Substitute Sounds for Ventriloquism and Speech Disorders

Jörg Metzner, Marcel Schmittfull, Karl Schnell

Balliol College, University of Oxford, OX1 3BJ, Oxford, UK
Institute of Applied Physics, Goethe-University Frankfurt
Max-von-Laue Str.1, 60438 Frankfurt am Main, Germany
Joerg.Metzner@balliol.ox.ac.uk, Schmittfull@gmx.de, Schnell@iap.uni-frankfurt.de

ABSTRACT

The restriction of articulation used by ventriloquists or caused by speech disorders can be compensated using substitute sounds. For a better understanding of these sounds, in this contribution the results of investigations of the substitute sounds by analysis and synthesis are presented. For that purpose the substitute sounds and their natural counterparts uttered by a ventriloquist are analyzed. Substitute sounds are also generated by articulatory synthesis and by a plaster model and are then compared to the original sounds. The results show that the perception of substitute sounds can reach the natural uttered sounds. The degree of the perceptive similarity depends on the substitute sound. The spectrum of the substitute sounds can replicate partially the spectrum of the natural sounds. However, for successful perception of the sound it is only necessary to reproduce the relevant spectral information. The knowledge about those sounds is interesting for speech therapy.
Index Terms: articulation, speech production

1. INTRODUCTION

Usually the production of sounds can use the whole articulatory space of the speech production system [1]. In special cases the articulators are restricted intentionally by ventriloquists or forced by speech disorders. The restrictions of the articulators affect particular sounds in their natural articulatory positions. The purpose of the substitute sounds is to compensate these restrictions by an elegant dexterous articulatory constellation and nevertheless produce the respective sounds as good as possible. Substitute sounds are essential for ventriloquism and might be used against speech disorders when only certain parts of the articulation cannot be used anymore. Ventriloquism is the art of speaking in such a way that a viewer cannot identify the speaking from observing the face and especially the lips. Therefore the articulatory restrictions affect the articulators, which can be seen by a viewer, namely the lips, the jaw position and in some cases the tongue tip.

2. SUBSTITUTE SOUNDS

2.1 Synthesis of ventriloquist's substitute sounds

For the investigations of substitute sounds they are synthesized by an articulatory synthesis system. For this purpose the simulation program tractsyn [2] is used allowing to adjust the articulators of a three dimensional articulatory model. The corresponding transfer function can be calculated from the resulting vocal tract areas, with a consideration of the nasal tract. The nasal tract is modelled in tractsyn with side cavities. From the magnitude response the formants and antiformants can be observed. Furthermore, a synthesized speech signal can be generated by a sequence of articulatory positions and movements. This feature allows the generation of a CV phone chain (consonant vowel) which is necessary for perceptive tests of non-stationary sounds.
The adjustment of the articulators of the speech production system is performed in a two stage process. Initially the articulators are adjusted with hints of literature and ventriloquists. Then a tuning to optimal articulatory positions is performed with respect to the formants of the transfer function and a perceptive test of synthesis results, considering the restriction of articulation. In addition to this approach, a search of new articulatory positions of substitutes is performed. It must be emphasized that the available information about the substitute sounds of ventriloquism is relatively scarce not at least because professional ventriloquists don't want to reveal their secrets about ventriloquism. Even though it is possible for conventional speakers to learn to ventriloquize, it requires long-term training to get used to using a completely different method to produce certain sounds. Substitute sounds are denoted by a prime, for example /b/ (normal) /b'/ (substituted).

2.1.1 Plosives /b/ /p/ and substituted /b'/ /p'/ by tractsyn

From a phonetic view and definition the closed lips are essential for the generation of the plosives /b/ and /p/ due to the bilabial building of constriction. These sounds are especially difficult to substitute for conventional speakers. Since the closing of the lips is not allowed in the case of ventriloquism a constriction behind the lips is chosen. However, the constriction's place and especially the shape of the whole tongue should be adjusted in a way that the first two formants are equal to the formants of /b/, differing from the plosive /d/. It has to be mentioned that the plosives are non-stationary sounds. Their resonances are effective in speech utterances by the bending of formants to the subsequent voiced sound corresponding to the locus-theory. Fig. 1 shows the models with corresponding transfer functions of /b/ and /b'/. The first spectral difference between /b/ and /b'/ is given by the third formant; the first two formants are reproduced correctly. Perceptive tests of the synthesized sequences [ba] and [b'a] show, that the substitute sound can hardly be distinguished

from the natural sound /b/ in a CV sequence. For comparison, the synthesized CV sequence with /d/ is also analytically and perceptively evaluated, resulting that the CV sequence with /d/ is distinctly different in the relevant spectrum frequency to the sequence with /b'/. However, it can be observed that in the higher frequency spectrum there are more similarities between /b'/ and /d/ than between /b'/ and /b/.

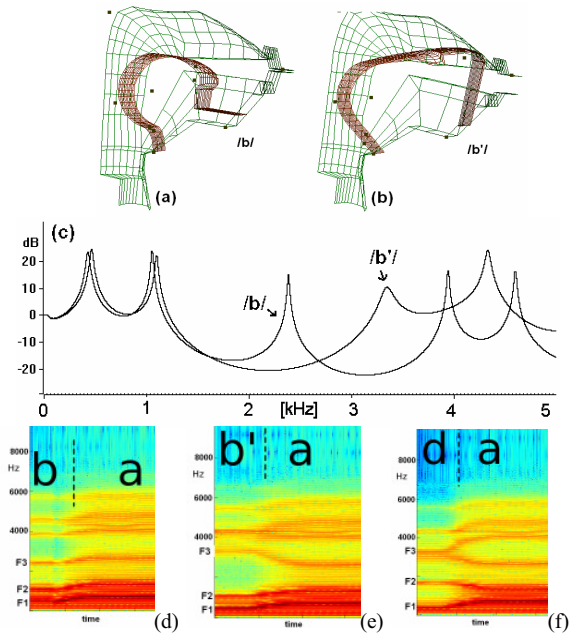

Figure 1: (a) and (b) are the articulatory models of [b] and [b']; (c) corresponding magnitude responses to the models (a) and (b); (d)-(f) spectrograms of the synthesized [ba], [b'a] and [da].

2.1.2 Fricative /v/ and substitute /v'/

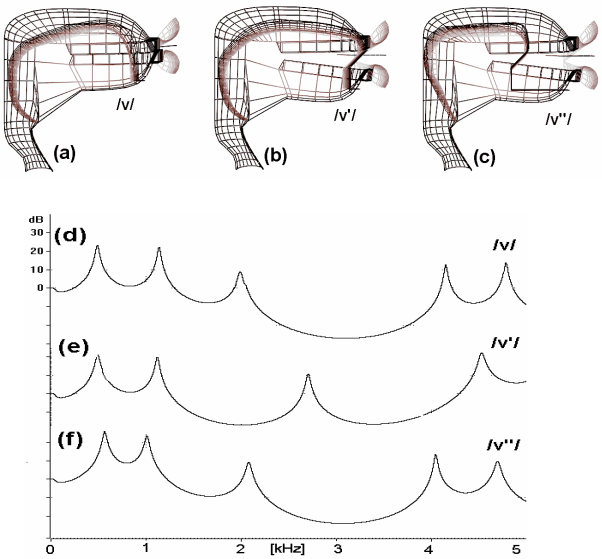

Figure 2: (a), (b), and (c) are the articulatory models of [v], [v'], and [v'']; (d)-(f) corresponding magnitude responses to the models (a)-(c).

The voiced fricative /v/ has a constriction at the alveolar position. The substitute sound /v'/ proposed from literature [3] has a constriction behind the lips. The articulatory model of /v'/ is shown in fig. 2(b) with transfer function in 2(e). It can be seen that the frequency of the third formant is increased by the substitution. Furthermore, a new substitute sound /v''/ is found improving the spectral fit. From the articulatory model of /v''/ in fig. 2(c) it can be seen that the tongue is placed backward in comparison to /v'/.

2.1.3 Nasal /m/ and substituted /m'/ /m''/

In the case of the production of nasals the closed mouth cavity serves as a side cavity of the speech production system. If a side cavity is coupled to a tube system, zeros and poles are introduced to the transfer function. There exist two known substitute sounds /m'/ and /m''/ for the nasal /m/ from literature. The first substitute sound /m'/ has a constriction a little bit behind the teeth whereas the second substitute sound /m''/ has a constriction near the velum. The articulatory models of /m/, /m'/, and /m''/ with corresponding magnitude responses from the glottis to the nostrils are shown in fig. 3. The frequency range up to 1 kHz can be reproduced relatively well by the substitutions. Over 1 kHz the spectrum can be approximated only partially by the substitute sounds, however, in the case of /m'/ the similarity to /m/ is better.

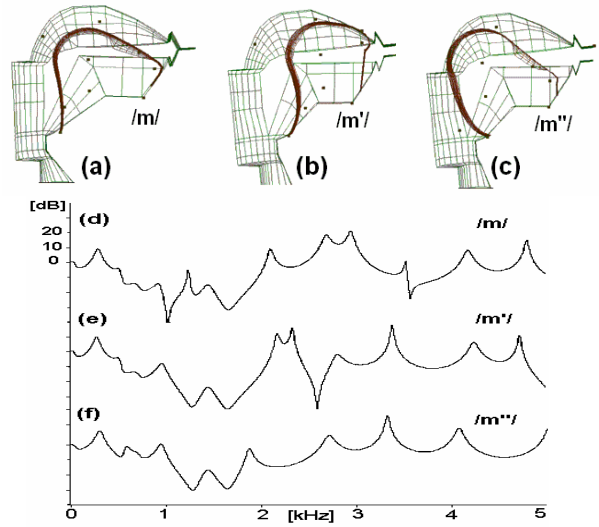

Figure 3: (a), (b), and (c) are the articulatory models of /m/, /m'/, and /m''/; (d)-(f) corresponding magnitude responses to the models (a)-(c).

2.2 Modeling of nasal substitute sounds in an experiment using a plaster model

In addition to the generation of the nasal /m/ and the substitutes by articulatory synthesis, the nasals are produced by a replica of the speech tract by a real plaster model. The plaster model represents the external boundaries of the coupled cavities of the speech production system: the mouth cavity, the pharynx, and the nasal tract. Side cavities of the nasal tract are also implemented (fig. 4). The size of the cavities is chosen with respect to the data of literature, for example [4, 5]. The mouth

cavity approximates the condition of the nasal /m/. The plaster model was chosen in competition with a metal based

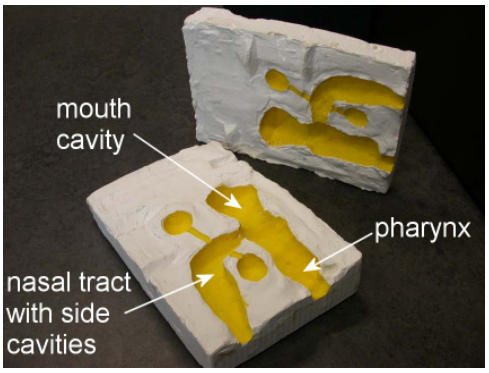

Figure 4: Plaster model of nasal /m/.

model, due to the losses. One critical point is the excitation of the model. For that purpose a special loudspeaker is chosen. The loudspeaker is closed and emits the sounds over a connected rubber tube which is joined with the plaster model. The radiated sounds from the plaster model can be recorded. For the spectral analysis of the recorded signals an impulse train is used. However, for a perceptive evaluation the periodic impulse train lacks in naturalness. Therefore the residual signal obtained by a linear prediction of a voiced speech sound is chosen, producing relatively natural sounding sounds. To obtain the spectral decrease of voiced speech the residual is filtered twice by a system with a real pole for de-emphasis.

The geometry of the plaster model represents the nasal /m/. To generate the substitute sounds the geometry of the mouth cavity is modified by plastic modeling material. For that purpose the mouth cavity is shortened with respect to /m'/ or /m''/. For comparison the mouth cavity is also modified for the nasal /n/. For all plaster models the same filtered residual with de-emphasis is used. The nasals /m/ and /n/ produced by the corresponding plaster models can be identified and are distinguishable. The produced substitute sound /m'/ can be accepted for the nasal /m/. A perceptive comparison of /m/, /m'/ and /n/ yield that the produced signals /m'/ are perceptually between /m/ and /n/, however, closer to the nasal /m/.

Fig. 5 shows the DFT-spectra of the generated signals by the plaster model of /m/ and /m'/ with impulse train excitation. The main zero caused by the mouth cavity can be observed at 1000Hz for /m/ and 1250 Hz for /m'/ which is marked in fig. 5.

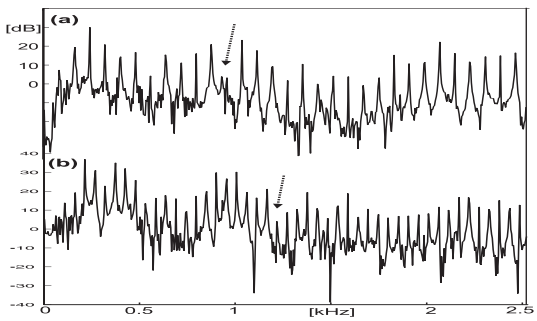

Figure 5: Spectra of the output of the plaster model (a) for /m/ and (b) for /m'/; the excitation is the impulse train.

The higher frequency of this zero is caused by the shortened mouth cavity. The resulting zeros can also be estimated from speech signals [6].

2.3 Synthesis of substitute sounds of speech disorders

For people with speech disorders it is very common that only the tip of their tongue is floppy and immobilized. It is not possible for these persons to raise their tongue tip, which is important for the production of the plosive /d/. Instead of the tip, the tongue blade can build the constriction. Fig. 6 (b) shows the articulatory positions of the substituted plosive /d'/ with a lowered tongue tip compensated by the tongue blade. The synthesized CV sequences with and without the substitution are perceptively comparable.

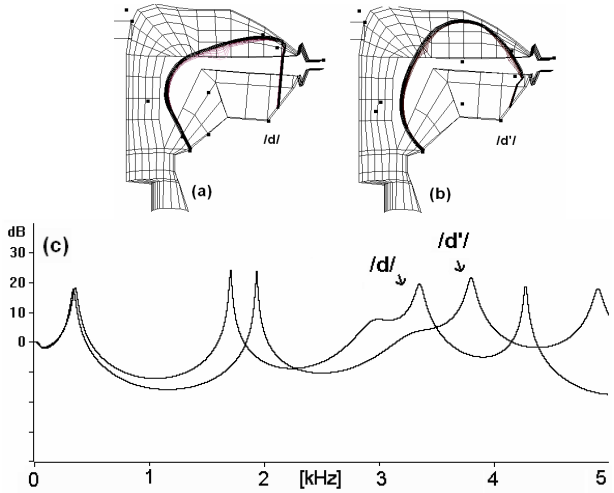

Figure 6: (a) and (b) are the articulatory models of /d/ and /d'/; (c) corresponding magnitude responses to the models (a)-(b).

2.4 Analysis of uttered substitute sounds

Besides of synthesis and model based simulations, utterances by a professional ventriloquist are analyzed. For that purpose the ventriloquist uttered selected sentences with multiple occurrences of critical sounds, which have to be substituted. Furthermore the ventriloquist uttered CV sequences each naturally and ventriloquized spoken. The CV sequences are analyzed in the spectral domain. For that purpose the spectral envelopes from overlapping frames of the speech signals are determined by linear prediction. To eliminate the spectral decrease of the voiced speech the frames are filtered by an adaptive pre-emphasis. The analyzed frames contain exactly two speech periods with an overlapping of one period; the periods of the speech signals are marked. The resulting spectral envelopes up to 6 kHz are depicted in fig 7. The lower magnitude responses of the figure correspond to the consonant whereas the upper magnitude responses tend to the vowel /a:/. In the case of /b/ it can be seen that the second formant has practically disappeared for /b'/ whereas the first and third formant correspond to those of /b/. This can be explained by an excessively wide bandwidth of the second formant of /b'/. The same situation exists for /p/. The spectral envelopes of the sounds /v/ and /v'/ in fig. 7(c) and (d) are shown. The first and

second formant are comparable, however, the frequency of the third formant of /v'/ is slightly higher than the formant frequency of /v/. This effect is more pronounced by the results of the articulatory synthesis simulations.

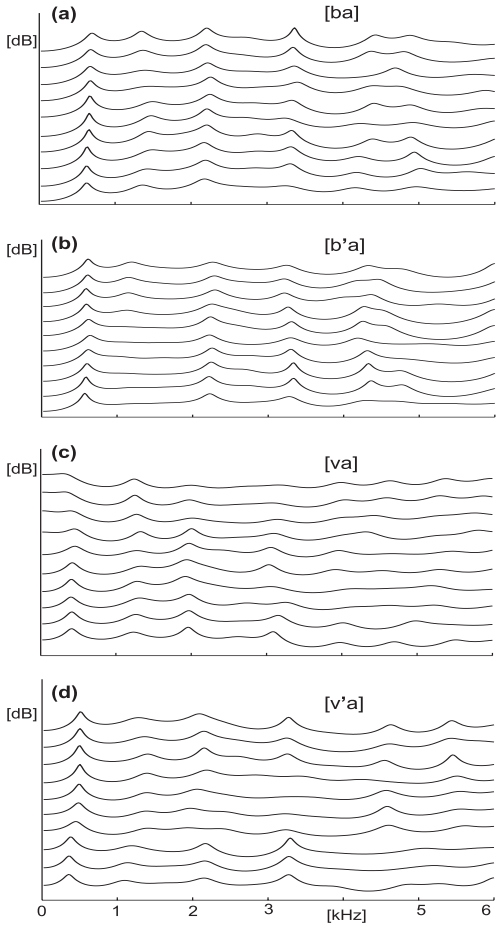

Figure 7: Spectral envelopes obtained by linear prediction (with pre-emphases) of utterances: CV sequences naturally uttered (a) for [ba], (b) for [b'a], (c) for [va], and (d) for [v'a].

2.5 Comparison of simulated and natural speech

A comparison of the results obtained by the synthesis of sounds and the analysis of real speech shows that for respective sounds the reproduction by their substitutions is perceptively comparable. The substitutions /b'/ and /v'/ are relatively close to /b/ and /v/ whereas the nasal /m'/ is perceptively between /n/ and /m/ but much closer to /m/. This is equally valid for the simulations and for the ventriloquist's isolated CV utterances. A perceptive inspection of substitute sounds which are embedded in sentences and uttered by the ventriloquist shows that the substitution of /m/ is a difficult task, so that occasional appearances of a substituted /m/ are perceptively close to /n/. One explanation is given by the higher articulatory effort of the production of the respective substituted sounds.

Moderate dissimilarities between the simulation and real speech are given by the partially different formant constellations of the substituted sounds, especially for /b/. This can be caused by different techniques of ventriloquists to generate the substitute sounds. The spectrum of the sounds cannot be perfectly reproduced by their substituted counterparts. An interesting point is that the strategies are different in the case of /b/. Whereas in the simulation the first two formants are reproduced at the expense of a higher third formant in the real utterances the intensity of the second formant is significantly reduced, beyond recognition.

Both the results of the simulations and the analysis of real speech indicate that by the substitute sounds the lower frequency range is better reproduced than the higher frequency range. This corresponds to the relevance for human speech recognition.

3. CONCLUSIONS

Substitute sounds are generated and analyzed in various manners. Furthermore, new substitution sounds were found. The degree of achieved perceptive similarities between the sounds and their substituted counterparts are unexpectedly high. Nevertheless differences in the spectral domain exist, especially in the higher frequency range. For a successful substitution the relevant spectral information should be reproduced as precisely as possible and the introduction of confusing spectral information should be avoided.

It was shown by simulation and by real speech recordings, respectively, that it is generally possible to replace certain sounds by perceptively equal substitute sounds. This may lead to new perspectives in the treatment of dysarthria patients.

4. ACKNOWLEDGEMENTS

The authors wish to thank especially the professional ventriloquist Patrick Martin for ventriloquial recordings and Peter Birkholz (University Rostock) for support concerning the articulatory synthesis.

This project has been awarded the German Federal President's prize for young researchers.

5. REFERENCES

[1] P. Ladefoged and I. Maddieson: "The Sounds of the World's Languages", Blackwell Publishers, Oxford UK & Cambridge USA, 1996.

[2] P. Birkholz, D. Jackel, and B.J. Kröger: "Construction and control of a three-dimensional vocal tract model" Proc. IEEE Conf. ICASSP-2006, Toulouse France.

[3] V. Vox: "I can see your lips moving", Retonios Magic, Casino, Switzerland.

[4] G. Fant G., "Acoustic Theory of Speech Production", Mouton, The Hague – Paris, Second Edition 1970.

[5] B.H. Story et al. "Vocal Tract Area Functions from Magnetic Resonance Imaging", J. Acoust. Soc. Am. Vol. 100 (1996), pp. 537-554.

[6] K. Schnell and A. Lacroix, "Parameter Estimation of Branched Tube Models by Iterative Inverse Filtering", Proc. IEEE Conf. DSP-2002, Santorini Greece, Vol. I, pp. 333-336, 2002.