

SEGMENTAL RE-SYNTHESIS OF CHILD SPEECH USING UNIT SELECTION

Sofia Strömbergsson

Department of Speech, Music and Hearing, School of Computer Science and Communication,
Royal Institute of Technology (KTH), Stockholm, Sweden
sostr@csc.kth.se

ABSTRACT

This report describes a novel approach to segmental re-synthesis of child speech, by concatenation of speech from different speakers. The re-synthesis builds upon standard methods of unit selection, but instead of using speech from only one speaker, target segments are selected from a speech database of many child speakers. Results from a listener evaluation suggest that the method can be used to generate intelligible speech that is difficult to distinguish from original recordings.

Keywords: speech synthesis, unit selection, child speech, listening evaluation

1. INTRODUCTION

Modified re-synthesis of recorded speech can be used for many different purposes. For example, modification through linear predictive coding (LPC) parameter manipulation has been used to create ambiguous realizations of speech sounds for experiments of categorical perception, and to create extrapolated realizations of speech sounds to exaggerate the difference between two phonemes [4]. A similar technique has also been used to generate “corrected” versions of children’s deviant /r/ productions [6]. Here, children and adolescents listened to recordings of themselves and of other children, where half of the words were incorrectly produced, and the other half “corrected” by LPC parameter modification. For each word, the children judged the correctness of the /r/ and the identity of the speaker. The LPC manipulations were done manually, and therefore there was a time span of 1-2 weeks between recording of the children and the listening task, and this gap may partly explain the difficulties the children had recognizing their own recordings. Thus, in order to fully understand how children (or adults) react to hearing corrected versions of their recorded incorrect speech, corrective re-synthesis should preferably be done in real-time.

A method for re-synthesizing segmentally modified versions of recorded speech could be valuable not only in speech and language intervention for children with deviant speech, but could also be used e.g. in second language learning. However, for these purposes, the technique must not only allow modification and re-synthesis of recorded speech in real-time, but the generated speech must also be conceivable realizations of speech that might have been produced by the recorded speaker himself/herself. In earlier studies involving re-synthesized speech, characteristics other than naturalness and preservation of perceived speaker identity have been prioritized, e.g. controlling intonation and syllabic rhythm [5] and increasing intelligibility of poorly articulated speech [2]. In these studies, re-synthesis was done through diphone synthesis and formant synthesis, respectively. If naturalness and preservation of speaker identity are prioritized, however, other synthesis methods are better suited.

The present study describes a novel approach to modified re-synthesis of phonemic segments, by standard methods of unit selection and concatenative synthesis, but where the concatenated speech segments come from different speakers. The purpose of the study was to find out if this re-synthesis method can be used to generate natural and comprehensible speech.

2. METHOD

2.1. Speech data

A corpus of recordings of 76 children was used as a speech database. The recording script used for all children consisted of 19 one word utterances (see Appendix).

The recorded children were 4 to 9 years old. 62 of the children had normal speech and 14 of the children were diagnosed with phonological impairment (PI), and had deviant production of either /k/ or /t/, which were often produced as [t] or [k], respectively.

The recordings were made at different schools and pre-schools, and always took place in a separate room with limited noise. All recordings were made by a Sennheiser m@b 40 headset, using a 16-kHz sampling rate and 16-bit resolution. The total number of utterances in the speech corpus was 1444 utterances. 132 of these were produced by the children with PI and judged by the author as having a deviant initial plosive.

2.2. Preparation of speech data

All data in the speech corpus was segmented and aligned with the HMM-based nAlign [7]. A concatenation position was defined at the middle of the first vowel in the utterance. At this position, three sets of acoustic features were extracted: F0, log power and MFCCs (with 13 coefficients).

2.3. Unit selection

The task for the unit selector was to find an initial segment u_{i-1} in the speech database that would best match a given remainder segment u_i . The concatenation cost C^c between these segments was calculated as follows (as in [1]):

$$(1) \quad C^c(u_{i-1}, u_i) = \sum_{j=1}^q w_j^c C_j^c(u_{i-1}, u_i)$$

Three sub-costs C_j were used in the unit selection (i.e. $q = 3$):

- Euclidean distance [8] in F0
- Euclidean distance in log power
- Mahalanobis distance [8] for the MFCCs

Different weights w_j were assigned to the different sub-costs. These weights were derived from a weight optimization procedure (described below). High penalties were given to combinations of segments where the vowel in u_{i-1} did not match the vowel in u_i , to avoid combinations of mismatching vowels. The segment u_{i-1} with the lowest concatenation cost was then selected from the speech corpus for concatenation with u_i .

2.4. Concatenation

Concatenation positions in u_i and u_{i-1} were adjusted to the zero-crossings closest to the middle of the vowel (within a range of 15 samples before or after), where the direction of the slope (negative or positive) was the same for u_i and u_{i-1} , to preserve wave continuation.

2.5. Weight optimization

15 original recordings were held out from the speech corpus and used as training material. To arrive at an optimal set of weights for the sub-costs for F0, log power and MFCC distance, a weight space search [1] was performed. Three different values were attempted for the different weights, in all possible combinations ($3^3 = 27$). For each weight set, the 15 training utterances were re-synthesized with the best fitting initial segment in the speech corpus. (The initial segments in the training material were never eligible for selection.) The Mel-Cepstral Distance (MCD; [3]) was used as an objective measure of the difference between synthesized utterances and training utterances. MCD was calculated frame-by-frame, as follows (for the frame k):

$$(2) \quad MCD(k) = \sqrt{\sum_{i=1}^{13} [MC_x(i, k) - MC_y(i, k)]^2}$$

where $MC_x(i, k)$ and $MC_y(i, k)$ are the i^{th} Mel-Cepstral coefficients of the vowel in the synthesized and the original (training) utterance, respectively. As the remainder part of the synthesized utterances is always identical to the remainder part of the training utterances, MCD was only calculated for the vowel part of the utterances. When the number of frames in the synthesized vowel and the training vowel was different, comparison was only performed up to the last frame of the shortest segment. An average MCD was calculated for all frames in the vowel to represent the difference between the synthesized utterance and the training utterance. The weight set that generated the set of synthesized utterances that were most similar to the set of training utterances was then selected as the optimal weight set.

2.6. Evaluation

A listening script of 60 stimuli was generated without supervision from the speech corpus, with the restriction that 20 stimuli were original recordings (10 normal and 10 deviant), and 40 stimuli were modified. Of the modified stimuli, 20 were modified normal recordings and 20 were modified deviant recordings. In all modified stimuli, the initial consonant was replaced by a different initial consonant (/t/ for /k/ and vice versa). Through this re-synthesis, the 20 deviant recordings were “corrected”, and the 20 normal recordings were “impaired”. An online listening test was constructed to first present 3 training

stimuli, and then the 60 stimuli in random order (different for different listeners). The task for the listeners was to report what they heard (by typing) and to judge whether the stimulus was an original recording or a modified recording. 52 adult listeners participated in the experiment.

3. RESULTS

As a measure of intelligibility of the stimuli, the listeners' identification accuracy of the initial consonants was used. (As half of the stimuli were produced with deviant speech, and as such, most often nonsense words, inconsistency in the listeners' spelling was expected. Therefore, only the initial consonant was regarded in the analysis.) Inter-rater agreement of the listeners' perception of the initial consonants was measured by Fleiss' kappa at 0.67. Figure 1 shows that original deviant consonants are ambiguous to the listeners, whereas for all other stimulus types, the identification accuracy is 80% or higher.

As a measure of naturalness, the listeners' judgment of whether a stimulus was an original or a modified recording was used. Here, inter-rater agreement was measured by Fleiss' kappa at 0.12. As shown in Table 1, 66% of the modified stimuli were perceived as original recordings. For

comparison, 21% of the original recordings were actually perceived as modified.

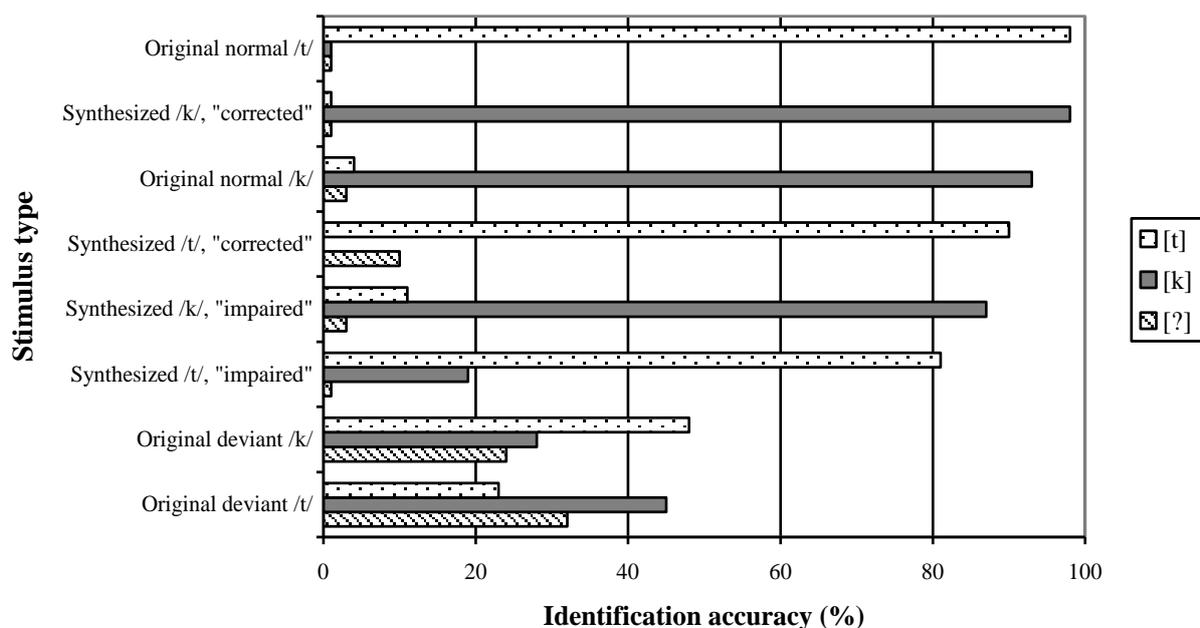
Table 1: Relative distribution of the listeners' perception of original and modified recordings.

Perceived	Actual	
	Original	Modified
Original	79%	66%
Modified	21%	34%

4. DISCUSSION

The finding that 80-98% of the modified stimuli are perceived correctly (by their initial consonant) indicates that the generated speech is intelligible. For the intended use with children with deviant speech, the difference in consonant identification accuracy between original deviant recordings and corrected re-synthesized recordings is perhaps the most interesting. Whereas the listeners' perception of the initial consonants in recordings of children's deviant speech is quite equivocal, corrected re-synthesized recordings are perceived as much less ambiguous. This suggests that correction by re-synthesis could be a way of generating unambiguous speech targets for children with deviant speech.

Figure 1: The listeners' identification accuracy of initial consonants, across the eight different types of stimuli, in descending order of accuracy. The different bars represent the three perceptual categories [t], [k] and [?] (any other sound), respectively.



As illustrated in Figure 1, the listeners' perception of deviant stimuli (original recordings or recordings "impaired" by re-synthesis) was generally less accurate than their perception of normal stimuli. Since around half of the listeners had not had any formal training in transcription, they might have had difficulties to refrain from interpreting what a child might have *intended* to say in typing what they heard. Taking this into account, the results for deviant stimuli might underestimate the actual comprehensibility, whereas the results for normal stimuli might overestimate the actual comprehensibility.

Most of the modified stimuli were perceived as original recordings. This finding, together with the low degree of agreement between the listeners (Fleiss' kappa at 0.12), are indications that it was difficult for the listeners to distinguish between original and modified stimuli. Thus, it is indeed possible to generate natural speech by concatenating speech from different speakers. But even if listeners were able to detect that an utterance has been modified, it is not clear whether this is really a problem in the intended use with children with deviant speech, assuming that the generated speech is still intelligible.

An additional conclusion that can be drawn from the results in this study is that quite good results can be achieved with recordings of sub-optimal quality. The recordings in this study were all done with rather simple recording equipment, and in naturalistic settings, rather than in a sound-proof studio. As these are conditions one could expect in a clinical setting, the technique (together with the speech corpus) could easily be implemented into a speech and language therapy tool. Such a tool would allow the generation of speech production targets that are tailor-made for each individual child, and for each individual utterance. Assumably, this would be a valuable resource in speech and language therapy.

5. ACKNOWLEDGEMENTS

This work was funded in part by the Foundation Promobilia. Many thanks to PhD Christoph Draxler for administering the web experiment.

6. REFERENCES

[1] Hunt, A.J., Black, A.W. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing ICASSP-96* Atlanta, GA, USA.

[2] Kain, A.B., Hosom, J-P., Niu, X., van Santen, J.P.H., Fried-Oken, M., Staehely, J. 2007. Improving the intelligibility of dysarthric speech. *Speech Communication* 49(9), 743-759.

[3] Kubichek, R.F. 1993. Mel-cepstral distance measure for objective speech quality assessment. *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 125-128.

[4] Protopapas, A. 1998. Modified LPC resynthesis for controlling speech stimulus discriminability. *136th Annual Meeting of the Acoustical Society of America* Norfolk, VA, USA.

[5] Ramus, F., Mehler, J. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America* 105(1), 512-521.

[6] Shuster, L. 1998. The Perception of Correctly and Incorrectly Produced /r/. *Journal of Speech Language and Hearing Research* 41(4), 941-950.

[7] Sjölander, K. 2003. An HMM-based system for automatic segmentation and alignment of speech. *Proc of Fonetik 2003* Umeå University, Sweden.

[8] Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge, UK.

7. APPENDIX

Recording script

	Word	Pronounced	In English
1	k	/ˈko:/	(the letter k)
2	kaka	/ˈkɑ:kɑ/	cake
3	kam	/ˈkɑm/	comb
4	karta	/ˈkɑ:tɑ/	map
5	katt	/ˈkɑt/	cat
6	kavel	/ˈkɑ:vəl/	rolling pin
7	kopp	/ˈkɔp/	cup
8	korg	/ˈkɔrj/	basket
9	kulle	/ˈkələ/	hill
10	kung	/ˈkøŋ/	king
11	tåg	/ˈto:g/	train
12	tak	/ˈtɑ:k/	roof
13	tant	/ˈtɑnt/	lady
14	tavla	/ˈtɑ:vla/	picture
15	tomte	/ˈtɔmtə/	Santa Claus
16	topp	/ˈtɔp/	top
17	tumme	/ˈtømə/	thumb
18	tunga	/ˈtøŋɑ/	tongue
19	tupp	/ˈtøp/	rooster