Techniques for Estimating Vocal Tract Shape for Speech Training Aids

ShilpaChaman
Assistant Professor
Electronics and Telecommunication Department,
St. Francis Institute of Technology, Mumbai,India

Abstract :The main goal of this paper is to review various techniques for mapping acoustical properties of speech to the geometry of vocal tract which is helpful in speech synthesis, speech recognition, coding, music control and in speech training aids. If the set of time-varying articulatory parameters are known then synthesis of speech from them is known as a direct problem. This direct problem is well understood but the estimation of vocal tract geometry from input speech known as inverse problem is still difficult to understand because of the non-uniqueness of acoustic to articulatory mapping. Different techniques used in speech training aids are discussed which may provide visual feedback of articulatory efforts to hearing impaired people. This may help them to overcome their speech disability.

IndexTerms **- Articulatory synthesis, Vocal tract estimation, Speech training aids**

I. INTRODUCTION

Articulatory synthesizer is a model for human speech production from articulatory parameters like lung pressure, jaw angle, nasality, tongue movement, velum opening, lip opening etc.In articulatory speech mimic, natural speech is generated in an articulatory synthesizer where methods for estimating its control parameters are considered. If the set of time-varying articulatory parameters like geometry of vocal tract and glottis are known, then synthesis of speech from them is known as a direct problem. This direct problem is well understood but the estimation of vocal tract geometry from input speech known as inverse problem is still a challenge because of the non-uniqueness of acoustic to articulatory mapping. This has attracted many researchers to estimate the articulatory parameters from acoustic information obtained from speech and to display these features for various purposes like speech training aids for hearing impaired people, musical control, text-to-speech synthesis, synthesis of best quality speech from the recovered shapes, for coding etc.

In hearing impaired people in spite of having proper speech production mechanism they are not able to speak because of the absence of auditory feedback. The hearing impaired often tries to speak by visualizing lip movements but they are not able to understand proper articulation. Several speech training aids are developed which may provide visual feedback of articulatory efforts to hearing impaired people. This may help them to overcome their speech disability.

This paper is structured as follows. Section II describes a simplified acoustic model of vocal cord and vocal tract. A review various techniques for speech synthesis and acoustic to articulatory mapping is discussed in section III, followed by a discussion on methods for improving the estimation of vocal tract shape in Section IV. A visual model for speech training aids is mentioned in Section V and Section VI concludes this paper and also provides future scope.

II. ACOUSTIC TUBE MODEL OF VOCAL TRACT

Human acoustic system of vocal tract and chord [1] is depicted in Fig. 1. The vocal tract can be considered as a non-uniform tube with varying cross sectional areafrom zero to 20 $cm^2$and of length 17 cm approximately. When a person speaks, the subglottal air pressure is applied, which leads to the oscillations of the vocal cord model and results in the glottal volumevelocity.Sound is radiated from the system which results in volume velocities at the mouth and nostrils.

Cross-dimensions of the human acoustic model are small as compared to sound wavelengths, therefore planar wave motion can be confined in the tract .The vocal tract can be approximatedas a variable-area tube when it is straightened out and hence the linear wave equation is valid. The cross-sectionalarea as a function of position along the tract, with x = 0at the glottis end of thetract, completely specifies the shape of the vocal tract. This is denoted by the area function, $A(x)$. Assuming no viscous or thermal losses, the pressure, $P(x,s)$ and the volume velocity, $U(x,s)$ in the tube satisfy the pairof first order differential equations (1), (2).

$$\frac{dP}{dx} = -\frac{\rho s}{A}U \qquad (1)$$

$$\frac{dU}{dx} = -\frac{As}{\rho c^2}P \qquad (2)$$

where,sis the complex frequency variable, ρ isthe density of air, and c is the velocity of sound. By Webster's Horn equation [2], the volumevelocity can be substituted from (1), to yield equation (3).

$$\frac{d}{dx}A\frac{dP}{dx} - \frac{s^2}{c^2}AP = 0 \qquad (3)$$

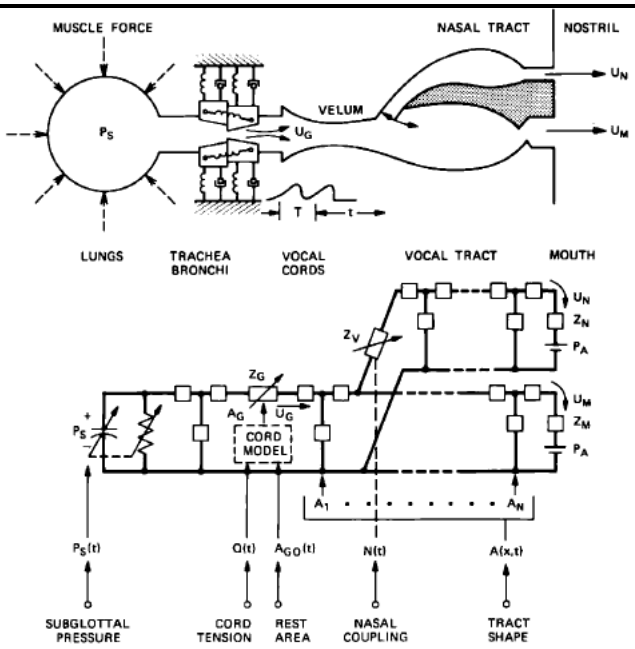

Fig. 1. *Acoustic tube model of human vocal tract*

Similarly, an equation for U(*x*,*s*) alone can be derived by eliminatingP(*x*,*s*). These equations relate pressure and volume velocity directly to area function. Equation [3] is modified to equation [4], if thermal and viscous losses M(*x*,*s*)and wall impedance N(*x*,*s*) are taken into consideration.

$$\frac{d}{dx}M(x,s)\frac{dP}{dx} - N(x,s)P = 0 \qquad (4)$$

Here the functions M(x,s) and N(x,s) can be computed in terms of A(x).

III. TECHNIQUES FOR SPEECH SYNTHESIS AND ESTIMATION OF VOCAL TRACT SHAPE

There are two main problems in the area of speech production. One is the direct problem of speech synthesis from a given set of time-varying articulatory parametersand other is the inverse problem of estimating vocal tract geometry from the natural input speech.The second problem is relatively difficult because of the non-uniqueness of acoustic to articulatory mapping.

3.1Direct Problem

Speech signal can be synthesized if the articulation information, like the area function *A(x)*, the wall impedance, and the loss parameters of the vocal tract, are specified. Then equation (1) or (2) canbe solved for any given boundary conditions at the lips andglottis. Thus, speech signal for a variety of sounds can be generated if a proper choice of boundary conditions is done.

If the case of computation of non-nasalized vowelsounds is taken with boundary condition at the lips is such that the tract is terminated with the radiation impedance, $Z_L(s)$. The solution for the pressure in the tract which satisfy this boundary condition is $H_p(x,s)$ and thevolume velocity at the glottis is unity. Let $H_u(x,s)$ be thecorresponding volume velocity. Then the volume velocity inthe vocal tract due to any other input $U_g(s)$ at the glottis is

$$U(x,s) = H_u(x,s)U_g(x,s) \qquad (5)$$

In particular, the volume velocity at the lips is obtained bysetting x = L, the length of the vocal tract. The function$H_u(L, s)$ is called the transfer function of the tract. Then the speech signal in the frequency domain is

) (6)

The inverse Laplace transform of S(L, s)of the function give the time domain speechsignal.Flanagan, Ishizaka, and Shipley in [4] created such type of articulatory speech mimic systems. Further they put a closed optimization loop around their articulatoryspeech synthesizer in [5] by comparing thespectra of the synthesized speech with given spectra of consecutivetarget speech frames. For eachframe, an optimization procedure tried to minimize an acousticdistance between the two speech signals, thus, in effect,estimating articulatory parameters by an analysis-by-synthesisprocedure. Further on these lines Schroeter et al. continued and created a new articulatory synthesizer [6],and an articulatory speech mimic [7]. Elsewhere, similarapproaches were taken (e.g., [8], [9]).

A major problem in articulatory analysis-by-synthesis procedure is the initialization of the optimization loop. One needto choose good startup parameters since mostoptimization algorithms will only find the local minimum of agiven cost function that is near the initial parameters. This can be achieved byemploying an acoustic-to-articulatory mapping. One possiblerealization of such a map is called articulatory codebook.

Articulatory Codebook: It is a table of corresponding acoustic and geometric vectors [10]. The acoustic representation is given as a key to look up (retrieve) the associated vocal-tract shape. Such articulatory codebooks provides a good set of start-up vectors for global optimization. In fact, if the codebook-lookup were good enough, one might avoid the iterative optimization altogether.

Non-Uniqueness: It can be seen that the acoustic input impedance of the tract uniquely specifies the area function while the transfer function does not. Two kinds of non-uniqueness can be defined. The first kind is due to the fact that different tract shapes may have (almost) the same transfer function. The second kind arises from the fact that the same speech spectrum may be produced by two different tract shapes with appropriately selected inputs at the glottis (vocal cords). Both types of non-uniqueness have to be

dealt with in an articulatory analysis/synthesis system.Direct problem of speech synthesis is well understood but the inverse problem of estimating vocal tract geometry from natural input speech is difficult because of the non-uniqueness of acoustic to articulatory mapping. In order to show this non-uniqueness lets discuss the inverse problem.

3.2 Inverse Problem

They employ techniques to estimating articulatory information especially the vocal tract area function A(x) using acoustic measurements i.e. from the analysis of the speech signal, but this inverse problem is slightly ambiguous because of the non-uniqueness of acoustic-to-articulatory mapping. The vocal tract shape can be computed using numerous direct and indirect methods.

3.2.1 Direct methods

In these methods by exposure to electromagnetic waves, the vocal tract shape is acquired by extractingits various articulatory features. Following are the various direct methods:

3.2.1.1 X-Ray

In this traditionally method [11],the movement of vocal tract is captured on high speed films using X-Ray beamsand it provides the best view the for speech research, but it is no longer in practice due to its possible harmful effects.

3.2.1.2 X–Ray Microbeam(XRMB)

In this method [12] the vocal tract shapes are estimated using a narrow beam of high energy X-Rayswhich track the motion of gold pellets glued at certain positions on the tongue, jaw, lips, and soft palate. This provides details of articulators in the mid-sagittal plane (side–view) and is also known as an articulograph. User can also control the rate of display of the articulograph, and it also displays the pitch, RMS trace of the audio signal, and spectrogram.

3.2.1.3Multi-Channel Articulatory (MOCHA)

In this method [13] both Electro Magnetic Articulograph (EMA) and Electro Palatograph (EPG) are used for various utterences.The EMA provides details of mid-sagittal plane similar to the XRMB sampled at a rate of 500 Hz. It uses 6 pellets to track the motion of vocal tract. The EPG, on the other hand, provides tongue–palate contact details at a sample rate of 200 Hz. It uses 62 contacts which are distributed along 8 rows over the upper palate. Contact made by the tongue on any one of the contact causes that particular contact to be shaded in the display. The contact made in the first three rows, the next two and the last three indicates an alveolar, palatal, and velar contact respectively in an utterance. There is a provision to simultaneously listen to the audio recording and observe articulatory data for these sentences.

3.2.1.4Ultrasound imaging

This method [14] is based on the application of ultrasound which produces an image by using the reflective properties of sound waves. Although its a non-invasive method without any harmful effects on the speaker but the images produced tend to be very noisy and estimation of places of articulation is prone to error especially for the base or the tip of the tongue.

3.2.1.5 Magnetic Resonance Imaging (MRI):

In this methodMRI is used for vocal tract area measurements directly [15]. A 3D volume could be constructed by scanning over a period of around 65 minutes and taking 26 slices. The area then can be estimated by counting the 3D voxels in a given section.Themajor drawback of this method is that the speaker has to be in supine position when articulating and also have to sustain the articulatory position during the scanning. This may cause speaker to get tired. Also the absence of simultaneous speech recordings for corresponding articulatory efforts hinders this method's usefulness for validation of vocal tract shape estimation.

3.2.2 Indirect methods

They employ various time-domain and frequency domain methods to estimate the vocal tract shape using acoustic measurements or from the analysis of the speech signal.

3.2.2.1 Time Domain Methods

One of the acoustic measurements for vocal tract shape estimation involves measurement of the acoustic impedance at the lips [16] by using a long impedance tube. The speaker has to articulate without phonation,and hence this method cannot be used for speech training.

Another method was proposed by Sondhi and Gopinath [17] which is based on the time domain specificationof the input impedance, $z_{in}(t)$. They showed thatthere is a unique one-to-one correspondence between $z_{in}(t)$ for $0 < t < T$ and A(x) for $0 < x < cT/2$. Further, themethod can be generalized to include the effect of losses andyielding walls [18], [19], provided that these losses are known.However, this method is notuseful for deriving A(x) from the speech signal, because oneneeds to make a measurement of the input impedance.

3.2.2.2 Frequency Domain Methods

Almost 65 years ago Borg [20] considered an ideal, lossless vocal tract and proved a remarkable result that allows computation of area function from the knowledge of certain sets of eigenvalues of boundary value problems associated with (2).

Ladefoged et al. [21] estimated vocal tract shapes from formant frequencies. The first three formants were extracted from speech and used for vocal tract shape estimation of vowels.As different vocal tract shapes may correspond to the same set of formant frequencies,constraints were imposed to exclude the vocal tract shapes which are not physically possible.

Methods based on inverse filtering of speech signal generally used linear predictivecoding (LPC) proposed by Wakita [22]. The method modelled vocal tractas a lossless acoustic tube with equal-lengthsegments of varying cross-sectional areas as shown in Fig.2. The analysis givesreflection coefficients which are used to obtain the area ratios at the section interfaces using the relation

$$\frac{A_i}{A_{i+1}} = \frac{1 + r_i}{1 - r_i} \qquad (7)$$

where, A_i is the area of i^{th} section and r_i is the reflection coefficient at the section interface of A_i and A_{i+1}.

Generally, the scaling iscarried out by assuming the glottis end of the vocal tract to have a normalized area of unity which is used as thereference area for scaling.

Fig. 2. *Vocal tract modelled as sections of equal length and varying area*

However, the change in the area at the glottis end during speech production introduces gross errors duringdynamic estimation of the vocal tract shape, as also seen in MRI images[23] for different utterances. Hence the assumption of a constantreference area at the glottis end cannot be considered.

IV. METHODS FOR IMPROVING ESTIMATION OF VOCAL TRACT SHAPE

The proposed solution to improve the LPC based estimation done by Wakita [22], is to use the area of mouth opening i.e. the inner lip contour area as the reference area for scaling purposes. Nayak et al. [24] estimated the required inner lip contour area from the video recording of speaker's face during speech utterance. The points corresponding to lip opening were manually marked and joined using straight line and the number of pixels within the polygon formed was used as its area. This was repeated for all frames of the video. The area values were normalized by the area obtained for the largest opening, which occurs during the utterance of vowel /a/. It was reported that the scaling of the vocal tract using the area of mouth opening resulted in better estimation of vocal tract shape area compared to the one obtained by using a constant reference area.

Jain et al. [25] continued on these lines and developed an image processing techniquebased on colour transformation andtemplate matching for consistent and accurate detection of the inner lip contour.This technique was robust against variations in illumination, skin hues across speakers and also not affected by the presence of tongue and teeth. It is compared with reference to manually estimated values and the results were much better. Block diagram of the technique used for detection of inner lip contour is depicted in Fig.3.

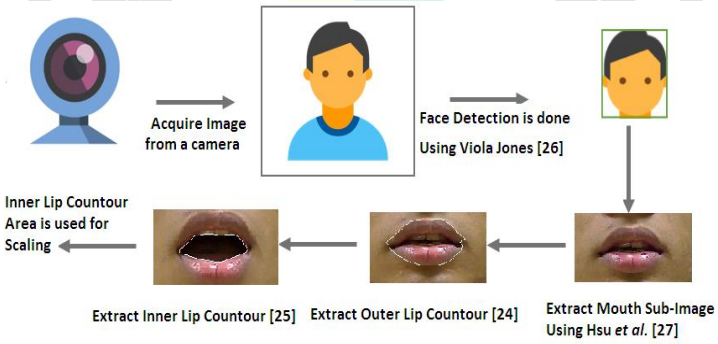

Fig. 3. *Block diagram for inner lip contour area detection*

V. VISUAL SPEECH TRAINING AID

Normal people can acquire the ability to control various articulators parameters like lung pressure, jaw angle, nasality, tongue movement, velum opening, lip opening etc.by the age offour since they receive both visual and auditory feedback. However, hearing impaired people do not have access tothe auditory feedback and hence they are not able to speak, in spite of having proper speech production mechanism. They have neither auditory loop nor any remembrance of speech by themselves.Lip reading technique also fails, since vowels & consonants with tongue movement hiddenin the mouth are not distinguishable to them by simply visualizing lip movements.

Speech-training systems can be designed based on visual or tactile feedback of acoustic parameters suchas speech intensity, fundamental frequency, spectral features or based on feedback of articulatoryparameters such as voicing, nasality, lip & vocal tract movement [28]-[29]. The tactile feedback is difficult to understand, delayed and unnatural whereas visual speech training aid provides better feedback as the person's voice and articulation can be immediately shown on the computer display. This way, hearing impaired person would be able to evaluate and correct their utterance or pronunciation based onexpected and actual parameters that are displayed to him. Like, they can compare the articulation of their vocal tractshapes with the reference articulation and suitably correct their articulation defects.

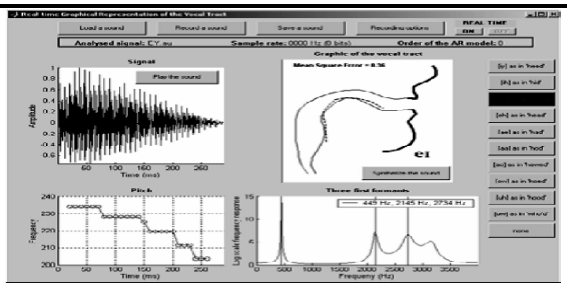

Fig. 4. *Speech training aid based on visual feedback*

VI. CONCLUSION AND FUTURE SCOPE

In this paper a review of various available techniques for mapping acoustical properties of speech to the geometry of vocal tract is done. After a brief introduction of the acoustic system of vocal tract, the direct and inverse problems are discussed. This led to the discussion of the important issue of non-uniqueness, that is, more than one tract shape can produce a given tract transfer function. Various ideas for alleviating this ambiguity are discussed. Also a discussion on how to reduce errors in estimated vocal tract shape, caused by variation in the area at glottis end is done. By considering lip area as a reference for scaling instead of glottis area, this problem can be solved. Finally the model of visual speech training aid is presented.

It remains to be seen if novel and improved applications of image processing and neural network are considered, the lip area estimation may further be improved and it may provide significantly better mappings than the other approaches.

REFERENCES

[1] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speechfrom a dynamic model of the vocal cords and vocal tract," Bell Syst. Tech. J., vol. 45, no. 3, pp. 199-229,1975.

[2] J. L. Flanagan, Speech Analysis Synthesis and Perception, 2nd ed. New York: Springer, 1972.

[3] M. M. Sondhi, "Model for wave propagation in a lossy vocal tract," J.Acoust. Soc. Am., vol. 55, no. 5, pp. 1070-1075, 1974.

[4] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Signal models for lowbit-rate coding of speech," J. Acoust. Soc. Am., vol. 68, no. 3, pp. 780-791, 1980.

[5] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speechfrom a dynamic model of the vocal cords and vocal tract," Bell Syst. Tech. J., vol. 45, no. 3, pp. 19S229, 1975.

[6] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domainarticulatory speech synthesizer," IEEE Trans. Acoust.. Speech, Signal Processing, vol. ASSP-35, no. 7, pp. 955-966, July 1987.

[7] J. Schroeter, J. N. Larar, and M. M. Sondhi, "Speech parameterestimation using a vocal tractkord model," in Proc. IEEE Inr. Con$ Acoust., Speech, Signal Processing, vol. 1, 1987, pp. 308-311.

[8] S. Maeda, "A digital simulation method of the vocal-tract system,"Speech Communicat., vol. 1, pp. 199-229, 1982.

[9] P. Meyer, R. Wilhelms, and H. W. Strube, "A quasiarticulatory speech synthesizer for the German language running in real time," J. Acoust.Soc. Am., vol. 82, no. 2, pp. 523-539, 1989.

[10] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," IEEE Trans. Acoust.. Speech, Signal Processing,vol. 36, no. 12, pp. 1812-1818, 1988.

[11] G. Fant, Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

[12] J. R. Westbury. X-ray Microbeam Speech Production Database User's Handbook (Version 1.0), June 1994. [Online]. Available: https://files.nyu.edu/ag63/public/fhs atelier/ubdbman.pdf (Last accessed in May, 2011).

[13] A. A. Wrench. "MOCHA multichannel articulatory database," 2008.[Online].Available:http://data.cstr.ed.ac.uk/mocha/README v1.2.txt (Last accessed in May, 2011).

[14] [B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2004, vol. 1, pp. I–685–688.

[15] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am., vol. 100, no. 1, pp. 537–554, 1996.

[16] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," J. Acoust. Soc. Am., vol. 41, no. 4, pt. 2, pp. 1002–1010, 1967.

[17] M. M. Sondhi and B. Gopinath, "Determination of vocal tract shapefrom impulse response at the lips," J. Acoust. Soc. Am., vol. 49, no. 6,

[18] M. M. Sondhi and B. Gopinath, "Determination of the shape of a lossyvocal tract," in Proc. Seventh Int. Congr. Acousr. (Budapest, Hungary), 1971.

[19] J. R. Resnick, "Acoustic inverse scattering as a means for determiningthe area function of a lossy vocal tract: theoretical and experimental model studies," Ph.D. dissertation, Johns Hopkins Univ., Baltimore,MD,1979.

[20] G. Borg, "EineUmkehrung der Sturm-LiouvilleschenEigenwertaufgabe," An inversion of the Sturm-Liouville eigenvalue problem," in German) ActaMathemutica, vol. 78, pp. 1-96, 1946.

[21] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," J. Acoust. Soc. Amer., vol. 64, no. 4, pp. 1027–1035, 1978.

[22] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acousticspeech waveforms," IEEE Trans. Audio Electroacoust., vol. AU-21, no. 5, pp. 417–427, 1973.

[23] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammart, J. Kim, S.Lee, K. S. Nayak, Y. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P.K. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research, " J. Acoust. Soc. Am., vol. 136, pp. 1307-1311, 2014.

[24] N. S. Nayak, R. Velmurugan, P. C. Pandey, and S. Saha, "Estimation of lip opening forscaling of vocal tract area function for speech training aids, " in Proc.18th National Conf. Commun., Kharagpur, 2012, pp. 521-525.

[25] S.Jain, P.C.Pandey, and RajbabuVelmurugan,"Lip Contour Detection for Estimation of Mouth Opening Area, " in Proc.5th National Conf. on Computer Vision,Pattern Recognition, Image Processing and Graphics,Patna, 2015.

[26] P. Viola and M. Jones, "Robust real-time face detection," Int. J. of Computer Vision, vol. 57, no. 2, pp. 137-154, 2004.

[27] R. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 696-706, 2002.

[28] C. S. Watson, M. Elbert and G. DeVane,"The Indiana Speech TrainingAid(ISTRA)", J. Acoust. Soc. Am., Vol. 81, Issue S1, pp.95, 1987.

[29] A.M. Oster, "Auditory and visual feedback in spoken L2 Teaching", Reports from the Dept of Phonetics, Umeå University, PHONUM 4, 1997.