THE EFFECT OF VOCAL DISGUISE ON SOME VOWEL FORMANT FREQUENCIES

Lynda Penny

Department of Speech Pathology
Flinders University of South Australia

Ronnie Grace and Karly Winkler

Heathfield High School
South Australia

ABSTRACT - The effect of assuming a vocal disguise on the formant frequencies of some vowels is examined. The material used was that of an authentic vocal line-up in which the identity of the speaker was in question. The subjects assumed commonly adopted vocal disguises

INTRODUCTION

People who feel that their personal safety or the security of their property is at risk may employ a recording device (video- or audio-) which may firstly act as a deterrent or later as an aid in apprehending an assailant, should the threat be realised. Recording instruments may also be used by the police and other agencies in surveillance, in the investigation and prevention of crime. Of course people whose activities may be the subject of such recordings are often aware of the possibility and may assume a disguise to hide their identity. Recordings may be made surreptitiously, in less than ideal conditions and may be of such poor quality that the identity of the speaker (and the content of the message) may be open to doubt. So it is that the testimony of the expert witness is called upon, to resolve uncertainty, to penetrate the disguise and validate suspicions as to the identity of the perpetrator of an action. The examination of objective recordings by expert witnesses is attractive in forensic work, since studies into the reliability of eyewitness (and earwitness) accounts, even when these are available, are not reassuring. Whether experts are in fact able to extract patterns from the speech signal which reliably identify a speaker who is attempting to avoid identification is a question still to be answered. This paper is a preliminary attempt to identify what happens to some aspects of the speech signal, namely the formants of some vowels, when a vocal disguise is adopted. Vowel formants were selected for measurement as they are such an obvious feature of the spectrogram, and because a great deal is known about them in natural speech. In particular, it was of interest to see whether the upper formants (F3 and F4) remained more or less constant over disguise, as these are usually assumed to reflect in part personal characteristics of the speaker, such as bone density and the shape of bony cavities, and should therefore resist major changes brought about by altering vocal tract shape in the course of assuming a disguise.

SUBJECTS

Two healthy young adults, one female and one male, both native born speakers of general Australian English prepared a tape for analysis.

MATERIALS

The subjects read text which had been the material used in an authentic 'voice line-up':-

I'll have that thanks.
Money in the bag.
I want to buy a bracelet for my wife.
I don't want one with a padlock.
Can I have a look at the diamond rings.
Don't be silly.
Get the pads and put them in the bag, and the money.
Put the money in the bag.

PROCEDURE

The subjects read the text in their normal voice, (N), assuming natural, if somewhat urgent, intonation.
Then in turn they adopted six vocal disguises which were selected to copy what people commonly do when disguising their voice. It was also hoped that the disguises would impact on vowel formants in clear-cut ways, so that the 'take-up' of the independent variable on the dependent variable could easily be seen. The disguises were:-

Helmet (Hm) A motor-bike rider's full helmet, very well stuffed, was worn in an effort to dampen resonances from the sinuses and bony structures of the face and skull. It was predicted that this might affect the higher formants more than the lower.

Foreign accent (A) The subjects adopted a 'foreign accent' that they felt comfortable with. In the case of the young woman this was a Southern US. accent and in the case of the young man it was a Scottish accent. It was simply assumed that the vocal tract resonances would be changed with the imitation of different vocal patterns.

Stairwell (S) Subjects bellowed into the microphone in a highly echoing concrete stairwell that extended up 4 floors. The common use of mobile telephones makes sending messages from such a location possible. It was assumed that the extra effort involved in making the transmission would alter all formants.

Helium (He) The well known effect of helium on the voice was extended to include an examination of its effects on vowel formants. Subjects inhaled by mouth the contents of a helium-filled balloon.

Creaky voice (C) After practice, the subjects assumed creaky voice (vocal fry) which dramatically changes fundamental frequency, though it is not entirely clear what effect this has on formants.

Tube (T) The subjects talked down a thick-walled hard cardboard tube, 2.5cm in internal diameter and 21.5cm long, thus effectively more than doubling the length of their vocal tract and consequently introducing additional formants.

EQUIPMENT

The tapes were made with a Sony DAT portable tape recorder and a Sony microphone, ECM-959DT.

The formants were measured from spectrograms displayed by the Kay DSP5500 Spectrograph. Single target vowels in stressed position for which at least two tokens were available were selected for measurement. Mean values were calculated for these tokens.

RESULTS

Four vowels, /æ/, /ʌ/, /ɪ/, /ʊ/ met the criteria for inclusion.
The results for each subject are displayed in Tables 1 and 2

DISCUSSION

Formant values for each subject's normal voice are compatible with the ranges suggested by Bernard (Bernard and Mannell, 1986) for men and by Penny (Penny, 1992) for women. Comparing formant values in disguised voice against normal voice it can be seen that:-

1. The helmet condition changes things very little.
2. For Subject 1 (young woman) the assumption of a foreign accent (Southern US) spread the range of values with F3 and F4 considerable higher. For S2 (young man) a Scottish accent produced a generally downward shift in F1 and very little change in F2, F3 or F4.
3. The stairwell condition elevated F1 but there is no clear effect discernible on the other formants.

Table 1: Subject 1(female) Formant values for natural voice and 6 voice disguises

Vowel /æ/

	N	Hm	A	S	He	C	T
F1	756	672	728	804	1000	888	632
F2	1888	1792	2028	1952	1796	1952	1216
F3	2800	2760	3184	3096	2504	3234	1888
F4	4004	3832	4480	4296	3125	4688	2708
F5							4128

Vowel /ʌ/

	N	Hm	A	S	He	C	T
F1	780	700	540	780	750	1000	630
F2	1680	1680	1600	1460	1350	2320	1070
F3	2960	2760	3600	3100	1960	3620	1480
F4	4140	4140	4980	4140	2410	5160	2540
F5							3360

Vowel /ɪ/

	N	Hm	A	S	He	C	T
F1	430	540	480	500	540	920	500
F2	2260	2220	2800	1660	1030	2020	1200
F3	2960	3000	3240	2780	2610	3800	2120
F4	3820	3960	4460	4020	2890	5295	2760
F5							3880

Vowel /ʊ/

	N	Hm	A	S	He	C	T
F1	490	580	520	580	610	900	520
F2	1420	1360	1620	1440	1220	2120	1030
F3	2360	2420	3080	2020	1960	3260	1360
F4	3570	3420	4340	2880	3190	4780	2420
F5							3490

Subject 2 (male) Formant values for natural voice and six voice disguises

Vowel /æ/

	N	Hm	A	S	He	C	T
F1	625	660	668	732	884	732	652
F2	1588	1636	1192	1688	1294	1572	1464
F3	2836	2620	2792	2860	2964	2932	2076
F4	3636	3580	3564	4184	2892	3644	2772
F5							3628

Vowel /ʌ/

	N	Hm	A	S	He	C	T
F1	640	640	550	670	750	1000	610
F2	1410	1410	1390	1470	1350	2320	1310
F3	3000	3000	2660	2910	1960	3620	2040
F4	3590	3870	3400	4220	2410	5160	2850
F5							3656

Vowel /ɪ/

	N	Hm	A	S	He	C	T
F1	440	440	420	560	540	920	400
F2	2000	1980	1920	1560	1030	2020	1940
F3	2570	2870	2670	2700	2610	3800	2100
F4	3690	3690	3785	4080	1960	5295	3130
F5							3880

Table 2 cont
Vowel /ʊ/

	N	Hm	A	S	He	C	T
F1	480	480	410	500	550	460	510
F2	1270	1270	1520	1470	1120	1370	1210
F3	2740	2740	2720	2560	1670	2260	2020
F4	3570	3570	3620	3620	3110	3570	2730
F5							3550

DISCUSSION (cont)

4. Helium seems to have had an elevating effect on F1 and F2 but a depressing effect on F3 and F4.
5. Creaky voice has the rather surprising effect of elevating all formants.
6. The tube, by effectively increasing the length of the vocal tract, changes its resonating characteristics markedly. A second formant is introduced lower than the natural F2, and F3, F4, and F5 are all lower, as would be expected.

CONCLUSION

If you wish to disguise your voice, and assuming that an expert phonetician will pay regard to formant frequencies, use a tube to extend the length of your vocal tract. Do not bother with a helmet. There is no evidence that F3 and F4 resist change more than the other formants.

REFERENCES

Barnard, J.R.L. & Mannell, R.H. (1986) *A study of /h-d/ words in Australian English* in Working Papers, a report of speech, hearing and language research in progress (School of English and Linguistics, Macquarie University: Sydney)

Penny, L. (1992) *Acoustic measurements of the diphthongs of women speakers of general Australian English* Proceedings of the Fourth Australian International Conference on Speech Science and Technology, (ASSTA: Canberra)

ENDNOTE

This project was arranged as part of the C.S.I.R.O. 'Young Scientist' programme. The organisers are to be thanked for introducing two young and enthusiastic students to the field of speech science.

ON-LINE SPEAKER ADAPTATION FOR HMM BASED SPEECH RECOGNISERS

B. Watson

Department of Electrical and Computer Engineering
University of Queensland

ABSTRACT – An investigation of a gradient-descent based training technique was performed for the on-line adaptation of hidden Markov models to new speakers in a speech recognition system. It was found to be successful for supervised speaker adaptation, improving the recognition performance on a 46 word task (alphabet, digits and control words) from 88.0% to 93.2% after adaptation with nine repetitions of each word. Unsupervised adaptation on the same task was unsuccessful. However, for an easier 20 word vocabulary, unsupervised adaptation improved the recognition performance from 97.7% to 99.0%.

INTRODUCTION

The Baum-Welch re-estimation procedure for hidden Markov models (HMMs) directly maximises the likelihood of the training observations given the model. It is suited to the batch estimation of model parameters. In order to update models as new data becomes available, a smooth on-line learning algorithm is desirable. Baldi and Chauvin have proposed such an algorithm, and have experimented with its use for HMM training in a molecular biology problem (Baldi et al., 1993). Baldi and Chauvin's approach to HMM training is based on the use of a gradient descent algorithm. Suppose we wish to maximise some function $\mathcal{L}$, that is dependent on a model parameter x. We can do this by calculating the gradient of the objective function $\mathcal{L}$, and making a small change (Δx) to the value of the parameter based on this. The value of the parameter x_n after training step n is:

$$x^n = x^{n-1} + \Delta x^n$$

where Δx^n is calculated using the partial derivative of the objective function with respect to the parameter, $\frac{\partial \mathcal{L}}{\partial x}$, evaluated at value x^n. That is:

$$\Delta x^n = \mu \left.\frac{\partial \mathcal{L}}{\partial x}\right|_{x^{n-1}} + \eta \Delta x^{n-1}$$

μ is a learning rate parameter, which controls the size of the parameter changes between training steps. It will affect both the speed of convergence, and the stability of the model parameter estimates. A momentum parameter, η, can also be introduced into the update calculations, in order to allow the use of a larger learning rate parameter, while still maintaining consistent updates, as it can provide partial averaging over the training observations (Hertz et al., 1991).

Consider an N state discrete output HMM with M possible outputs. Let the state of the model at time t be denoted as s_t, and the output at time t be denoted as o_t. The model parameters are: initial state probabilities, $\pi_i = Pr(s_1 = i)$; transition probabilities, $a_{ij} = Pr(s_{t+1} = j|s_t = i)$; and discrete output probabilities, $b_{ik} = Pr(o_t = k|s_t = i)$. The parameters of the HMM will be referred to collectively as λ. That is, $\lambda = \{A, B, \pi\}$, where A, B, and π, are the sets of transition, output, and initial state probabilities, respectively. Because all of these parameters are probabilities, when summed over the appropriate range, they will sum to 1.

For the HMM, we wish to update the model corresponding to a particular utterance, O, so that it maximises the likelihood of the model producing that utterance, $L = Pr(O|\lambda)$. Gradient descent on the log-likelihood is numerically better conditioned than gradient descent on the likelihood (Levinson et al., 1983) so the log-likelihood was used as the objective function for model training.

For both the transition and output probabilities, a normalised-exponential representation is introduced, to ensure that as probabilities they will have values between 0 and 1, and sum over the appropriate ranges to 1. The transition and output probabilities are written in terms of new parameters, w_{ij} and v_{ij} respectively, which are the values before normalisation:

$$a_{ij} = \frac{e^{\varphi w_{ij}}}{\sum_k e^{\varphi w_{ik}}}$$

$$b_{ij} = \frac{e^{\varphi v_{ij}}}{\sum_k e^{\varphi v_{ik}}}$$

φ is a parameter of the normalisation function which, can be absorbed into the learning rate.

The derivative of the log-likelihood with respect to w_{ij} and v_{ij} can then be derived (see (Baldi and Chauvin, 1994)) so that updates can be calculated. The derivatives are:

$$\frac{\partial \log L}{\partial w_{ij}} = \varphi[\sum_t \gamma_{t,i,j} - \sum_t \gamma_{t,i} a_{ij}] \quad (1)$$

$$\frac{\partial \log L}{\partial v_{ij}} = \varphi[\sum_t \xi_{t,i,j} - \sum_t \gamma_{t,i} b_{ij}] \quad (2)$$

where $\gamma_{t,i,j} = Pr(s_t = i, s_{t+1} = j|O,\lambda)$, $\gamma_{t,i} = Pr(s_t = i|O,\lambda)$, and $\xi_{t,i,j} = Pr(s_t = i, o_t = j|O,\lambda)$. These values can be computed using the forward and backward algorithms.

SUPERVISED ADAPTATION

Although our main interest in the proposed gradient descent technique is for on-line unsupervised learning, we began by analysing its behaviour for supervised adaptation of speech recognition models in order to determine: what models constitute the best prototype models, which are to be used as the starting point for speaker adaptation; how the performance of the adaptation algorithm depends on the amount of adaptation data; and what values of the learning rate and momentum parameters give the best results.

Experiments were performed on the Texas Instruments 46 word speech database. This database contains recordings from sixteen speakers - eight male and eight female. Each speaker has uttered the letters of the alphabet, the digits from zero to nine, and ten "control" type words (enter, erase, go, help, no, rubout, repeat, stop, start, yes). The data from each speaker has been divided into a training set of nine or ten utterances, and a test set of sixteen utterances of each word.

The TI-46 database was selected for experimentation because it represents a small but challenging task. Accurate alphabet recognition is difficult because the vocabulary contains a number of highly confusable words such as the e-set (B, C, D, E, G, P, T, V, Z (for American English)).

The speech waveforms from the TI-46 database were processed in a standard manner for input to the HMM recogniser. The speech waveforms in the database were sampled at 12.5 kHz with 12-bit quantisation. These waveforms were segmented into frames of 512 samples (40 milliseconds) with a new frame starting every 128 samples (10 milliseconds). The samples were pre-emphasised with a filter whose transform function is $1 - 0.95z^{-1}$, and Hamming windowed. For each frame, twelfth order LPC cepstral coefficients were calculated, along with first and second order delta coefficients.

Left-to-right hidden Markov models with ten states were used to model each of the vocabulary words. Supervised adaptation of each of the models, for each speaker, was performed by using each training utterance to update the relevant model, using values computed according to equations (1) and (2). In figure 1, the recognition performance following supervised adaptation is presented as a function of the learning rate and the amount of adaptation data used. The results presented were obtained when the starting models were speaker independent, but had only been trained using the data from the other speakers of the same gender as the target speaker. Genderless speaker independent models were observed to give recognition performances approximately 1% lower than those obtained using gender specific models. However, in both cases the adaptation improved the recognition performance substantially.

Adapting from the gender based speaker independent models, with a learning rate of between 0.18 and 0.26, the recognition performance improved from 88.0% to 93.2% following adaptation using nine