Moderating Strong Accents

Dr L Baghai-Ravary
Phonetics Laboratory, University of Oxford

ladan.baghai-ravary@phon.ox.ac.uk

Abstract

Modern labour-intensive communication-based industries, such as call centres, are increasingly outsourced to Asian countries where a dialect of English is widely spoken, and the pool of suitable staff is large. Despite the distances involved, this is highly cost effective, but is not without its drawbacks. In particular, when UK speakers hear a strong accent they often react negatively. This is partly due to the assumption that a strong accent normally indicates a lack of experience with, and a poor knowledge of, the language, and it may be felt that the call will somehow be delayed or its meaning misunderstood [1, 2].

Although the former assumption is usually unjustified, the latter may be true; unless the listener is attuned to the accent in question, it may be difficult to understand the speaker without frequent requests for repetition and clarification. This applies to the speech of both parties: both the UK-based caller and the Asian call centre operative. A difference in accents can significantly impair communication in both directions. An automated real-time system to reduce the misunderstanding between speakers with significantly different accents would be of great value to these industries.

1 - Introduction

This presentation examines the differences and similarities between strong non-native accents of Indian English and Southern British English, from an engineering and signal processing perspective. Both timing and acoustic factors are considered. It goes on to outline those aspects of speech which are primarily associated with accent, and those which determine “speaker identity”.

The practicality of separating accent from speaker identity will be discussed in some detail, along with the implications for any potential real-time system.

A preliminary system is then be proposed to ameliorate the perceived differences between the accents, so as to aid the comprehension by the listener, and improve the naturalness of their communication. The aim is not to completely remove all traces of accent, but simply to suppress or “moderate” the more extreme aspects of it. This should, in turn, allow the dialogue to proceed more quickly and with fewer requests for repetition, which would otherwise be frustrating to both parties to the conversation.

The speaker's voice characteristics are combined with a knowledge of how their accent compares to the target accent, to produce a signal with timing and acoustic characteristics more typical of the target, but still with many recognisable characteristics of the original speaker. This is different to what is conventionally termed “voice morphing”, as the speech preserves the personal characteristics of the speaker. The aim is for the speaker to change accent, but retain as much as possible of their own voice.

In the proposed method, the relationships between the characteristics in the two accent “domains” would be quantified using a statistical analysis of a database of speech collected from two subjects with similar voice characteristics, but with very different accents. The average of, and the degree of variability in, the power spectrum and the phoneme durations would be calculated and a mapping constructed between one accent domain and the other. The effects of differences in vocal tract length can be suppressed by vocal tract length normalisation. The result is a compromise between suppression of the original accent and retention of the speaker identity.

2 - Accents, Dialects, and Real-World Constraints

The phenomenon of human verbal communication is extremely complex and has defied all attempts at accurate and complete mathematical modelling. The levels of information communicated by a single sentence are many and varied. In just a few seconds of speech, the listener can learn not only the semantic content of the spoken words, but also various more-or-less reliable indications of many other factors including the emotional state, subconscious intent, state of health, age, gender, social class, level of education, and ethnicity of the speaker. Most of these auxiliary factors are conveyed by primarily non-verbal means.

The ultimate aim of this work would ideally be to separate out a subset of these non-verbal factors, and modify those related to accent while minimising any effect on the others. However it must be acknowledged that this is impossible to achieve completely and perfectly, given the current level of understanding of human speech production, perception, and intelligence in general. Nonetheless, some encouraging results are presented later in this paper, showing that although a perceived accent cannot (yet) be completely changed without either destroying intelligibility or introducing unacceptable delays into any conversation, it *can* be moderated so that it is perceived as more natural by a speaker of a different accent.

There follows a discussion of the various signs which identify any particular accent. For the purposes of this paper, we will define the term “accent” as pertaining to the articulatory or acoustic realisation of the words being spoken, but not the choice of the words themselves. Thus any difference between the speech of two speakers which manifests itself as a different vocabulary, or a different grammar, will be attributed to a difference in “dialect” rather than “accent”, and so will be outside the scope of this paper.

2.1 - Phonemics

The purely acoustic differences between accents can be characterised in terms of the shape and position of the articulators within the vocal tract. These manifest themselves in the broad spectral shape of the acoustic signals.

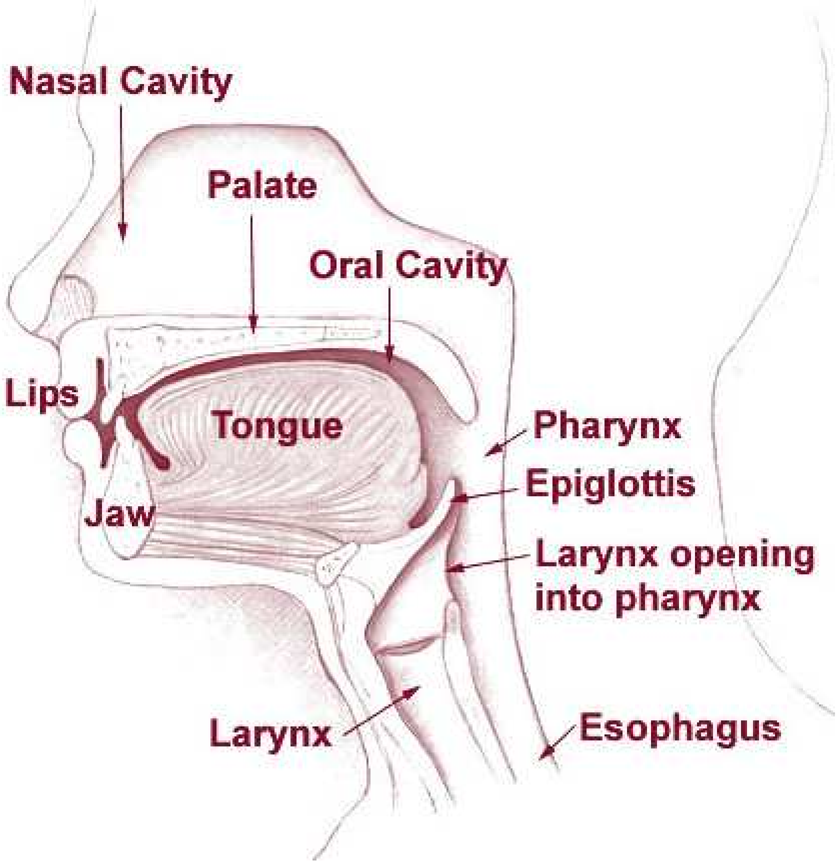

Figure 1: The Human Vocal Tract

(from http://training.seer.cancer.gov/head-neck/anatomy/overview.html)

There have been a number of studies of this aspect of the differences between accents of

English, and in particular between Indian English and British English [3, 4, 5].

It should be pointed out that most of these studies have taken British English as the “reference” accent, and reported ways in which the Indian accent is an deviation from it (i.e. in what contexts the British phonemes are “mispronounced” or “misinterpreted”). Relatively few studies have looked at the problem from the other side, and examined in what ways the British pronunciation differs from a standard Indian English reference. For that reason, most of the discussions in this paper will refer to the phonemes of Standard Southern British English, and will use the standard SAMPA notation [6], as shown in Table 1.

Consonants		**Vowels**		**Diphthongs**	
p	pen, spin, tip	A:	arm, father	eI	day
b	but, web	i:	see	aI	my
t	two, sting, bet	I	city	OI	boy
d	do, odd	E	bed	@U	no
tS	chair, nature, teach	3:	bird	aU	now
dZ	gin, joy, edge	{	lad, cat, ran	I@	near, here
k	cat, kill, skin, queen, thick	V	run, enough	E@	hair, there
g	go, get, beg	Q	not, wasp	U@	tour
f	fool, enough, leaf	O:	law, caught	ju:	pupil
v	voice, have, of	U	put		
T	thing, breath	u:	soon, through		
D	this, breathe	@	about, winner		
s	see, city, pass				
z	zoo, rose				
S	she, sure, emotion, leash				
Z	pleasure, beige				
h	ham				
m	man, ham		**Other symbols**		
n	no, tin	"	Primary stress,e.g. "happy" /"h{pi/		
N	singer, ring	%	Secondary stress, e.g. "battleship" /"b{tl=%SIp/		
l	left, bell	.	Syllable separator		
r	run, very	=	Syllabic consonant, e.g. /"rIdn=/ for ridden		
w	we				
j	yes				

Table 1: SAMPA Phonetic Symbols for Standard Southern British English

Summarising all the differences between “Indian English” and “Standard Southern British English” is made more difficult by the wide variety of dialects and accents of English to be found in India, even though many of them are perceived as being very similar by linguistically naïve British listeners. The number of native languages spoken within the Indian subcontinent is very large (29 languages with over 1 million native speakers, according to the 2001 Census of India [7]), and the diversity of these linguistic backgrounds ensures that both perception and production of many phonemes is far from uniform across all speakers.

However, in the remainder of this section we will list some of the most common and clearly audible differences which adversely affect intelligibility.

2.1.1 - Vowels and Diphthongs

The sounds which are commonly referred to as “vowels” include two distinct groups: those whose frequency content is approximately constant throughout, and those which have a frequency content which changes smoothly between the beginning of the sound and the end. From here on, we will refer to the former (steady) sounds as “vowels” or “monophthongs”, and the latter (dynamic) sounds as “diphthongs”.

- A number of discrete British English diphthongs and vowels can be used interchangeably in Indian English. For example /E/ (“b**e**t”) and /{/ (“b**a**t”), and /Q/ (“c**o**t”) and /O:/ (“c**augh**t”). This can be especially problematic where the lack of these distinctions introduces ambiguity to the meaning of the words.
- In some cases the diphthong /OI/ (“b**oy**”) can also be replaced by the vowel /{/ (“b**a**t”), and in some dialects, schwa (i.e. /@/), /3:/ (“b**ir**d”) and /V/ (“r**u**n”) are all pronounced as /A:/ (“f**ar**m”). In terms of spelling, a word-final “a” is invariably pronounced /A:/ in Indian English, whereas British English usually uses /@/.
- In British English, unstressed vowels are often reduced to a schwa (/@/ as in “**a**brupt”), whereas in Indian English they are usually fully articulated and sometimes stressed as well. For example “abrupt” may be pronounced /”eIbrVpt/ or /”EbrVpt/ rather than /@b”rVpt/. Similarly, the vowel /E/ is often replaced by the diphthong /eI/.
- Finally, some diphthongs in British English are replaced with monophthongs with a frequency content mid-way between that of the initial and the final portions of the British diphthong.

Most of these differences appear to be due to the relevance of the respective phonetic distinctions in British English to the native language of the Indian speaker. If the same (or similar) phoneme boundaries are significant in the native language, then the differences between British and Indian accents are reduced. If there is a significant phonetic boundary in British English but not in the Indian speaker’s native language, or vice versa, then the difference between the two accents tends to be larger.

2.1.2 - Consonants

With one or two exceptions, the differences between Indian and British English consonants are rather more difficult to express in terms to the British English phoneme set. That is because many of these differences appear to be due to mechanical aspects of the articulatory manoeuvres used to produce the sounds, rather than the inability of the speaker to perceive acoustic cues which may be important in the target accent, but not their own. That is to say that they cannot generally be expressed as a simple substitution of one British English phoneme for another; many involve substitution of the British phoneme with a sound which is not normally present in British English at all.

The consonant-related differences include:

- The production of the “liquids” /l/, /r/ and /w/ is strongly affected by the native language/dialect of the speaker. /w/ is frequently produced as /v/ or a similar phoneme from the speaker’s native accent. /l/ is generally articulated differently from the British English variants: the “dark L” is rarely used in Indian English, and the /l/ phoneme is frequently articulated further back in the mouth (i.e. it is more “velar”) than the British version. This also occurs with the /n/ phoneme.
- The syllabic consonants, /l=/, /m=/ and /n=/, are usually articulated as /@l/, /@m/ and /@n/ in Indian English.
- In Indian English, many fricatives are articulated as “obstruents” (where the airflow from the lungs is temporarily obstructed) with aspiration, rather than as true fricatives.
- There is general confusion about voiced fricatives – /Z/ (“plea**s**ure”), /dZ/ (“**j**oy”), /D/ (“brea**the**”), /z/ (“ro**se**”) etc., and they are frequently substituted for one another.
- The /N/ (“thi**ng**”) phoneme is often followed by an explicit /g/ (“**g**et”) phoneme in Indian English, and double letters in the orthographic transcription of a word often result in an extended duration of the corresponding phoneme. Neither of these phenomena are observed in British English.

2.2 - Prosody

As well as the acoustic realisation of the phonemes, different accents often exhibit different timing patterns, rhythm, intonation and stress. These features can be different at the phoneme, word, or sentence levels.

2.2.1 - Timing and Rhythm

The duration of Indian English phonemes tends to be more heavily modulated than British

English – some phonemes are significantly extended but others, often in bursts, and especially around consonant clusters, can be shortened.

It has been suggested that Indian languages are “syllable timed” (where syllables tend to have similar durations), rather than being “stress timed” like British English (where the time between consecutive stressed syllables tends to be equal, but the duration of the individual syllables varies). However this simplified interpretation of the differences between timings has proved difficult to substantiate with any objective evidence. Nonetheless there are clear differences in phoneme durations, and some of those differences do appear to be related to the syllable stress.

2.2.2 - Intonation and Stress

Similarly intonation in Indian English tends to be more varied than in British English, but also the relationship between pitch and stress can be quite different. In some cases it is even reversed – stressed syllables in Indian English can have a lower pitch than unstressed ones. This makes Indian English sound unnatural to British ears – as though all syllables were stressed. When combined with the different duration relationships mentioned in section 2.2.1, this can have a significant effect on the perceived naturalness and intelligibility of the speech.

2.3 - Perception

The perception of phonemes varies according to the linguistic environment and history of the listener. However, as already touched upon in Section 2.1.1, the perception of the *speaker* also affects their *pronunciation*. Depending on the extent and quality of any training they may have received (as well as their own aptitude), non-native speakers of English may be able to produce sounds which are more or less close to those of native English speakers, but in most cases, they cannot *perceive* the differences between these sounds and the nearest equivalent ones in their own native language. Such speakers produce these learned sounds by consciously modifying their articulation. In some ways this is akin to a form of a consciously learned “exemplar model“ - a theory of speech production which involves the memorisation of patterns of articulation at a larger scale than individual phonemes. Exemplar models are currently under investigation at the Phonetics Laboratory, University of Oxford.

Thus any mismatch between the linguistic environments and histories of the speaker and the listener may result in a perceived accent to the speech. As with production, the effect of perception can be related either to acoustic or temporal characteristics of the speech.

2.4 - Examples

As an example of some of the differences mentioned above, this section includes two example utterances of the phrase “Alternate (er) Energy Sources”. The first was taken from unscripted conversation of an Indian speaker, while the second was from a British English speaker who was instructed to say the same phrase, in a similar way to the Indian speaker (i.e. with similar intent). These are shown graphically in Figures 2 and 3 respectively.

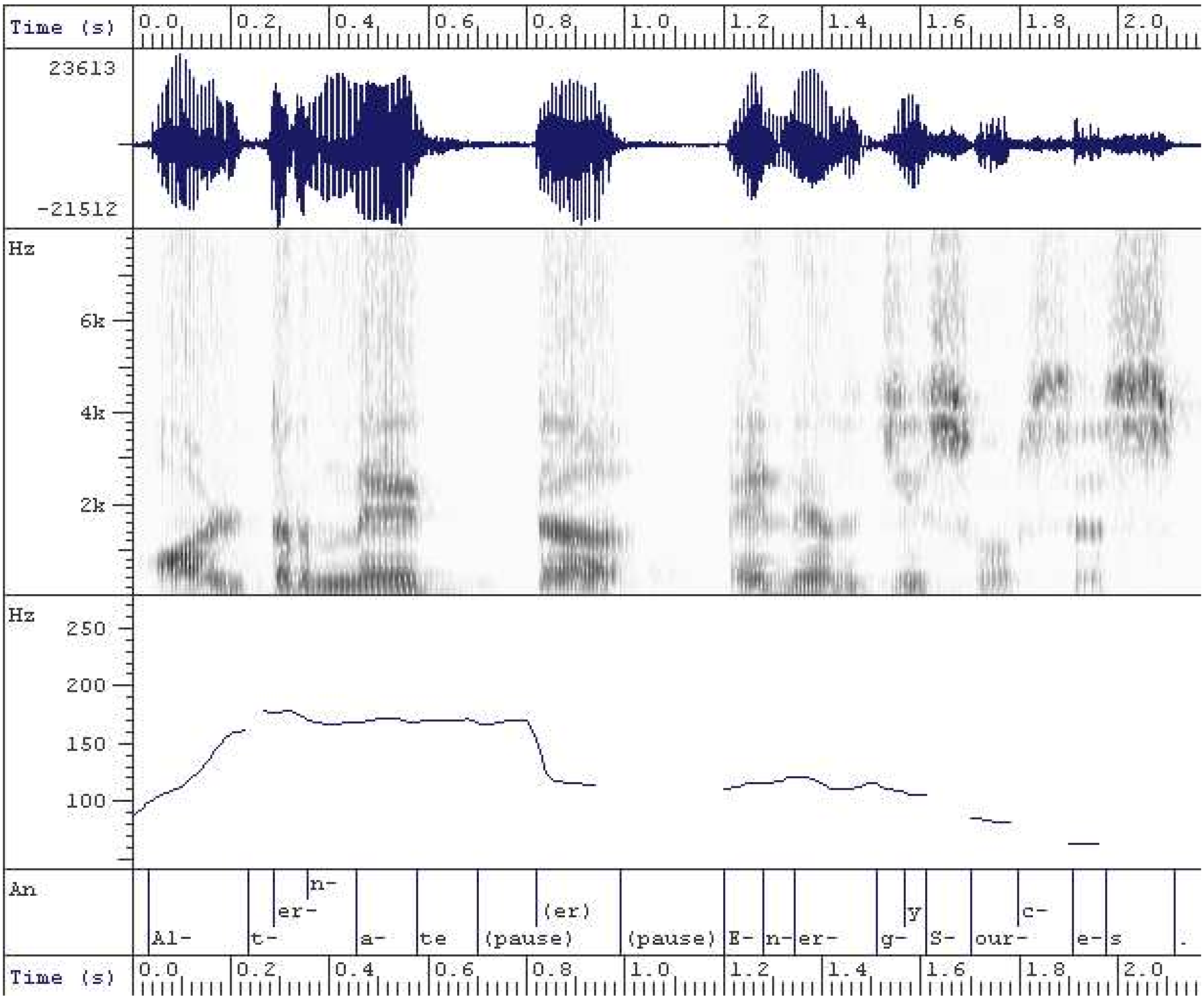

Figure 2: Unscripted Indian English utterance, "Alternate (er) Energy Sources"
Top row: recorded signal (waveform), Second row: broad-band spectrogram,
Third row: estimated pitch track, Bottom: word-fragment labels.

In these diagrams, the second rows are “broad-band spectrograms”. These show the temporal changes in speech energy at relevant frequencies. The vertical axis is the frequency of interest and the darkness of the image at each point reflects the power (intensity) of the signal at that point in time and at that frequency. They are termed “broad-band” because they are designed to have a fine time-resolution, and the main consequence of this is that they cannot simultaneously exhibit a high resolution along the frequency axis (i.e. they only quantify the energy content over a relatively broad band of frequencies).

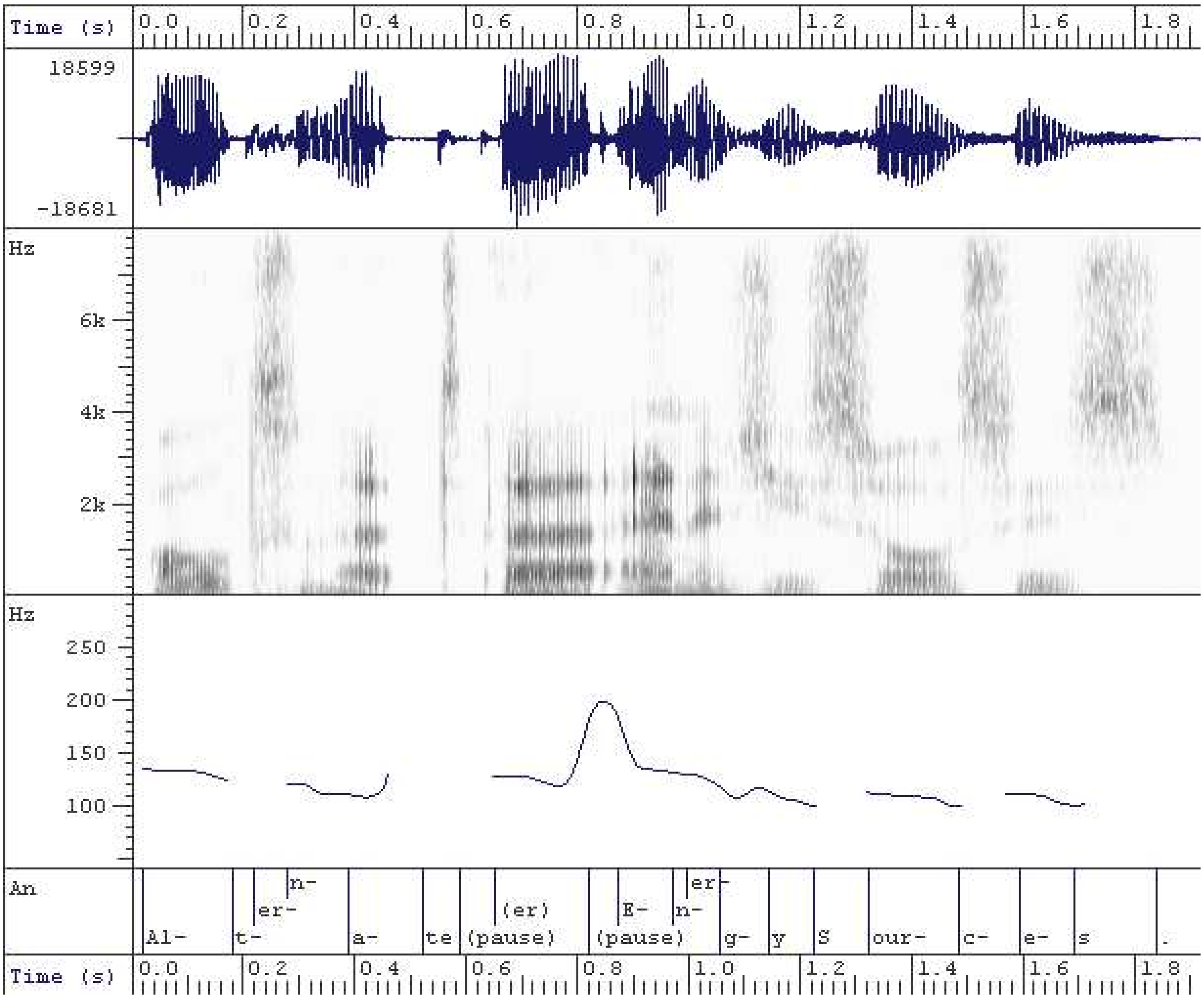

Figure 3: Native British English copy utterance, "Alternate (er) Energy Sources"
Top row: recorded signal (waveform), Second row: broad-band spectrogram,
Third row: estimated pitch track, Bottom: word-fragment labels.

It can be seen from these figures that, although there is a definite similarity between the two utterances, there are some aspects which are clearly different. In particular, the pitch track, as extracted via the SFS software from University College London [8], has a number of clear differences.

Note here that the “bump” around the middle of the pitch track in Figure 3 is an artefact of the algorithm used to extract the pitch – that region of the signal is essentially aperiodic, and so should really be represented as a break in the line.

However, comparing the rest of the two pitch tracks, it is clear that almost the entire British English pitch track is a gentle decline from approximately 135 Hz to 100 Hz throughout the whole phrase. Within each word (and even syllable), the pitch tends to fall slightly from its initial value, and so overall the pitch appears to fall at a steady rate throughout.

By contrast, during the initial syllable of the Indian English utterance, the pitch rises very rapidly from less than 80 Hz to over 160 Hz. It then stays around 170 Hz for the whole of the first word before dropping abruptly to 120 Hz. Following that the pitch drops noticeably with each new syllable (or run of voiced phonemes), but does *not* change significantly within them. Towards the end of the phrase, where the pitch is dropping more rapidly between syllables, the energy of the Indian English speech also decreases much more than in the British English.

In terms of duration, in this example, there is a clear difference, but only in the *word* which is unstressed (“sources” in this example). In the Indian English example, *all* the words show a variation in phoneme duration related to the stress of the respective phoneme within each word. The same is true for the British example in words which are stressed within the sentence, but for unstressed words the phonemes are of almost uniform duration.

Finally, the pronunciation of *some* phonemes is clearly different. In this example, the /l/ and /r/ phonemes are most obviously different. In the British English example, the /l/ of “Alternate” is hardly discernible at all – it affects the pronunciation of the /Q/ preceding it, but the vowel sound is almost completely steady. By contrast the Indian example is more fully articulated, with a steady change in the power spectrum throughout the /l/ phoneme.

3 - Systems

A number of components, which, taken together, could be used to produce an automatic accent converter, have been proposed over the years (see the references in [9]). Most (but not all) of these methods for performing such a conversion have concentrated on the acoustic realisation of the phonemes without allowing for the variations in timing and intonation. In particular, [9] itself did consider these issues too, and provided insights into the relative importance of the different “conversions” for the perceived degree of accent, speech quality, and speaker identity.

However, a typical system which involves phoneme-dependent transformations would require large vocabulary, continuous speech recognition (LV-CSR) optimised for the speaker’s accent, followed by synthesis with a synthesiser developed specifically for that of the listener.

This has limited their usefulness in “real-world” applications, where the constraints on the system in terms of processing delay (latency) make recognition inherently inaccurate. To achieve high phoneme or word recognition accuracy, it is generally necessary to analyse a complete utterance, which implies a delay of several seconds before the modified speech can be produced.

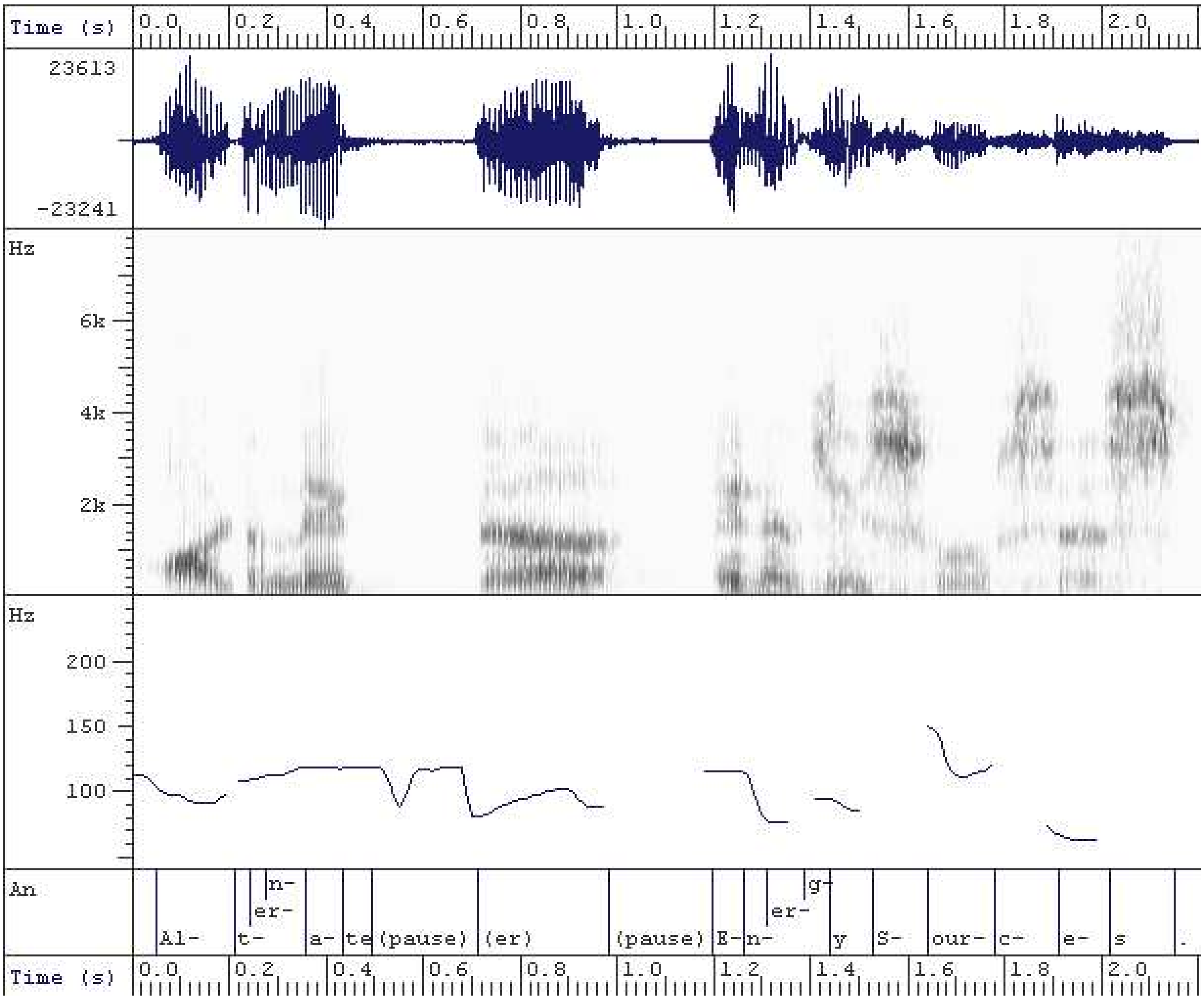

Figure 4: The Indian English utterance shown in Figure 2 manipulated to more closely resemble the British English utterance of Figure 3
Top row: recorded signal (waveform), Second row: broad-band spectrogram, Third row: estimated pitch track, Bottom: word-fragment labels.

By concentrating on sub-phonemic features – whether duration (tempo), intonation, and acoustics – speech spoken with one accent can be made to sound more natural to a listener acclimatised to another, without any need for low-latency high-accuracy (phoneme or word) recognition. As an example of this, the utterance of Figure 2 has been manipulated solely in terms of pitch, duration and effective vocal tract length. The results are shown in Figure 4 These parameters were controlled without reference to the identities of the words and phonemes, or the semantics of the phrase.

The perceived quality of this example of accent modification still leaves something to be desired due to various limitations of the software used to perform the modification. However, even though the processing was independent of the phonemes themselves, it does make the speech sound noticeably less “foreign”, while retaining a clear sense of speaker identity.

4 - Conclusions

Complete transformation of one accent to another, with low enough latency to facilitate natural communication, is not possible with current technology. Nonetheless, accent *moderation* can be achieved without explicit transformation of individual phonemes, and so is a more realistic aim.

By developing algorithms which operate in terms of short-term durations (tempo), intonation, and sub-phonemic acoustics, it appears to be possible to ease telephone communication between people with very different accents and hopefully reduce misunderstandings and apprehension in the participants. This approach has the additional advantage that characteristics of the original speaker are retained in the modified speech.

This work is, however, at a very early stage and it will be some time before the capabilities of such algorithms can be demonstrated conclusively.

References

[1] P. M. Schmid and G. H. Yeni-Komshian, “The Effects of Speaker Accent and Target Predictability on Perception of Mispronunciations”, *Journal of Speech, Language, and Hearing Research*, vol. 42, pp. 56-64, February 1999.

[2] R. K. Bansal, “The Intelligibility of Indian English” , CIEFL Monograph 4, Hyderabad, 1969.

[3] J. C. Wells, “Accents of English”, Cambridge University Press, 1982.

[4] O. Maxwell and J. Fletcher, “Acoustic and Durational Properties of Indian English Vowels”, World Englishes, vol. 28, iss. 1, pp. 52-69, February 2009.

[5] A. Sen, “Pronunciation Rules for Indian English Text-to-Speech System”, *Workshop on Spoken Language Processing*, Mumbai, January 2003.

[6] J. C. Wells, “SAMPA Computer Readable Phonetic Alphabet”, http://www.phon.ucl.ac.uk/home/sampa/, October 2005.

[7] “Census of India 2001”, http://www.censusindia.gov.in/, 2001.

[8] M. Huckvale, “Speech Filing System”, http://www.phon.ucl.ac.uk/resource/sfs/, February 2008.

[9] D. Felps et al., “Foreign accent conversion in computer assisted pronunciation training”, *Speech Communication*, vol. 51, pp. 920-932, 2009.