Corrective re-synthesis of deviant speech using unit selection

Sofia Strömbergsson
Dep. Of Speech, Music and Hearing,
KTH (Royal Institute of Technology),
Stockholm, Sweden
sostr@csc.kth.se

Abstract

This report describes a novel approach to modified re-synthesis, by concatenation of speech from different speakers. The system removes an initial voiceless plosive from one utterance, recorded from a child, and replaces it with another voiceless plosive selected from a database of recordings of other child speakers. Preliminary results from a listener evaluation are reported.

1 Background

Modified re-synthesis of recorded speech can be used for 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 (Protopapas, 1998). A similar technique has also been used to generate "corrected" versions of children's deviant /r/ productions (Shuster, 1998). 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. However, as manipulations were done manually, there was a time span of 1-2 weeks between recording of the children and listening, and this gap could 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 (Ramus & Mehler, 1999) and increasing intelligibility of poorly articulated speech (Kain et al, 2007). 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 74 children producing one word utterances was used as a speech database. The recording script used for all children contained 19 words, with 10 beginning with /kV/ and the other 9 beginning with /tV/ (see Appendix). The recorded children were 4 to 9 years old.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 214–217

60 of the children had normal speech and 14 of the children were diagnosed with phonological impairment (PI), and had problems with 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 1406 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 (Sjölander, 2003). 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 (13 Mel-Frequency Cepstrum 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 (Hunt & Black, 1996):

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

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

- Euclidean distance (Taylor, 2008) in F0
- Euclidean distance in log power
- Mahalanobis distance (Taylor, 2008) 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 as the optimal segment 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 (Hunt & Black, 1996) 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; Kubichek, 1993) 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):

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

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 automatically, 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. In all modified stimuli, the initial consonant was replaced by a different initial consonant (/t/ for /k/ and vice versa). Through this re-synthesis, 20 of the modified stimuli were "corrected" and 20 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. 38 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 0.64. Figure 1 shows that original deviant consonants are ambiguous to the listeners, whereas for all other stimulus types, the identification accuracy is around 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 0.13. As shown in Table 1, 62% of the synthesized stimuli were perceived as original recordings. For comparison, 24% of the original recordings were actually perceived as modified.

	Actual	
Perceived	Original	Modified
Original	76%	62%
Modified	24%	38%

Table 1. Relative distribution of the listener's perception of original and modified recordings.

4 Discussion

The finding that around 80-90% of the synthesized 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 original deviant recordings 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.

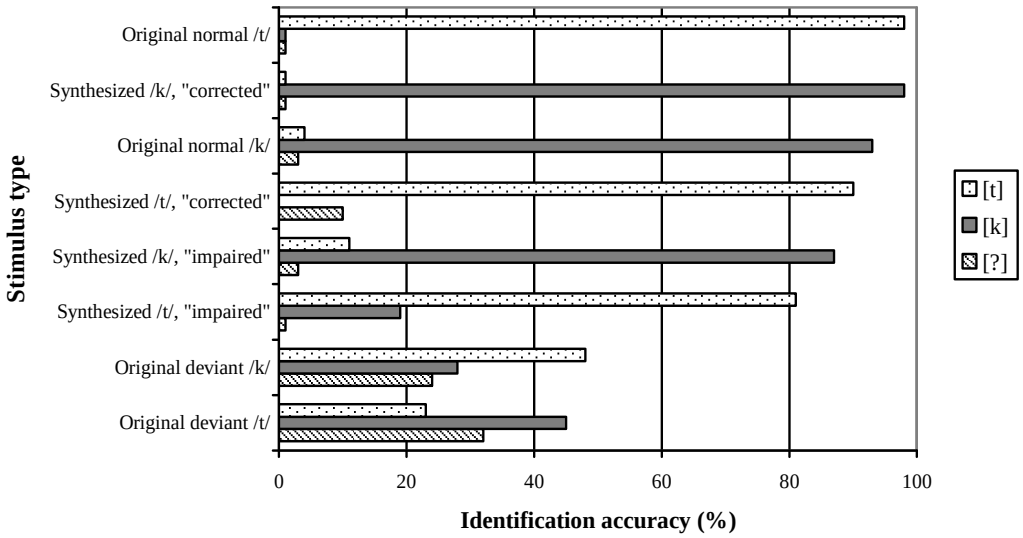

Figure 1. The listeners' identification accuracy of initial consonants, across the eight different types of stimuli, in descending order of accuracy. The different bars represent the three perceptual categories [t], [k] and [?] (any other sound), respectively.

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 0.13), 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.

Acknowledgements

The web experiment was implemented and administered by PhD Christoph Draxler at the Institute of Phonetics and Speech Processing in Munich, Germany.

References

Andrew J. Hunt & Alan W. Black. 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.

Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, X, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2007. Improving the intelligibility of dysarthric speech. *Speech Communication.* 49(9): 743-759.

Robert F. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*.

Athanassios Protopapas. 1998. Modified LPC re-synthesis for controlling speech stimulus discriminability. *136th Annual Meeting of the Acoustical Society of America*. Norfolk, VA, USA.

Franck Ramus and Jacques Mehler. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105(1): 512-521.

Linda I. Shuster. 1998. The Perception of Correctly and Incorrectly Produced /r/. *Journal of Speech Language and Hearing Research*, 41(4): 941-950.

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

Paul Taylor. 2009. *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge, UK.

Appendix A. Recording script

	Word	Pronounced	In English
1	k	/koː/	*(the letter k)*
2	kaka	/kɑːka/	*cake*
3	kam	/kam/	*comb*
4	karta	/kɑːʈa/	*map*
5	katt	/kat/	*cat*
6	kavel	/kɑːvəl/	*rolling pin*
7	kopp	/kɔp/	*cup*
8	korg	/kɔrj/	*basket*
9	kulle	/kʉlə/	*hill*
10	kung	/kʉŋ/	*king*
11	tåg	/toːg/	*train*
12	tak	/tɑːk/	*roof*
13	tant	/tant/	*lady*
14	tavla	/tɑːvla/	*picture*
15	tomte	/tɔmtə/	*Santa Claus*
16	topp	/tɔp/	*top*
17	tumme	/tʉmə/	*thumb*
18	tunga	/tʉŋa/	*tongue*
19	tupp	/tʉp/	*rooster*

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955