Modified re-synthesis of segments in isolated words: A theoretical background

Sofia Strömbergsson

January 18, 2008

Abstract

This paper presents a theoretical background to an imagined application assumed to be used in a speech and language therapy setting, where the assumed user is a child who produces systematically deviant speech. The application expects as input an isolated word spoken by the child in a specific way (i.e. as a specific sequence of phones), and where one of the phonemes is substituted by another phoneme. The application aligns the speech signal to a transcription of the word, modifies the erroneous segment so that it sounds like the "correct" phoneme, and concatenates the modified segment with the unmodified parts of the word. The assumed output produced by the application is thus a modified and resynthesised version of the speech signal the child provided as input – an approximation of what the given word would have sounded like if the child had produced it correctly. This report presents a theoretical background to each of these modules, reviewing different techniques that have been used in similar applications earlier. Ideally, this background to the challenges involved in segmentally modified resynthesised speech can guide the decision on how – if at all! – to embark on the practical implementation of the application.

1 Introduction

Computer technology can be used to assist speech and language therapy for children with deviant or delayed speech and language development. In some areas, computer technology provides possibilities that go beyond what a human therapist (or teacher) can offer. For example, some programs (e.g. SpeechViewer and Box of Tricks, see [1]), provide immediate visual feedback on speech production, thereby enhancing auditive proprioceptive feedback. By modifying his/her speech production, the child manipulates both acoustic features *and* visual features visible on the screen. This has proved to be particularly beneficial for children with hearing impairments [1]. Furthermore, articulatory models (i.e. visual representations of the articulation of specific sounds) can illustrate information that is not normally visible in a human speaker, to provide guidance to the user as to how to place articulators to produce a specific sounds. An example of this is the ARTiculation TUtoR (ARTUR), developed at KTH [2]. However, although there are potential benefits of computer-assistance in speech and language therapy, computers are only rarely used in clinical practice in Sweden today.

In normal as well as deviant speech 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 [3]. 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 [4]. As expressed by Locke and Pearson [4] (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, final consonant deletions and assimilations. Some of these children might well percieve phonological distinctions that they themselves do not produce, while others have problems both in percieving and producing a distinction.

Based on the above, it seems reasonable to assume that enhanced feedback of *one's own* speech might be particularly valuable to children with phonological difficulties, in increasing their awareness of their own speech production. For instance, how would a child with phonological difficulties react to hearing what his/her speech would have sounded like if s/he had produced speech "correctly" (preferrably in comparison with his/her current speech production)? Studying the effects of performing such an exercise might help to gain more insights to the nature of the phonological difficulties these children have, as well as providing implications for clinical intervention. Technically, this kind of exercise could be implemented in an application which takes as input the child's production of a specific word, and produces as output a modified version of this speech sample. Hence, this could be called "modified resynthesis".

1.1 Earlier applications of modified resynthesis

Modified resynthesis has been used as a way of creating stimuli for perceptual experiments, e.g. to produce syllables where specific speech sounds have been transformed into intermediate and ambiguous forms between two prototypical phonemes [5]. 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 [6]. In these types of applications, however, stimuli have been created once and there has been no need for real-time processing.

Modified resynthesis has also been used in efforts to increase intelligibility in dysarthric speech (i.e. poorly articulated speech, often due to reduced oral muscle strength and/or control in the speaker) [7] [8]. Kain et al. describe how they analysed dysarthric speech in terms of F_0, formant frequencies and energy, how these values were modified to resemble desired targets, and how transformed speech was generated using formant synthesis. However, the authors did not strive to perform the analysis and resynthesis on-the-fly. Moreover, naturalness and voice resemblance were not prioritised, as the main focus was increased intelligibility.

The computer-assisted language learning system VILLE [9] includes an exercise that involves modified resynthesis. Here, the segments in the speech produced by the user are manipulated in terms of duration, i.e. stretched or shortened. This application obviously shares several traits with the application suggested in this paper, and it will be referred to later.

1.2 The imagined application

The hypothetical application described in this paper is assumed to be used in a speech and language therapy setting. The assumed user is a child who produces systematically deviant speech and who might benefit from hearing how his/her speech would have sounded if s/he could pronounce words accurately. The application could be used in a therapy setting, with the child and therapist working together by the computer in a relatively quiet environment (typically a small room in a health care unit or at a school). The child is assumed to have a microphone close to his/her mouth (e.g. a headset).

The application expects as input an isolated word spoken by the child in a specific way (i.e. as a specific sequence of phones). For example, the child might produce the word "kotte" as /kɔkɛ/, instead of the correct /kɔtɛ/ (e.g. as a consequence of assimilation). Thus, the application "knows" that the target pronunciation is /kɔtɛ/ and that the child's attempt will be /kɔkɛ/ already before the child has begun to speak.

The assumed output produced by the application is a modified and resynthesised version of the speech signal the child provided as input. This speech signal is an approximation of what the given word would have sounded like if the child had produced it correctly. Thus, the phone-specific features of the erroneous speech segment (i.e. the medial /k/ in the example with "kotte") are modified and transformed (into a /t/), while speaker characteristic features of this segment are preserved. The parts of the word that the child pronounced correctly, i.e. the parts preceding and following the erroneous segment, are assumed to pass through the application unmodified. Ideally, output is produced in (near) real-time. The path from the input word spoken by the child to the output signal produced by the application is assumed to include four different modules, as illustrated in Figure 1.

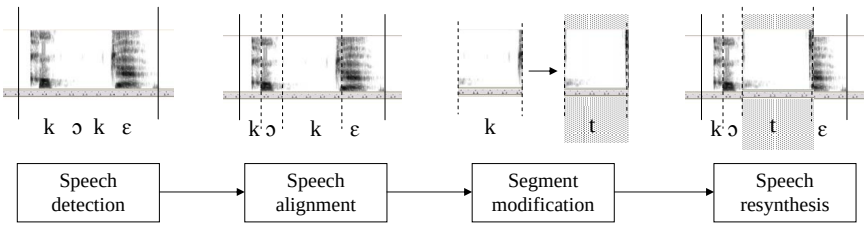

Figure 1: Graphical representation of the modified resynthesis application. (Please note that this representation is a simplification; as will be argued later, the most appropriate way of segmenting the speech wave is probably into diphones rather than into phones. The segment to be modified will thus consist of two diphones rather than of one phone.)

This report describes a theoretical background to each of these modules, reviewing different techniques that have been used in similar applications earlier. In Section 2, the different modules of the imagined application are described, one by one: Speech detection (Section 2.1), Speech alignment (Section 2.2), Modification of the erroneous segment (Section 2.3) and Resynthesis (Section 2.4). Finally, a conclusion is presented in Section 3. Ideally, this background to the challenges involved in segmentally modified resynthesised speech can guide the decision on how – if at all! – to embark on the practical implementation of the application.

2 Modules

2.1 Speech detection

The first step in the suggested application is to identify the start and end points of the isolated word spoken as input. The stronger the background noise and/or the weaker the speech signal (i.e. the smaller signal-to-noise ratio, SNR), the more difficult it is to distinguish the speech signal from the background noise. In recognition of isolated words, inaccurate detection of word endpoints is a major cause of error, and simply defining an energy level threshold (where everything louder than the threshold is treated as speech and everything below the threshold as noise), may not be sufficient [10]. To distinguish weak word-final sounds from background noise, to ignore "impulsive" noise before the word (e.g. smacks, inhalation etc.), and to handle silent periods within words (typically occlusion phases in plosives, are examples of challenges even in relatively quite environments [11].

Considering that the suggested application is assumed to be used in a relatively quiet environment, and the child is assumed to use a microphone close to his/her mouth, the SNR can be expected to be quite high. And as the input words are known in advance, even very long silent occlusion phases within a word can quite confidently be treated as word-internal silent gaps, since the application "knows" that the gap will be followed by more speech. In other words, the duration threshold can be very long. So, even if the speech detection step might not be trivial, the real challenges in designing the suggested application do probably not lie in this step, but in later modules.

2.2 Speech alignment

As the phoneme sequence of the incoming signal is already known, the task for the second module is not to recognise what segments it contains, but rather to locate the boundaries between the (known) segments. In several speech resynthesis applications, where real-time processing is not prioritised, phoneme segmentation is done manually [6] [8]. However, for our purposes, (near) direct feedback is necessary, and therefore segmentation must be done automatically. This task is often referred to as *forced alignment*, and usually builds upon techniques used in automatic speech recognition (ASR). In concatenative synthesis, accurate positioning of boundaries is, however, more important than for ASR purposes [12], as misplaced boundaries between phones might cause discontinuities in the concatenated speech. As our intended application also involves a

step of concatenative synthesis, we are also dependent on phoneme (or diphone) boundaries being placed as accurately as possible.

The most widely used techniques for automatic segmentation today are Hidden Markov Model (HMM)-based [13]. Here, an acoustic model of phoneme HMMs is derived from a corpus of spoken data with associated phonemic transcriptions. Alignment is then forced between the incoming speech signal and the HMMs associated with the phones in the known phone sequence. Obviously, the better the input speech corresponds to the training data, the more confident the alignments. Although this approach is considered the most reliable, it is limited when it comes to the *exact* placement of boundaries between phonemes [12] [13]. In order to overcome the limitations at phoneme transitions, several researchers have suggested that an initial HMM-based segmentation could be refined, e.g. by using boundary models derived from a smaller training database of manually labelled and segmented speech [14], or by recognising spectral discontinuities as cues to phonemic boundaries [12].

2.2.1 The training corpus

Assuming we use a statistically based approach to speech segmentation, we need a corpus of speech from which to derive acoustic models. For our application, the ideal training corpus would consist of speech produced by children the same age as the intended user, and in the same kind of acoustical environment. However, even if the acoustic models used are based on child speech, automatic recognition of child speech is still more difficult - and much less studied - than recognition of adult speech [15] [16]. A key problem in collecting a representative database of child speech is *variability*; children's voices (and motor skills) are developing and changing while adult's voices are more stable.

Available at KTH is the PF_Star corpus [17], a Swedish child-speech database, with collected speech samples from 200 children aged 4-8 years, from the Stockholm region. The acoustic environment is similar to that of our intended application; all recordings were done in small rooms with only the child and an adult present. The corpus consists of sentences (imitated after an adult, since many of the children couldn't read) and digit sequences. Although these types of utterances are somewhat different from the expected input in our assumed application, using a child-speech corpus might still be a better choice than using an adult-speech corpus.

2.3 Modification of the erroneous segment

Once the incoming speech signal has been segmented, and the erroneous segment has been identified, the next step is to modify this segment. Ideally, we want to modfiy only phoneme-specific features and preserve speaker-specific features, and we are therefore interested in a way of separating the two types of information. This is an interest shared with the field of *voice conversion*, where the goal is to modify an utterance spoken by one speaker, so that it sounds as if spoken by another speaker [18]. But what features in the speech signal are phoneme-specific and what features are speaker-specific? As it turns out, there has been little success in isolating specific acoustic parameters that capture all voice-characteristic features in a speech signal [19].

A technique that is often used in voice conversion is Linear Predictive Coding (LPC), to estimate what information in the speech signal corresponds to the excitation signal (the "source"), and what information corresponds to the resonances of the vocal tract (the "filter") [20]. Spectral modification, e.g. modification of formant frequency values, is then done by passing the source part of the speech signal (the LPC residual) through a filter that corresponds to the spectral shape of a target speech sound. This technique, Residual Excited LPC (RELP), (although somewhat modified) was for example used by Kain et al. [8] in their effort at improving intelligibility in dysarthric speech through modified resynthesis. Protopapas [5] also used RELP to create intermediate stimuli between two specific phonemes, and to extrapolate beyond the typical phonemes, thus creating "exaggerated" versions of the phonemes.

There are some limitations in using RELP in speech modification, however. Ideally, if the LPC residual were an accurate and "clean" representation of the voice signal, its spectral shape would be sloping without any formant peaks or valleys. The LPC residual is just an approximation, however, and passing it through a different filter than the original often leads to distortions in the modified speech [20]. Different ways of improving the results have been suggested, e.g. in [20].

The type of segments, source and target, and the acoustic distance between them will affect how modification is performed. An obvious limitation of the RELP technique is that it has only been applicable to voiced speech sounds. For example, Kain et al. [8] only resynthesised voiced regions of the input speech, and let unvoiced frames pass unmodified to the output. Acoustic features that are usually mentioned as speaker-characteristic are often related to the speaker's voice (e.g. fundamental frequency, properties of the glottal source spectrum), and the vocal tract resonance (e.g. formant frequencies and bandwidths, amplitude spectra of vowels and nasals) [11] [19]. Is it safe, then, to assume that unvoiced phonemes do not carry any speaker-characteristic information? If so, could we simply replace (instead of modifying) any unvoiced source segment with any unvoiced target segment produced by any other speaker? Or even with a formant-synthesised target segment? Attempts have been made at integrating waveform concatenation and formant synthesis, showing particularly promising results for splicing together synthesised (non-nasal) obstruents and natural speech segments [21]. Another example of successfull integration of formant synthesis and natural speech segments was presented in [22], where naturalness of formant synthesis was improved by replacing unvoiced segments synthesised by the formant synthesiser with their natural (recorded) correspondences. According to Hertz's findings, the integration of synthesised and natural segments does not affect the perception of voice quality, as long as stressed syllable nuclei are still made up from natural segments. If this proves right, it would make the task of modifying unvoiced phonemes considerably easier, assuming we also find a way of handling transitions from and to surrounding voiced sounds appropriately. For voiced phonemes, however, or for source-target phoneme pairs which are very different, the modification is assumably much more complex.

So far, it has been assumed (at least implicitly) that the speech segment to be modified is a phone. However, most concatenative synthesis systems use segments that include the transitions between phones, e.g. diphones, as the basic units. So, the segment to be modified is actually not only the phone, but should probably include the preceding and following semi-phones as well.

2.4 Speech resynthesis

Once the erroneous diphone segments have been modified, the last task is to "transplant" this modified speech segment into the original word. However, this is not a trivial question of cutting and pasting, but rather a complex task of merging the modified and the unmodified parts together. Obtaining smooth transitions between concatenated speech segments is a key challenge within concatenative synthesis, especially when the database is small, with only a limited number of segments to choose from [23]. A major difference between common speech synthesis systems and the resynthesis application suggested here can be illustrated with the previously referred example word "kotte". In a large speech corpus, you will assumably find many examples of the sequence [-ɔt-], and the challenge is to choose which one could best match the preceding end point. In our case, there is only one candidate for the [-t-] segment (i.e. the output from the modification step), and the task will be to bridge the gap between the unmodified speech and the modified signal as well as possible. However, as some concatenative speech synthesis systems actually do apply some modification to the original segment to smooth the concatenation point to the preceding signal, some techniques have been developed that might fit our purposes.

Pitch-synchronous overlap and add (PSOLA) is a common technique to adjust pitch and duration of a segment [23] [11]. Here, pitch markers are inserted at certain positions in the glottal cycle, and segments are joined at these positions, in a pitch-synchronous manner. Before they are joined together, the segments are tapered towards the end, and overlapped. Pitch modification can be done by reducing or increasing the length between pitch markers, and duration can be modified by removing or repeating pitch pulses. For PSOLA synthesis to work well, the positions of the pitch markers have to be accurate. A way to achieve reliable pitch markers is by recording glottal activity simultaneously with speech, by using a laryngograph. However, using a laryngograph will not be an option in our application, and we might therefore risk ending up with misplaced pitch markers.

Some approaches aimed at smoothing spectral discontinuities between segments have been described in [23]. Spectral smoothing can be done either by modifying the existing audio frames, or by adding frames to interpolate between the segments. In *optimal coupling*, the segment boundaries are not fixed to a certain point, but rather are specified as being located within certain frame ranges. The exact segment joint is determined during synthesis, as the combination of a start and end point that yields the best spectral fit in the concatenated signal [23]. Although the simplicity of this approach might be appealing, a disadvantage is that boundary points might sometimes be pushed too far, so that essential parts of a segment are lost. *Waveform interpolation* (WI) is a technique of averaging between the end point of a preceding segment to the starting point of a following segment. However, WI only brings small improvements and works best with segments with similar spectral envelopes [23]. WI can also be used with linear prediction methods, to bridge the gap between the LP residuals of two bordering frames. Of course, the spectral components of the LP-filter of the same bordering frames could be interpolated independently of the interpolation of the LP-residual. When applied to pitch synchronous windows, this method has shown promising results [23]. A quite different approach, also described in [23], is to mask discontinuities with noise. This approach takes advantage of an

psychoacoustic phenomenon, the *continuity effect*. Similarily to what happens when we look at a scene while moving past a picket fence, and still perceive all information as continuous, adding noise to mask the segment joints in concatenative synthesis produces the same effect, although auditorily [23]. However, this approach has not yet been extensively studied and one could guess that although it might be advantageous for the intelligibility of the synthesised speech, naturalness might be disturbed. And for our purposes, where acoustic information is produced to assist the child in perceiving acoustic details, masking any details might be counter-productive.

Most of the smoothing techniques described here are appropriate for use with voiced speech, but not designed particularly to handle unvoiced speech [23]. In PSOLA synthesis, pitch markers can obviously be placed only in voiced regions of the speech signal. For unvoiced speech regions, markers can be placed at arbitrary positions at a constant rate, as the positions of the analysis windows are not as critical as in voiced regions [11]. Special care is however needed for stop consonants, considering that misplaced pointers might yield frames containing both the last part of the occlusion phase and the first part of the explosion phase.

3 Conclusion

The task of implementing a system for segmentally modified resynthesised speech is complex. Each module described in this paper represents a different field of research in itself. However, to a certain extent, we could make use of earlier experiences and existing systems. For example, up to the point of speech alignment, the imagined application could follow the same steps as e.g. the duration modification exercise in VILLE [9] referred to earlier. Here, an aligner tool available at KTH [24] is used to time-mark the phone boundaries in the waveform of the utterance. However, as the speech modification in VILLE only involves stretching and shortening of segments, the exact placements of phone boundaries are probably less sensitive than they are for our purposes. Moreover, as the assumed users of VILLE are adult speakers, while the assumed users of the application described here are children, alignments would probably be more reliable if the aligner could build upon acoustic models derived from child speech data. Fortunately, child speech data is also available at KTH (the PF-Star corpus), but it remains to be studied if it is better to use acoustical models that build on child speech data of the "wrong" format (i.e. not isolated words), than adult speech data of the "correct" format.

It is assumed that the type of segment to be modified, the modification target and the context of the segment will affect both the processing and resulting speech quality. Most methods for speech modification and concatenative (re)synthesis have been designed to treat voiced speech sounds. There could be two reasons for this; either unvoiced speech sounds are too complex to even try to handle, or handling voiced speech sounds is more acute since they cause more problems in speech modification and (re)synthesis. As I find the second explanation more likely, it seems reasonable to begin with treating unvoiced speech sounds in the practical implementation of the suggested application. (Again, without overlooking the transitions from and to neighbouring voiced regions.) Moreover, the risk of introducing speech discontinuities is assumably smaller

when "transplanting" an initial or final segment onto a word than inserting a medial segment, as there is only one concatenation point instead of two. Therefore, if practical implementation is embarked upon, it would seem reasonable to limit the application to the handling of unvoiced segments in word-initial or word-final position, at least as a starting point. Considering that children usually master speech sounds in word-initial and word-final positions later than in word-medial positions [25], this starting point can be doubly motivated.

References

[1] Öster, A-M. (2006) Computer-Based Speech Therapy Using Visual Feedback with Focus on Children with Profound Hearing Impairmnets. Doctoral Thesis in Speech and Music Communication, Stockholm.

[2] Bälter, O., Engwall, O., Öster, A-M. & Kjellström, H. (2005) Wizard-of-oz test of ARTUR - a comptuerbased speech training system with articulation correction. *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, 36-43.

[3] Hewlett, N. (1992) Processes of development and production. In P. Grunwell (Ed.), *Developmental Speech Disorders* (pp. 15-38), London: Whurr.

[4] 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.

[5] Protopapas, A. (1998) Modified LPC resynthesis for controlling speech stimulus discriminability. 136^{th} *Annual Meeting of the Acoustical Society of America*, Norfolk, VA, October 13-16.

[6] 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.

[7] Kain, A.B., Niu, X., Hosom, J., Miao, J., van Santen, J.P.H. (2004) Formant resynthesis of dysarthric speech, In *IEEE Workshop on Speech Synthesis*, Pittsburg, PA, 25-30.

[8] 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*, 743-759.

[9] Wik, P. (2004) Designing a virtual language tutor, in *Proc of the XVIIth Swedish Phonetics Conference, Fonetik 2004* (pp. 136-139). Stockholm University.

[10] Karray, L. & Martin, A. (2003) Towards improving speech detection robustness for speech recognition in adverse conditions, *Speech communication, 40:3*, 261-276.

[11] Holmes, J. & Holmes, W. (2001) *Speech synthesis and recognition* (2nd ed), London/New York: Taylor & Francis.

[12] Kim, Y.-J. & Conkie, A. (2002) Automatic segmentation combining an HMM-based approach and spectral boundary correction, In *ICSLP-2002*, 145-148.

[13] Jarifi, S., Pastor, D. & Rosec, O. (2008) A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis, *Speech Communication, 50*, 67-80.

[14] Wang, L., Zhao, Y., Chu, M., Zhou, J. & Cao, Z. (2004) Refining segmental boundaries for TTS database using fine contextual-dependent boundary models, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May, Vol. I, 641-644.

[15] Giuliani, D. & Gerosa, M. (2003) Investigating recognition of children's speech, *IEEE International Conference on Acoustics, Speech and Signal Processing*, April, Vol. 2, 137-140.

[16] Potamianos, A. & Narayanan, S. (2003) Robust recognition of children's speech, *IEEE Transactions on speech and audio processing*, Vol. 11:6, 603-616.

[17] Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S. & Wong, M. (2005) The PF_STAR Children's Speech Corpus, in *Proc Interspeech 2005*.

[18] Kain, A. & Macon, M.W. (2001) Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, 813-816.

[19] Kuwabara, H. & Sagisaka, Y. (1995) Acoustic characteristics of speaker individuality: Control and conversion, *Speech Communication, 16*, 165-173.

[20] Wouters, J. & Macon, M.W. (2001) Control of spectral dynamics in concatenative speech synthesis, *IEEE Transactions on Speech and Audio Processing, 9:1*, 30-38.

[21] Hertz, S. (2002) Integration of rule-based formant synthesis and waveform concatenation: A hybrid approach to text-to-speech synthesis, in *Proc. IEEE 2002 Workshop on Speech Synthesis*. Santa Monica, USA, September 2002.

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

[23] Chappell, D.T. & Hansen, J.H.L. (2002) A comparison of spectral smoothing methods for segment concatenation based speech synthesis. *Speech Communication, 36*, 343-373.

[24] Sjölander, K. (2003) An HMM-based system for automatic segmentation and alignment of speech, in *Proc of Fonetik 2003* (pp. 93-96). Umeå University.

[25] Linell, P. & Jennische, M. (1980) *Barns uttalsutveckling*, Stockholm: Liber.