Translating sounds

Silvia Barreiro Bilbao
Universidad Nacional de Educación a Distancia
sbarreiro@flog.uned.es

1. Introduction

Automatic speech recognition is one of the three main areas that can be distinguished in the field of speech processing, together with coding and synthesis. The general purpose of speech recognition is "to single out the message" (Chollet 1994: 133). However, it can have a simpler design, that is, to identify a particular word. '*Word*' in the context of isolated word recognition means "a word or a short phrase that can be treated for recognition purposes as a single unit" (Wolf 1976: 173).

Nowadays, teaching English pronunciation can be complemented by the use of software packages in which speech recognition programs are included. However, there is no doubt that more information is needed concerning the advantages and disadvantages of their use in the process of learning and teaching L2 pronunciation (see Sustarsic 2001), especially when many teachers or tutors do not really know how a speech recogniser works (See Wolf 1976; Ladefoged 2001; Ainsworth 2005; Christensen *et al*. 2005; Torres 2006 or Benzeguiba *et al*. 2007 for an extensive explanation of how speech recognition programs work and how their performance can be improved by the use of phonetic knowledge).

The present paper aims to provide information on how sounds are recognised and *translated* into words by speech recognisers; In the first part we will explain briefly the architecture of a template-based recogniser and the problems usually encountered with this type of speech recogniser. In the second part we will offer a detailed description of the experimental sessions carried out to check the performance of an isolated-word template-based recogniser on a small vocabulary with the added problem of multiple speakers with different levels of proficiency in English. The analysis will also include comments on the patterns in the ordering of the scores. All this information can be relevant when integrating these systems into the process of teaching and learning L2 pronunciation.

2. Description of a template-based recognition system

The architecture of a template-based recognition system, as described by Wolf (1976), contains two phases.

The first phase, the *training phase*, starts with a speaker saying particular words which are converted into some sort of spectral representation, i.e., acoustic parameters are generated in accordance with those words. Finally, the resulting numbers are stored in a computer file which contains the acoustic parameters of the words and labels for their identification. Examples of acoustic patterns (called *templates*) for all the desired words to be recognised have to be stored in the machine.

The second phase, the *recognition phase*, starts with some incoming unknown words whose spectrum is calculated. Afterwards, the matching is carried out, that is, the unknown acoustic parameters of the incoming word are compared against all the words stored in the machine. Every time a comparison is made a number comes up which tells us how close these two sets of spectra match, and eventually one would be a closer match than the others.

This apparently simple process faces many problems associated to this type of speech recogniser (Chollet 1994; Holmes 2001), which can be summarised as follows:

1. *Segmentation*: there is a problem of how to isolate words from a string. It is very easy for the system to miss the beginning and/or end of words with a period of silence within them (stop gaps), or words which start and/or end with weak sounds like fricatives.

2. *Speaker dependency*: another important limitation is that the system is dependent on *one* speaker, the person who has recorded the templates of certain words. Consequently, if another speaker uses the speech recogniser the system matches a person’s voice against another person’s voice, creating normalisation problems. In other words, there will be mismatches between what the new speaker says and what is

stored in the machine. Those problems do not arise because the new user is saying a different word but because he/she may have a different accent or may use different voice quality, among other reasons.

3. *Discrimination*: the system is limited not only because all the desired words have to be recorded in advance, but also because it looks only for the *best match*. Therefore, the larger the entry vocabulary is, the greater the chance that the system chooses the wrong word. As a result, we end up with high great error rates.

This is almost inevitable with words which are phonetically similar, such as minima pairs, for example, '*pat*' and '*cat*'. The recognition system knows nothing about the phonetics of the word or about segmentation, and what it matches is the entire word rather than its segments. So, parts of the word that we know are irrelevant will contribute to the matching between the two words. As an example, if we say the word '*cat*' the system will match the '*c*' against all the initial consonants of the stored words, the '*a*' against all the '*a*'s and the '*t*' against all the final consonants, and comes out with a score. In the end, the recogniser chooses the overall pattern that is the best match. Therefore, the '*at*' part contributes to establish the difference, although we know it is not relevant in this case.

The limitations associated with minimal pairs can also be seen in those cases where the stored vocabulary is extended. If we have 100 words, for instance, the system has to do 100 matches, and if it is increased even more, say to 100,000 words, the system has to do 100,000 matches. Therefore, the chances of producing wrong matches will be greater as the average distances between the words are going to get smaller and smaller.

All these problems have to be kept in mind when carrying out experiment studies, as the one proposed in the following section.

3. Experimental study

The performance of an isolated-word template-based recogniser was checked on a small vocabulary in two experiments with different speakers with different levels of proficiency in English.

3.1. First experiment

This first experiment tested the system's performance by recording the number of correct and incorrect answers as well as by studying which templates were used by the recogniser.

3.1.1. Method

3.1.1.1. *Material*: 20 English words were chosen for this experiment, shown in Table 1 with their phonemic transcription between slashes.

Table 1. Vocabulary stored in the word recogniser

Face /feɪs/
Lace /leɪs/
Lea /liː/
Toe /təʊ/
Go /gəʊ/
Toes /təʊz/
Knee /niː/
Bee /biː/
Older /ˈəʊldə/
Shoulder /ˈʃəʊldə/
Shoulder-pad /ˈʃəʊldə pæd/
Finger /ˈfɪŋgə/
Linger /ˈlɪŋgə/
Fingerprint /ˈfɪŋgəprɪnt/
Halibut /ˈhælɪbət/
Hatchet /ˈhætʃɪt/
Eyebrow /ˈaɪbraʊ/
Rainbow /ˈreɪnbəʊ/
Cauliflower /ˈkɒli $_{o}$flaʊə/
Television /ˈtelɪ $_{o}$vɪʒən/

As can be seen in Table 1, most of the words contain *plosives and/or fricatives*, which, in principle, may be more difficult to detect by the recogniser. Also, there is *phonetic similarity* thanks to the use of minimal pairs and/or compounds of words already present in the study, although there is a set of words which are relatively less similar phonetically. Besides, the *number of syllables* per word is different.

3.1.1.2. *Recording and creating templates*: Two female speakers were chosen as participants in the experiment. The first one, called V.G., was bilingual in Greek and English and the second one, L.E. was a native speaker of English.

They produced each word in isolation in order to create the templates. All the resulting forty templates were stored in the same file, although the format of the word included the initials of the person who said it, so that we knew exactly which templates the system used to recognise incoming words.

Once the templates were created, the system performance was tested out by presenting unknown words to the recogniser produced by four female speakers: one Spanish speaker (called S.B.) with an advanced level of English and one Greek (K.G.) with an intermediate level, apart from the two speakers who created the templates. Thus, there were differences according to the level of proficiency in the L2.

The informants were asked to choose some words randomly from the vocabulary and produced them aloud to the recogniser through a digital microphone.

3.1.2. Results

The system responses showed in Table 2 indicates that, on the whole, the performance of the system was not bad as the mean number of correct responses across the four speakers was 79.2% (s.d. 26.2). Some of the wrong words recognised by the system were predictable as the incoming unknown word and the recognised word were phonetically similar. In fact, '*finger*' and '*linger*' are minimal pairs, differing only in the first phoneme. The same can be applied to '*toe*' and '*go*' and '*catch it*' and '*hatchet*'. As far as '*fingerprint*' and 'shoulder-pad' are concerned, the system delivered a similar word but with the omission of the unstressed final syllables. In general, vowels were better recognised than consonants (see error analysis in Sustarsic 2001).

Nevertheless, it is difficult to understand how the system worked when trying to recognise the words '*halibut*' and '*bee*'. In some occasions, the unknown word and the recognised words differed not only phonetically but also in number of syllables contained in the word.

Table 2. Results of testing the word recogniser

Unknown Word	**Speaker S.B.**	**Speaker L.E.**	**Speaker V.G.**	**Speaker K.G.**
Fingerprint	Fingerprint.LE	Finger.VG Finger.LE	Fingerprint.VG	Fingerprint.VG
Finger		Linger.LE		
Shoulder		Shoulder.LE	Shoulder.VG	Shoulder.VG
Shoulder-pad	Shoulder-pad.LE	Shoulder.LE	Shoulder-pad.VG	Shoulder.VG
Lace		Lace.LE	Lace.VG	
Face		Face.LE	Face.VG	
Toe			Toe.VG	Go.VG
Toes			Toes.VG	Toes.LE
Bee	Bee.LE	Bee.LE Fingerprint.VG		
Knee	Knee.LE			
Go			Go.VG	Go.LE
Cauliflower		Cauliflower.LE	Cauliflower.VG	Cauliflower.LE
Halibut		Halibut.LE Go.LE		
Hatchet	Hatchet.LE			
Catch it *	Hatchet.LE			

* This word was added while the recogniser was being tested out in order to see how the system dealt with a word which did not a stored template.

The analysis of the individual percentages (Table 3) showed that there was no a clear correlation between level of proficiency and of accuracy. First of all, two of the speakers shared 100% recognition: the bilingual person (V.G.) who had recorded the templates and the

Spanish speaker (S.B). Secondly, the English speaker (L.E.) who had also recorded the templates did not get 100 % word recognition. Many reasons can explain these surprising results which be stated in the conclusion section. Nevertheless, the actual cause could not be specified. Finally, the results obtained by the Greek informant (K.G.) were not unexpected as her English pronunciation showed a very strong foreign Greek accent.

Table 3. Percentage of correct responses obtained for each speaker

Speaker S.B.	Speaker L.E.	Speaker V.G.	Speaker K.G.
100%	45.5%	100%	71.4%

In relation to the word recogniser use of templates spoken by the same or different speaker, the system was generally consistent in the use of the templates spoken by the same speaker (Table 4). In fact, the Spanish speaker's unknown words were always matched with the English one's templates whereas the bilingual speaker's (V.G.) realisations were always matched with her own templates. In a similar way, the words uttered by the Greek speaker with an intermediate level of English (K.G.) were matched with the bilingual one's templates more than half the time. As an unexpected exception, the English speaker's unknown words were not always matched with her own templates.

Table 4. Percentage of responses obtained for each speaker

Speaker S.B.	Speaker L.E.	Speaker V.G.	Speaker K.G.
100% LE	81.8% LE	100% VG	57.1% VG
	18.2% VG		42.9% LE

3.1.3. Conclusions

This first experiment -in which we tested the system's performance by recording the number of correct and incorrect answers as well as by studying which templates were used by the recogniser- can be summarised as follows:

a. On the whole, the performance of the system could be considered good, as correct word recognition was above 75% across the four users regardless of their level of proficiency in the L2. It is worth pointing out that even when '*catch it*' was presented to the system, it was matched with a template which shared some phonetic similarity with it.

b. The recogniser tended to use the templates spoken by the same speaker. Furthermore, when two new users introduced unknown words, they were normally matched with one of the speakers who had created the templates in a systematic way.

c. When errors were analysed, it was clear that the system was not trying to segment the signal into phoneme-size units and then recognise them individually, but trying to recognise entire words spoken individually. What's more, it was not relying on the phonetics or the syllabic structure of the words, as humans do. The elements omitted were superior to the number of additions. Besides, similar error patterns were found across the speakers: the native speaker (L.E.) and the speaker with an intermediate level of proficiency in English (V.G.). Nevertheless, it is fair to say that errors were found in half the users.

d. There were some unexpected errors with one of the speakers who had previously created the templates. The actual cause of these errors could not be identified. In fact, they may have been caused by the combination of a few factors. Firstly, the templates could have not been stored properly; secondly, the speaker could have changed her speaking rate, or even voice quality, and thirdly the presence of a large compression could have caused matching problems, among other reasons.

3.2. Second experiment

The second experiment aimed at finding out if any patterns of scores emerged from the tables of best-matched distances to an input word as well as investigating the relative size of the smallest distances for the correct and the incorrect words.

3.2.1. Method

The procedure was the same as the one followed in the first experiment but this time four of the vocabulary words (*one* per speaker) were tested.

3.2.2. Results

3.2.2.1. *Native speaker of English (L.E.)*: As can be seen in Table 5 that shows the English speaker's results with the word '*cauliflower*', the first best match was the correct word produced by the same speaker who created the template, that is, by her. Besides, the template '*cauliflower*' produced by V.G. was the second best match which means that the system chose the right templates regardless of the speaker who created them. The difference between the correct and the first incorrect answer was only 0.94 dB[1], being the maximum distance between the first match and the last one 1.55 dB.

Table 5. Best-matched distances to '*cauliflower*' spoken by L.E.

Unknown Word	Speaker L.E.
Cauliflower	**1.89 cauliflower.LE**
	2.61 cauliflower.VG
	2.83 eyebrow.LE
	3.06 finger.LE
	3.07 rainbow.LE
	3.08 older.VG
	3.10 linger.LE
	3.11 eyebrow.VG
	3.15 go.LE
	3.17 toe.VG
	3.27 finger.VG
	3.28 shoulder.LE
	3.30 go.VG
	3.33 older.LE
	3.44 halibut.LE

[1] It represents the average over the 19 channels covering the spectrum up to 5kHz, and over all the frames.

On the whole, the system chose the right speaker in 60% of the total number of words.

We could not see any other pattern of ordering in the scores apart from the fact that the first five words, given as the best matches, had more than one syllable, although they did not share phonetic similarity.

3.2.2.2. *Bilingual speaker (V.G.)*: Her results with the word '*lace*' (Table 6) shows that the first best match was the correct word produced by the speaker who created the template. In this occasion, the template '*lace*' produced by the other speaker (L.E.) was the fourth best match after two incorrect words, meaning that the distance between that template and the first one was greater than those of the two words in the middle. The difference between the correct and the first incorrect answer was only 0.76 dB. On the whole, the maximum distance between the fist match and the fifteenth was 1.6 dB.

Table 6. Best-matched distances to '*lace*' spoken by V.G.

Unknown Word	Speaker V.G.
Lace	**1.72 lace.VG**
	2.48 face.VG
	2.51 lea.VG
	2.67 lace.LE
	2.76 television.VG
	2.84 lea.LE
	2.94 knee.VG
	3.04 bee.LE
	3.09 face.LE
	3.10 toes.LE
	3.13 bee.VG
	3.13 halibut.VG
	3.27 knee.LE
	3.28 older.LE
	3.32 fingerprint.VG

The system chose the right speaker in 53.3% of the 15 words, being very noticeable in the first four words.

Analysing the words given as the best five matches in order to see any pattern of ordering, it is relevant the fact that the first four were similar phonetically and in the number of syllables. Vowels and consonants had an accuracy rate of 75% identification. No pattern was found from the fifth word onwards.

3.2.2.3. *Non-native speaker but with advanced level of English (S.B.)*: The Spanish speaker's results with the word '*finger*' (Table 7) reveals that the first best match was the correct word produced by the speaker whom she used to be associated with, i.e., the English speaker (L.E.), being the template '*finger*' produced by the other speaker (V.G.) was the second best match, meaning that the recogniser once more chose the right templates regardless of the speaker who created them. The difference between the correct and the first incorrect answer was only 0.22 dB. The maximum distance between the fist match and the last one was 0.97 dB.

Table 7. Best-matched distances to '*finger*' spoken by S.B.

Unknown Word	**Speaker S.B.**
Finger	**2.10 finger.LE**
	2.12 finger.VG
	2.32 linger.LE
	2.61 shoulder.LE
	2.72 shoulder-pad.VG
	2.77 fingerprint.VG
	2.80 halibut.VG
	2.81 shoulder.VG
	2.81 older.LE
	2.83 cauliflower.LE
	2.85 older.VG
	2.88 linger.VG
	2.89 shoulder-pad.LE
	3.07 toes.LE
	3.07 toe.VG

The system chose the English speaker L.E. in 46.7% of the total number of words, and the remaining 53.3%, the bilingual speaker

V.G. This almost random choice was expected because the Spanish speaker did not create any template.

Looking at the five first words given as the best matches, they all shared the characteristic that they had more than one syllable, even more, the first four words ended up with the same sound /ə/. Besides, the third best match formed a minimal pair with the unknown word as they only differed in the first phoneme. After the sixth word we could not find any pattern of ordering of the scores.

3.2.2.4. *Non-native speaker with an intermediate level of English (K.G.)*: Table 8 shows the Greek informant's results with the word '*older*'.

As seen in the Table 8 the first best match was the correct word produced by the speaker whom she was normally associated with, i.e., the bilingual speaker (V.G.). As happened with the previous speaker analysed, the second best match was the template '*older*' produced by the other speaker (L.E.). The difference between the correct and the first incorrect answer was only 0.36 dB, being the maximum distance between the fist match and the last one was 0.89 dB.

Table 8. Best-matched distances to '*older*' spoken by K.G.

Unknown Word	Speaker K.G.
Older	**2.18 older.VG**
	2.34 older.LE
	2.54 cauliflower.LE
	2.67 finger.VG
	2.74 go.LE
	2.76 shoulder.LE
	2.79 eyebrow.LE
	2.80 linger.LE
	2.90 finger.LE
	2.92 shoulder.VG
	2.99 eyebrow.VG
	3.04 halibut.VG
	3.06 linger.VG
	3.06 shoulder-pad.VG
	3.07 rainbow.LE

The system chose the English speaker L.E. in 53.3% of the total number of words, and remaining 46.7% of the time, it chose the bilingual speaker. Once again this almost random choice was expected because this speaker with an intermediate level of English (K.G.) had not created any template.

The analysis of patterns of ordering of scores by looking at the five first words given as the best matches showed that most words shared the characteristic of having more than one syllable and ending in schwa. Also, 99% of the 15 words had more than one syllable.

3.2.3. Conclusions

The summary of this second experiment -in which we investigated the relative size of the smallest distances for the correct, and the incorrect words, as well as the pattern of ordering of the scores- can be expressed in the following terms:

a. The best match given by the recogniser was always the correct one, and the second best match was correct in three out of the four unknown words tested as well.

b. When the subjects who created the templates presented unknown words to the recogniser, the system used their templates as the first best match. As far as the other two speakers were concerned, the choice in the first best match was normally the expected one, in accordance with the previous testing, i.e., the Spanish speaker's words were matched with those uttered by the native whereas the Greek speaker's words, with those expressed by the bilingual person. However, when analysing the speaker chosen in the fifteen words in the tables, it was clear that that the percentage of using either the templates created by the English or those created by the bilingual person was around 50%, for *all* the four speakers, regardless of who created the templates. Therefore, the system did not show a total dependence on the speaker.

c. The analysis of the first five best matches given by the recogniser revealed that the tendency was to choose words which either shared some phonetic feature or had the same number of syllables. Besides, the phonetic similarity was based on the vowels rather than on the consonants.

d. The smallest differences between the correct and the first incorrect answer varied between 0.22 dB and 0.94 dB. This number was really small, and it did not increase that much over the 15 best matches given by the system every time it was tested. The largest difference between the first word and the fifteenth one was 1.6 dB. It is known that the average distance between words gets smaller as the vocabulary size increases, so that if the vocabulary were 100, the distances between words would be far fewer. Consequently, the error rate given by the recogniser would have increased considerably. Let's remember here that these numbers do not really represent our intuitions of distance; it would be interesting to compare machine distance with perceptual distance.

4. Final comments

This paper aims to provide more information of how an isolated-word, template-based recogniser works on a small vocabulary and with multiple speakers with different levels of proficiency in English. Its performance was tested out together with the description of the problems associated to this kind of speech recognisers.

It is fair to say, however, that the successful performance of this recogniser was limited, as it was based on a very small set of isolated words. Consequently, more testing is required and in a more systematic way in order to let us record more precise and reliable conclusions concerning speech recognisers. Nevertheless, error proved the prediction concerning the performance of the system in relation to minimal pairs and, in a lesser degree, to speaker-dependency. Furthermore, as all the findings were obtained on a very small number of samples we could not conclusively identify clear error patterns across the speakers.

In order to apply automatic speech recognition programs successfully to teaching and learning a foreign language, they have deal with many issues, such as, continuous speech, large vocabulary or speaker-independence, among others. Regarding this matter, significant improvements have been made in current sophisticated statistical recognisers.

In the context of teaching a language, what we have to bear in mind is that speech recognisers may never reach 100% correct recognition, which makes one suggest that our conception of the programs' usefulness should be changed. These programs should not be used only to make our students repeat and talk to the machine to achieve 100 per cent accuracy, as they can be easily disappointed if they not reach that level or they might feel penalised by their mistakes. Let's remember at this point that the program had problems with a native speaker who had created the templates; therefore, problems would be greater with non-native students. On the contrary, we should practise positive criticism of our students' mistakes and analyse the error patterns in relation to the language differences. For instance, a lack of aspiration of an English /p/ in certain contexts would be recognised a [b], etc. This method would encourage our students not only to become aware of their own pronunciation but also to improve it.

What is clear is that these programs and speech recognisers should be a complementary tool and never the substitute of a 'real' tutor or speaker. The process of speech communication requires perception (from the listener) and production (from the speaker). The productive part performed by the speaker gives significant acoustic and visual cues to the listener, and that would not be overlooked when technology is applied to languages, mainly in the process of learning or teaching a L2 pronunciation.

References

Ainsworth, W.A. 2005. "Can phonetic knowledge be used to improve the performance of speech recognisers and synthesisers?". In *The Integration of Phonetic Knowledge in Speech Technology*, W.J. Barry and W.A. Van Dommelen (eds.). Dordrecht: Springer [*Text, Speech and Language Technology* 25]. 13-20.

Benzeguiba, M., De Mori, R., Deroo, O, Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. and Wellekens, C. 2007. "Automatic speech recognition and speech variability: A review". *Speech Communication* 49. 10-11: 763-786.

Chollet, G. 1994. "Automatic Speech and Speaker Recognition: Overview, Current Issues and Perspectives". In *Fundamentals of Speech Synthesis and Speech Recognition*, E. Keller (ed). Chichester: John Wiley & Sons. 129-147.

Christensen, H., Lindgren, B. and Andersen, O. 2005. "Introducing phonetically motivated, heterogeneous information into automatic speech recognition". In *The Integration of Phonetic Knowledge in Speech Technology*, W.J. Barry and W.A. Van Dommelen (eds.). Dordrecht: Springer [*Text, Speech and Language Technology* 25]. 67-86.

Holmes, J.N. and Holmes, W. 2001. *Speech Synthesis and Recognition*. 2nd edition. London: Taylor & Francis.

Ladefoged, P. 2001. *Vowels and Consonants: An Introduction to the Sounds of Language*. Oxford: Blackwell.

Sustarsic, R. 2001. "Using a speech recognition program in teaching English pronunciation". *Proceedings of the Phonetics Teaching and Learning Conference*: 47-50.

Torres, M.I. 2006. "El reconocimiento del habla". In *Los sistemas de diálogo*, J. Llisterri and M.J. Machuca (eds.). Bellaterra-Soria: Universitat Autònoma de Barcelona, Servei de Publicacions-Fundación Duques de Soria [Manuals de la Universitat Autònoma de Barcelona, *Lingüística* 45]. 81-98.

Wolf, J.J. 1976. "Speech Recognition and Understanding". In *Digital Pattern Recognition*, K.S. Fu (ed). Berlin: Springer-Verlag. 167-203.