Nature Identical Prosody – data-driven prosodic feature assignment for diphone synthesis

Peter Juel Henrichsen
Danish Center for Applied Speech Technology (DanCAST)
Copenhagen Business School, Denmark
pjh.ibc@cbs.dk

Abstract

Today's synthetic voices are largely based on *diphone synthesis* (DiSyn) and *unit selection synthesis* (UnitSyn). In most DiSyn systems, prosodic envelopes are generated with formal models while UnitSyn systems refer to extensive, highly indexed sound databases. Each approach has its drawbacks; such as low naturalness (DiSyn) and dependence on huge amounts of background data (UnitSyn). We present a hybrid model based on high-level speech data. As preliminary tests show, prosodic models combining DiSyn style at the phone level with UnitSyn style at the supra-segmental levels may approach UnitSyn quality on a DiSyn footprint. Our test data are Danish, but our algorithm is language neutral.

1. Introduction

We outline a new method for improving the prosodic quality of artificial voices based on concatenative synthesis, inheriting the perceived naturalness of the massively data-demanding unit selection synthesis (UnitSyn) while maintaining the rational design of the conceptually simpler diphone synthesis (DiSyn).

The DiSyn engine is based on a sound database of a highly systematic design. The database can be described as a matrix *P*x*P*, where *P* is the phone inventory of the target language *T*. Each cell in the matrix is inhabited by a sound file representing a diphone (excluding those never occurring in *T*). Synthesis, then, amounts to diphone-splicing and post-processing. Due to the rational layout of the database, the footprint of the DiSyn system is moderate.

In the UnitSyn engine, in contrast, parsimony is traded for naturalness by including (huge amounts of) samples of connected speech in the database. Input text to the UnitSyn system with exact matches in the database are reproduced flawlessly (resembling playback rather than resynthesis), naturalness declining gracefully with the distance between input and best database match. In practical use UnitSyn systems tend to fluctuate between playback quality (very high) and sub-DiSyn quality (poor). In contrast, DiSyn systems deliver a moderate, but far more consistent quality.

	DiSyn (diphone)	**UnitSys (unit selection)**	**NIP (hybrid)**
Database preparation	*limited*	*labour-intensive*	*moderate*
Footprint	*moderate*	*very large*	*moderate*
Naturalness	*low*	*medium*	*medium*
Consistency	*high*	*low*	*high*

Table 1. Prosodic models (concatenative synthesis)

Our NIP algorithm (Nature Identical Prosody) combines the compact design of the DiSyn database with the data-driven prosodic plasticity of the UnitSyn. NIP can be applied in existing DiSyn systems, in contrast to other recently suggested hybrid synthesis systems (e.g. Oparin 2008, Aylett 2008, Guner 2011).

We first introduce Grønnum's prosodic model for the Danish sentence as well as our data-driven alternative; then we report on an experiment showing that a DiSyn-style algorithm informed by speech data may approach the UnitSyn prosodic quality.

2. Theory-driven prosody assignment

Following Grønnum (1978, 1985, 1992, 1998), the Danish stress group (SG) consists of one or more syllables. The rules of prosody assignment are:

I. an initial stressed syllable (all others unstressed),
II. from I, an F0 upstep to the 2nd syllable,
III. from II, a general (possibly linear) F0 fall,
IV. an optional final F0 upstep to the following SG

Henrichsen (2006) suggests this formalization:

i. $F_m = F_0 - \frac{m}{m_{TOTAL}}\left(F_0 - F_{m_{TOTAL}}\right)$

ii. $F'_{m,u'} = F_m + UP_{m,u'}$

iii. $F_{m,u',u} = F'_{m,u'} - \frac{u-1}{1-u}\left(F'_{m,u'} - F''_{m,u'}\right)$

iv. $F''_{m,u'} = F_{m+1} - UP_{m+1,u'}$

F_m is the fundamental frequency for the (full vowel of the) initial syllable of the *m*th SG; u' is the number of unstressed syllables in the *m*th SG; $F'_{m,u'}(F''_{m,u'})$ is F0 for the first (last) unstressed syllable in the *m*th SG; defined for $u'>1$ ($u'>2$); $F_{m,u',u}$ is F0 for the last unstressed syllable of the *m*th SG; defined for $u'>3$ and $u>1$. The arbitrary constants F_0, m_{TOTAL} and $F_{m_{TOTAL}}$ are all associated with linguistics properties; F_0 and $F_{m_{TOTAL}}$ are the upper and lower bound of the speaker's normal F0 range (possibly, but not necessarily a function of the sentence length too; Grønnum is not very specific here); m_{TOTAL} is the total number of SGs. The upstep function *UP* is introduced in the full papers.

3. Data-driven prosody assignment

NIP prosody assignment is based on pattern matching in a background corpus of read-aloud texts. The corpus does not include the actual sound files, but selected

annotation tiers only (using the Danish PAROLE corpus, Henrichsen 2007). One speech second is thus represented by 10 8-bit numbers or so, as opposed to the 48,000 16-bit sound samples typical of UnitSyn - a data reduction of four orders of magnitude.

What data types are necessary and sufficient for reliable pattern matching? Based on pilot experiments, we settled on tiers A1 (acoustic) and L1-L5 (linguistic).

A1. Fundamental frequency (logarithmic measures)

L1. Orthographic form (dictionary approved)
L2. Phonetic form
L3. Stress pattern (stressed=2, 2ndary=1, unstr.=0)
L4. Part-of-Speech (PAROLE-style tags)
L5. Word freq. (in a 28M corpus of balanced texts)

For Danish, L1 and L5 together provide almost 100% lexical disambiguation. L2 and L3, in contrast, may vary considerably with the syntactic and semantic context. L5 was included experimentally, assuming that high-frequency tokens are more likely to appear de-stressed or time condensed than low-frequency words, grouping words otherwise unrelated in L1-L4.

3.1 The NIP algorithm presented by an example

Consider an input string *I* "du vil gerne op til slottet" (*you'd like to get (up) to the Castle*). *I* is analysed (automatically) in the dimensions L2-L5.

$L1_I$	du	vil	gerne	op	til	slottet
$L2_I$	[du]	[ve]	[gáRn0]	[Cb]	[te]	[slCd-D]
$L3_I$	2	1	2-0	2	1	2-0
$L4_I$	PRO	AUX	ADJ	ADJ	PREP	$CN_{SG,DEF}$
$L5_I$	-7.6	-5.7	-8.0	-6.3	-4.1	-11.6

Using $L1_I$-$L5_I$ as a search expression, a matching utterance *U* is identified in the NIP database:

$L1_U$	han	har	ikke	noget	imod	indvandrere
$L2_U$	[han]	[hA]	[eg0]	[nc0D]	[imoD?]	[envAndCC]
$L3_U$	1	**1**	**2-0**	**2**-0	0-**1**	**2**-1-0-0
$L4_U$	**PRO**	**AUX**	**ADJ**	PRO	**PREP**	$\mathbf{CN}_{PL,-DEF}$
$L5_U$	**-5.1**	**-4.5**	-4.6	**-6.6**	-8.3	**-9.9**

Observe that, in the sound related tiers L1 and L2, *I* and *U* are unrelated; tiers L3-L5, however, show a distinct similarity (values shown in **bold**).

3.2 The NIP algorithm, summarized

Quantifying over input windows and database windows (both up to 7-place), a prosodic envelope is distilled by superimposing all envelope contributions ('envelope' = one F0 data point for each syllable) weighted by the corresponding GP^7 value. The resulting envelope is normalized wrt. permitted F0 range, duration, etc.

GP Geometrical proximity

Tier values: $GP_{TIER}(w,w')$ for $TIER = L1_x \,..\, L5_x$

Tier vectors: $GP^{VEC}(V, V') = (\sum_{i \in TIER} GP_i\,(x_i\,, x_i\,')) \,/5$

Windows: $GP^7(W,W') = (\sum_{n=1..7} GP^{VEC}(V_{W,n}\,, V_{W',n}))/7$

4. Experimental evidence

16 Danish test subjects graded a suite of test sentences varied systematically for length (2-8 SGs), synthesized with the DiSyn voice Gizmo (developed at DanCAST with the festival toolkit) using prosodic models m1-m5.

m1. Grønnum's model formalized as in 2
m2. Model of DiSyn voice Carsten (www.mikrov.dk)
m3. Model of UniSyn voice Sara (www.pdc.dk)
m4. NIP based model as presented in this paper
m5. Human read-aloud version re-synthesized

The test subjects were asked to evaluate the five instancies of each sentence for naturalness: "Order the versions from best to worst" and "Grade each version as excellent/good/mediocre/bad"

As expected, all subjects preferred m5 over all other models, showing the test set-up to be reliable. Excluding m5 from the test set, these patterns emerged:

{m3,m4} were preferred over {m1,m2} by all 16 subjects, suggesting that current theory-driven models of Danish prosody are inferior to data-driven models.

13 subjects had m1>m2, suggesting our formalization of Grønnum's model to be superior to the one used in Carsten (the leading commercial DiSyn based synthetic voice for Danish).

9 subjects had m4>m3 (m4 being preferred for sentences containing several infrequent content words), suggesting that NIP-based prosodic models may offer attractive alternatives to the full-blown UnitSys system.

5. Conclusion

We do not claim NIP-driven diphone synthesis to be superior to Unit Selection *as such*. More reference data should still provide better synthesis everything else being equal. However, based on our experiments we suggest that the standard claim of huge sound databases as *necessary* remedies to the failing prosodic naturalness of diphone synthesis be reconsidered.

References

Aylett, M. P. & J. Yamagishi (2008) Combining Statistical Parametric Speech Synthesis and Unit-Selection for Automatic Voice Cloning; LangTech-2008, Rome.

Grønnum, N. (1998). Intonation in Danish. In D.Hirst et al (eds) Intonation Systems. Cambridge Univ. Press.

Guner, E.; Cenk Demiroglu (2011) A Small-footprint Hybrid Statistical and Unit Selection TTS Synthesis System for Turkish; Computer and Information Sciences vol. II; Springer

Henrichsen, P.J. (2006) Danish Prosody, Formalized. In J.Toivanen et al (2006)

Henrichsen, P.J. (2007) The Danish PAROLE Corpus - a Merge of Speech and Writing; in J. Toivanen et al (2007)

Oparin, I.; V.Kiselev; A.Talanov (2008) Large Scale Russian Hybrid Unit Selection TTS. SLTC-08. Stockholm.

Toivanen, J.; P. J. Henrichsen (eds) (2006/2007) Current Trends in Research on Spoken Language in the Nordic Countries (vol 1/2). Oulu Univ. Press.