramento. MIT.
- Santen, J. van (1998): "Timing", in R. Sproat (ed.): *Multilingual text-to-speech synthesis: the Bell Labs approach*. Dordrecht: Kluwer, 115-140.
- Silverman, K., M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert & J. Hirschberg (1992): "ToBbi: a standard for labelling English prosody", in *Proceedings of ICSLP 2*, 867-870.
- Tokuda, K., H. Zen, & A. W. Black (2002): "An HMM-based speech synthesis system applied to English", in *Proceedings of IEEE Speech Synthesis Workshop*, 227-230.
- Wouters, J., & M. Macon (1998): "A perceptual evaluation of distance measures for concatenative speech synthesis", in *Proceedings ICSLP 1998*, vol. 6, 2747-2750.
- Ye, H., & S. Young (2006): "Quality-enhanced voice morphing using maximum likelihood transformations", in *IEEE Transactions on Audio, Speech and Language Processing* 14(4), 1301-1312.