

Modified re-synthesis of initial voiceless plosives by concatenation of speech from different speakers

Sofia Strömbergsson

Department of Speech, Music and Hearing, School of Computer Science and Communication, KTH, Stockholm

Abstract

This paper describes a method of re-synthesising utterance-initial voiceless plosives, given an original utterance by one speaker and a speech database of utterances by many other speakers. The system removes an initial voiceless plosive from an utterance and replaces it with another voiceless plosive selected from the speech database. (For example, if the original utterance was /tat/, the re-synthesised utterance could be /k+at/.) In the method described, techniques used in general concatenative speech synthesis were applied in order to find those segments in the speech database that would yield the smoothest concatenation with the original segment. Results from a small listening test reveal that the concatenated samples are most often correctly identified, but that there is room for improvement on naturalness. Some routes to improvement are suggested.

Introduction

In normal as well as deviant phonological development in children, there is a close interaction between perception and production of speech. In order to change a deviant (non-adult) way of pronouncing a sound/syllable/word, the child must realise that his/her current production is somehow insufficient (Hewlett, 1992). There is evidence of a correlation between the amount of attention a child (or infant) pays to his/her own speech production, and the phonetic complexity in his/her speech production (Locke & Pearson, 1992). As expressed by these authors (p. 120): “the hearing of one’s own articulations clearly is important to the formation of a phonetic guidance system”.

Children with phonological disorders produce systematically deviant speech, due to an immature or deviant cognitive organisation of speech sounds. Examples of such systematic deviations might be stopping of fricatives, consonant cluster reductions and assimilations. Some of these children might well perceive phonological distinctions that they themselves

do not produce, while others have problems both in perceiving and producing a phonological distinction.

Based on the above, it seems reasonable to assume that enhanced feedback of one’s own speech might be particularly valuable to a child with phonological difficulties, in increasing his/her awareness of his/her own speech production. Hearing a re-synthesised (“corrected”) version of his/her own deviant speech production might be a valuable assistance to the child to gain this awareness. In an effort in this direction, Shuster (1998) manipulated (“corrected”) children’s deviant productions of /r/, and then let the subjects judge the correctness and speaker identity of speech samples played to them (which could be either original/incorrect or edited/corrected speech, spoken by themselves or another speaker). The results from this study showed that the children had most difficulties judging their own incorrect utterances accurately, but also that they had difficulties recognizing the speaker as themselves in their own “corrected” utterances. These results show that exercises of this type might lead to important insights to the nature of the phonological difficulties these children have, as well as providing implications for clinical intervention.

Applications of modified re-synthesis

Apart from the above mentioned study by Shuster (1998), where the author used linear predictive parameter modification/synthesis to edit (or “correct”) deviant productions of /r/, a more common application for modified re-synthesis is to create stimuli for perceptual experiments. For example, specific speech sounds in a syllable have been transformed into intermediate and ambiguous forms between two prototypical phonemes (Protopapas, 1998). These stimuli have then been used in experiments of categorical perception. Others have modulated the phonemic nature of specific segments, while preserving the global intonation, syllabic rhythm and broad phonotactics of natural utterances, in order to study what

acoustic cues (e.g. phonotactics, syllabic rhythm) are most salient in identifying languages (Ramus & Mehler, 1999). In these types of applications, however, stimuli have been created once and there has been no need for real-time processing.

The computer-assisted language learning system VILLE (Wik, 2004) includes an exercise that involves modified re-synthesis. Here, the segments in the speech produced by the user are manipulated in terms of duration, i.e. stretched or shortened, immediately after recording. At the surface, this application shares several traits with the application suggested in this paper. However, more extensive manipulation is required to turn one phoneme into another, which is the goal for the system described here.

Purpose

The purpose of this study was to find out if it is at all possible to remove the initial voiceless plosive from a recorded syllable, and replace it with an “artificial” segment so that it sounds natural. The “artificial” segment is artificial in the sense that it was never produced by the speaker, but constructed or retrieved from somewhere else. As voiceless plosives generated by formant synthesizers are known to lack in naturalness (Carlson & Granström, 2005), retrieving the target segment from a speech database was considered a better option.

Method

Material

The Swedish version of the Speecon corpus (Iskra et al, 2002) was used as a speech database, from which target phonemes were selected. This corpus contains data from 550 adult speakers of both genders and of various ages. The speech in this corpus was simultaneously recorded at 16 kHz/16 bit sampling frequency by four different microphones, in different environments. For this study, only the recordings made by a close headset microphone (Sennheiser ME104) were used. No restrictions were placed on gender, age or recording environment. From this data, only utterances starting with an initial voiceless plosive (/p/, /t/ or /k/) and a vowel were selected. This resulted in a speech database consisting of 12 857 utterances (see Table 1 for details). Henceforth, this

speech database will be referred to as “the target corpus”.

For the remainder part of the re-synthesis, a small corpus of 12 utterances spoken by a female speaker was recorded with a Sennheiser m@b 40 microphone at 16 kHz/16 bit sampling frequency. The recordings were made in a relatively quiet office environment. Three utterances (/tat/, /kak/ and /pap/) were recorded four times each. This corpus will be referred to as “the remainder corpus”.

Table 1. Number of utterances in the target corpus.

	<i>Nbr of utterances</i>
Utterance-initial /pV/	2 680
Utterance-initial /tV/	4 562
Utterance-initial /kV/	5 614
Total	12 857

Re-synthesis

Each step in the re-synthesis process is described in the following paragraphs.

Alignment

For aligning the corpora (the target corpus and the remainder corpus), the NALIGN aligner (Sjölander, 2003) was used.

Feature extraction

For the segments in the target corpus, features were extracted at the last frame before the middle of the vowel following the initial plosive. For the segments in the remainder corpus, features were extracted at the first frame after the middle of the vowel following the initial plosive. The extracted features were the same as described by Hunt & Black (1996), i.e. MFCCs, log power and F0. The Snack tool SPEATURES (Sjölander, 2009) was used to extract 13 MFCCs. F0 and log power were extracted using the Snack tools PITCH and POWER, respectively.

Calculation of join cost

Join costs between all possible speech segment combinations (i.e. all combinations of a target segment from the target corpus and a remainder segment from the remainder corpus) were calculated as the sum of

1. the Euclidean distance (Taylor, 2008) in F0
2. the Euclidean distance in log power

3. the Mahalanobis distance (Taylor, 2008) for the MFCCs

F0 distance was weighted by 0.5. A penalty of 10 was given to those segments from the target corpus where the vowel following the initial plosive was not /a/, i.e. a different vowel than the one in the remainder corpus. The F0 weighting factor and the vowel-penalty value were arrived at after iterative tuning. The distances were calculated using a combination of Perl and Microsoft Excel.

Concatenation

For each possible segment combination ((/p|t|k/) + (/ap|at|ak/), i.e. 9 possible combinations in total), the join costs were ranked. The five combinations with the lowest costs within each of these nine categories were then concatenated using the Snack tool CONCAT. Concatenation points were located to zero-crossings within a range of 15 samples after the middle of the vowel following the initial plosive. (And if no zero-crossing was found within that range, the concatenation point was set to the middle of the vowel.)

Evaluation

7 adult subjects were recruited to perform a listening test. All subjects were native Swedes, with no known hearing problems and naïve in the sense that they had not been involved in any work related to speech synthesis development.

A listening test was constructed in Tcl/Tk to present the 45 stimuli (i.e. the five concatenations with the lowest costs for each of the nine different syllables) and 9 original recordings of the different syllables. The 54 stimuli were all repeated twice (resulting in a total of 108 items) and presented in random order. The task for the subjects was to decide what syllable they heard (by selecting one of the nine possible syllables) and judge the naturalness of the utterance on a scale from 0 to 100. The subjects had the possibility to play the stimuli as many times as they wanted. Before starting the actual test, 6 training items were presented, after which the subjects had the possibility of asking questions regarding the test procedure.

Statistical analysis

Inter-rater agreement was assessed via the intraclass correlation (ICC) coefficient (2, 7) for syllable identification accuracy and naturalness rating separately.

Pearson correlations were used to assess intra-rater agreement for each listener separately.

Results

The results of the evaluation are presented in Table 1.

Table 1. Evaluation results for the concatenated and original speech samples. The first column displays the percentage of correctly identified syllables, and the second column displays the average naturalness judgments (max = 100).

	% correct syll	Naturalness
Concatenated	94%	49 (SD: 20)
Original	100%	89 (SD: 10)

The listeners demonstrated high inter-rater agreement on naturalness rating (ICC = 0.93), but lower agreement on syllable identification accuracy (ICC = 0.79).

Average intra-rater agreement for all listeners was 0.71 on naturalness rating, and 0.72 on syllable identification accuracy.

Discussion

Considering that the purpose of this study was to study the possibilities of generating understandable and close to natural sounding concatenations of segments from different speakers, the results are actually quite promising. The listeners' syllable identification accuracy of 94% indicates that comprehensibility is not a big problem. Although the total naturalness judgement average of 49 (of 100) is not at all impressive, an inspection of the individual samples reveals that there are actually some concatenated samples that receive higher naturalness ratings than original samples. Thus, the results confirm that it is indeed possible to generate close to natural sounding samples by concatenating speech from different speakers. However, when considering that the long-term goal is a working system that can be implemented and used to assist phonological therapy with children, the system is far from complete.

As of now, the amount of manual intervention required to run the re-synthesis process is large. Different tools were used to complete different steps (various Snack tools, Microsoft Excel), and Perl scripts were used as interfaces between these steps. Thus, there is still a long way to real-time processing. Moreover, it is still limited to voiceless plosives in sentence-initial positions, and ideally, the system should

be more general. However, considering that children usually master speech sounds in word-initial and word-final positions later than in word-medial positions (Linell & Jennische, 1980), this limitation should not be disqualifying on its own.

The speech data in this work came from adult speakers. New challenges can be expected when faced with children's voices, e.g. increased variability in the speech database (Gerosa et al, 2007). Moreover, variability in the speech of the intended user - the child in the therapy room - can also be expected. (Not to mention the variability from child to child in motivation and ability and will to comply with the therapist's intervention plans.)

The evaluation showed that there is much room for improving naturalness, and fortunately, some improvement strategies can be suggested. First, more manipulations with weighting factors might be a way to assure that the combinations that are ranked the highest are also the ones that sound the best. As of now, this is not always the case. During the course of this investigation, attempts were made at increasing the size of the target corpus, by including word-initial voiceless plosives within utterances as well. However, these efforts did not improve the quality of the output concatenated speech samples. The current system does not involve any spectral smoothing; this might be a way to polish the concatenation joints to improve naturalness.

Looking beyond the context of modified resynthesis to assist therapy with children with phonological impairments, the finding that it is indeed possible to generate natural sounding concatenations of segments from different speakers might be valuable in concatenative synthesis development in general. This might be useful in the context of extending a speech database if the original speaker is no longer available, e.g. with new phonemes. However, it seems reasonable to assume that the method is only applicable to voiceless segments.

Acknowledgements

This work was funded by The Swedish Graduate School of Language Technology (GSLT).

References

Carlson, R. & Granström, B. (2005) Data-driven multimodal synthesis. *Speech Communication* 47, 182-193.

- Gerosa, M., Gioliani, D. & Brugnara, F. (2007) Acoustic variability and automatic recognition of children's speech. *Speech Communication* 49, 847-835.
- Hewlett, N. (1992) Processes of development and production. In Grunwell, P. (ed.) *Developmental Speech Disorders*, 15-38. London: Whurr.
- Hunt, A. and Black, A. (1996) Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of ICASSP 96 (Atlanta, Georgia)*, 373-376.
- Iskra, D., Grosskopf, B., Marasek, K., Van Den Heuvel, H., Diehl, F., and Kiessling, A. (2002) *Speecon - speech databases for consumer devices: Database specification and validation*.
- Linell, P. & Jennische, M. (1980) *Barns uttalsutveckling*, Stockholm: Liber.
- Locke, J.L. & Pearson, D.M. (1992) *Vocal Learning and the Emergence of Phonological Capacity. A Neurobiological Approach*. In C.A. Ferguson, L. Menn & C. Stoel-Gammon (Eds.), *Phonological Development. Models, research, implications.*, York: York Press.
- Protopapas, A. (1998) Modified LPC resynthesis for controlling speech stimulus discriminability. *136th Annual Meeting of the Acoustical Society of America, Norfolk, VA, October 13-16*.
- Ramus, F. & Mehler, J. (1999) Language identification with suprasegmental cues: A study based on speech resynthesis, *Journal of the Acoustical Society of America* 105, 512-521.
- Shuster, L. I. (1998) The perception of correctly and incorrectly produced /r/. *Journal of Speech, Language and Hearing Research* 41, 941-950.
- Sjölander, K. *The Snack sound toolkit*, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden. Online: <http://www.speech.kth.se/snack/>, 1997-2004, accessed on April 12, 2009.
- Sjölander, K. (2003) An HMM-based system for automatic segmentation and alignment of speech. *Proceedings of Fonetik 2003 (Umeå University, Sweden)*, PHONUM 9, 93-96.
- Taylor, P. (2008) *Text-to-Speech Synthesis*, Cambridge University Press.
- Wik, P. (2004) Designing a virtual language tutor. *Proceedings of Fonetik 2004 (Stockholm University, Sweden)*, 136-139.