The Use of Australian-English Vowel Formant Data Sets in Forensic Speaker Identification.

PANZE WORKSHOP, SST 04.

Tony Alderman
Phonetics Lab (Arts) & Australian National Dictionary Centre
Australian National University,
alderman@webone.com.au

Abstract

This paper demonstrates how vowel formant data sets of Australian English, from Bernard (1970, 1989) & Cox (1999) can be used to model the reference distribution in Bayesian Forensic Speaker Identification. In this paper I show how young male Australian same-speaker pairs can be discriminated from different-speaker pairs on the basis of their vowel formants in sub-sets of their /tense monophthongs/ (Alderman, 2004), using two different reference distributions – Bernard's 1960's data set of male speakers of Australian English, and Cox's 1990's data set of male speakers of General Australian English. The newer data is shown to yield better discrimination rates, although for most parameters the two data sets yield similar results.

1. Introduction

This paper demonstrates how vowel formant data sets of Australian English, from Bernard (1970, 1989) & Cox (1999) can be used to model the reference distribution in Bayesian Forensic Speaker Identification. In the real world, Forensic Speaker Identification (FSI) typically involves the comparison of one or more samples of an unknown voice with one or more samples of a known voice. The data for comparison may consist of recordings of telephone calls, surveillance videos, records police interviews and the like. The court wants to determine whether the two samples have come from the same person or not, and thus be able either to identify the suspect as the offender or exonerate them; the forensic expert is usually asked to make some kind of statement regarding the similarities/differences between the two voices.

It is now widely accepted that the proper way to evaluate the evidence in many areas of forensic identification, like FSI, is from a Bayesian perspective (Robertson & Vignaux, 1995; Aitken 1995; Rose, 2002; González-Rodríguez, Ortega-García & Sánchez-Bote 2002). In FSI this typically involves comparing suspect and offender speech samples against a reference (also called background) distribution of the relevant population to determine both their similarity and typicality (Rose, 2002). The aim is to estimate a likelihood ratio (LR) which quantifies the strength of the evidence in favour of same-speaker provenance.

The Likelihood Ratio is a way of quantifying the strength of evidence supporting one of two competing hypotheses. This is not the same as providing a statement of the probability of the hypothesis given the evidence; this requires access to prior odds, which are not generally available to the analyst. Further, it is not the place of the analyst to provide such a statement in a legal context, as this is the proper domain of the judiciary (Robertson and Vignaux 1995).

The LR shows the ratio of the probability of evidence assuming one hypothesis, divided by the probability of evidence assuming the competing hypothesis. This is shown at (1), where *p* represents probability, H a hypothesis (usually of identity), H_A a hypothesis in competition with H (some statement of non-identity), and E represents the evidence under consideration.

$$LR = \frac{p(E|H)}{p(E|H_A)} \qquad (1)$$

In FSI, H is typically the prosecution hypothesis that the samples have been produced by the same speaker, and H_A the defence hypothesis that the samples were produced by different speakers. A LR greater than 1 shows a higher relative probability of the evidence given the prosecution hypothesis, and a value less than one the opposite. The magnitude of the distance of the LR from unity indicates how strongly the evidence supports one of the two hypotheses.

In order to accurately assess the similarity of the two speech samples, as well as evaluate the strength of evidence supporting the different-speaker hypothesis, a

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after abstract only review

sample which accurately represents the population must be used, in order to evaluate just how typical of the different speakers in question the samples being analysed are. Two samples may show remarkable similarity to each other, but this in itself is not sufficient to support the hypothesis that they were produced by the same speaker. It must also be evaluated just how typical this feature is in the relevant population. If it is very common to have such a feature in the population, the similarities between the samples become less supportive of the same-speaker hypothesis.

The reference distribution is therefore an essential part of the evaluation of FSI evidence, and the identification of appropriate data for use as a reference population is therefore essential to the application of FSI in Australia. In the late 1960's John Bernard collected and analysed the speech of 170 male speakers of Australian English (AE) (Bernard, 1970, 1989). The first 3 formants of their vowels were measured. The speakers were also categorised according to one of the three accent types proposed by Mitchell and Delbridge (Mitchell and Delbridge 1965). As the different-speaker hypothesis can take a number of forms (different speaker, different broad speaker, etc.), this means that the data set is potentially useful for a number of different variations of the different-speaker hypothesis based on accent descriptions.

A potential problem with the use of the Bernard data for FSI in Australia, however, is that it is over thirty years old, and it is accepted that language changes over time. While Bernard has been shown to be useful as a reference population in FSI – same-speaker pairs can be discriminated from different-speaker pairs with it - (Alderman 2004; Kirkland 2003; Rose 2003), an important question to answer is whether a newer data set, reflecting contemporary vowel targets in AE, would be more suitable as a reference distribution for the actual practice of FSI in Australia. Cox's (1999) data, while only sampling speakers classified as speaking General AE, holds some promise in this regard. It is also structured for comparison with the Bernard data, meaning that it is also comparable in structure to the experimental data collected in Alderman (2004).

The aim of the experiment in this paper is to show how young male Australian same-speaker pairs can be discriminated from different-speaker pairs on the basis of their vowel formants in sub-sets of their /tense monophthongs/ (Alderman, 2004), using two different reference distributions – Bernard's and Cox's. This is done using target values for the first three formants of the /tense monophthongs/ of AE as recorded by Bernard. The relative performance of the data sets used for the reference distribution is examined with relation to magnitude of LRs, and also the LR as a discriminant function.

It is shown that both data sets perform well as a reference distribution, but that there are some differences in the performance of discrimination in some parameters which results in the Cox data performing slightly better than the Bernard data.

2. Methods of Modelling the Reference Distribution

2.1. The Normal Approach

There is more than one method for calculation of the LR using continuous acoustic parameters such as formants. LRs can be estimated using both analytical and empirical approaches – this experiment uses an analytical approach using derived LR formulas. As noted in Rose, Lucy, & Osanai (2004), empirical approaches are often adopted in automatic Forensic Speaker Recognition.

Within an analytical framework to the estimation of LRs in FSI, a number of formulae are available. All have strengths and limitations, and their usefulness varies depending on the structure of the data available for use as a background sample. The first of these assumes normality in the variables' distributions, and makes use of the mean and standard deviation of the sample. There are a number of variants of this formula, but Lindley (1977) derives one such version, which also – rather unrealistically for speech - assumes equal variance for both samples. This formula is shown at (2).

$$\mathrm{LR} \approx \frac{\tau}{\mathrm{a}\sigma} \;\mathrm{x}\; \underbrace{e^{\left\{-\frac{\left(\bar{x}-\bar{y}\right)^2}{2\mathrm{a}^2\sigma^2}\right\}}}_{\text{similarity term}} \;\mathrm{x}\; \underbrace{e^{\left\{-\frac{\left(w-\mu\right)^2}{2\tau^2}+\frac{\left(z-\mu\right)^2}{\tau^2}\right\}}}_{\text{typicality term}}$$

$\bar{x}$ = mean of questioned sample; $\bar{y}$ = mean of suspect sample

μ = mean of reference sample

σ = standard deviation of questioned and suspect samples

τ = standard deviation of reference sample

$z=\left(\bar{x}+\bar{y}\right)/2$

$w=\left(m\bar{x}+n\bar{y}\right)/\left(m+n\right)$

m = number in questioned sample

n = number in suspect sample

$\mathrm{a}=\sqrt{1/m+1/n}$

(2)

Lindley, in his glass fragment paper, notes that the formula used assumes a normal background distribution, and that the samples being compared share the same variance. This presented problems regarding the use of

LR calculations on data sets that do not exhibit a normal distribution. Rose (2002: 321) argues that a corpus of data representing speech is one such example of a distribution which may display a deviation from normality (Rose 2002: 321). Figure 1 shows a histogram distribution for F2 of /o:/ from Bernard's data set), with a normal curve superimposed. As can be seen, it appears bimodal.

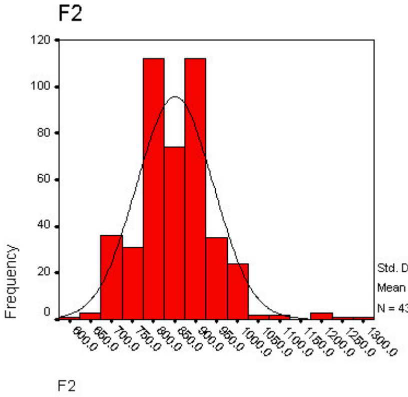

Figure 1: Histogram of F2 of /o:/ from Bernard's data (all accents combined)

This non-normality of the reference distribution has consequences for the level of accuracy of the LRs, as a non-normal distribution may have common values at different distances from the mean, meaning that the likelihood ratio calculated on the basis of a normal distribution may be too high or low. A benefit of the normality model, however, is that only a mean and standard deviation are required to model the distribution, and secondly, that it has been shown to work (Alderman 2004; Rose 2003; Kirkland 2003).

2.2. Kernel Density

Kernel Density Estimation is capable of modelling non-normal background distributions, but requires accurate estimation of within-speaker variance for a reliable LR calculation, as well as access to all of the tokens comprising the data set, not just the mean and standard deviation. (Aitken 1995: 184). There are, however, difficulties associated with the use of kernel density estimation to model the Bernard data.

The Bernard data comprises a maximum of three tokens for any given vowel for any given speaker. This results in poor estimation of the within-speaker variance. Although Alderman (2004) found kernel density estimation promising, the difficulty in accurate estimation of the within-speaker variance remains a problem.

Additionally, the formula given at (2) was found in Alderman (2004) to provide better discrimination for some combinations of parameters. This formula was also used in the calculation of the LRs for this experiment. It is hoped that the estimation of within-speaker variance will be overcome, to allow the proper use of kernel density estimation within FSI in Australia, as it clearly provides a more accurate representation of the reference distribution for use in the calculation of the LRs.

Both the normality and kernel density approaches assume independence of the parameters, which has not been established for this data. This is because, whilst theoretically this may bias the generated LR, this is an empirical test to see which formants, and combinations of formants, provide the best performance for the practice of FSI in Australia. Results showing correct discrimination of speakers suggest that while some dependence may exist between the parameters, this does not result in incorrect results, and thus the issue becomes less salient to the experiment at hand. Further, this "Idiot's Bayes" approach, assuming independence of parameters under analysis, has been shown to perform better than methods incorporating dependence of parameters (Rose et. al 2004).

2.3. The Experiment

The experiment uses data collected for the experiments in Alderman (2004). This data consists of formant centre frequencies (F1, F2, & F3) extracted from recordings of eleven male speakers of AE, including a pair of identical twins, aged between 18 and 26 years of age. Tokens of the tense monophthongs were collected in a controlled environment. Vowel tokens were elicited in a /h_d/ context, in a stressed sentence-final position (e.g. "that wasn't very hard", for /a/). This structure was intended to provide a comparable structure to the Bernard data, in controlling for the effect on F-pattern of the perivocalic segments, although it has some shortcomings in conforming to the criterion for optimal speech samples used in FSI analysis (Rose, 2002).

Each speaker was recorded on two occasions, separated by at least two weeks. This introduces within-speaker variation, a feature of naturally occurring speech which is a crucial desideratum for FSI experiments of this kind (Rose, 2002). Twelve tokens of each vowel were elicited in each recording session.

The two non-contemporaneous recordings of the eleven speakers provide a total of 231 comparisons (eleven same-speaker comparisons and 220 different-speaker comparisons). LRs were calculated for F1, F2, and F3 of each of the tense monophthongal phonemes /i/, /a/, /o/, /ʉ/, and /ɜ/, giving at total of 15 individual LRs for each speaker pairing for analysis. Combinations of LRs were calculated to find the best performance in terms of discriminating between same-speaker and different-speaker pairs. The combinations of parameters found to perform speaker discrimination

Accepted after abstract only review

most successfully were then examined in terms of the magnitude of LRs generated.

The Cox data has only speakers of General AE, and for more systematic comparison with the Bernard data, only the speakers classified by Bernard as speakers of General AE are used in the experiment.

3. Results

Table 1 presents LR results for all of the individual parameters used. This constitutes 15 individual parameters in all (F1, F2, and F3 of 5 vowels). The columns marked ‘B’ are the results using the Bernard data for the reference distribution; the columns marked ‘C’ are the equivalent results using Cox’s data as the reference distribution. The Cox data is also presented in italic font. The SS row lists the percentage of same-speaker pairs correctly discriminated (out of eleven comparisons), and the DS row shows the same information for the 220 different-speaker pairings. Please note that Table 1 spans across to the next column.

The LR$_{test}$ is an overall measurement of the success of the tests carried out. It is a ratio of the number of correctly discriminated same-speaker pairings to the number of incorrectly discriminated different-speaker pairings, and is an overall measure of the strength of evidence (the results) supporting the hypothesis that a same-speaker pairing will be resolved with a LR greater than 1, compared to the hypothesis that a different-speaker pairing will be resolved with LR>1. The LR$_{test}$ thus takes the form of a ratio of the conditional probabilities p(LR>1 | SS) / p(LR>1 | DS). This works the same as other LRs, in that values greater than 1 suggest greater relative support for same-speaker discrimination with LR>1, than different-speaker discrimination with LR>1. Values have been rounded to 1 decimal place.

Table 1: LR results for individual parameters

F1	/i/		/a/		/o/		/ʉ/		/ɜ/	
Set	**B**	***C***	**B**	***C***	**B**	***C***	**B**	***C***	**B**	***C***
SS	45.45	*45.45*	72.73	*72.73*	54.55	*54.55*	36.36	*36.36*	63.64	*63.64*
DS	65.00	*69.09*	74.55	*74.55*	70.45	*69.55*	65.91	*66.82*	76.82	*76.36*
LR$_{test}$	1.3	*1.5*	2.9	*2.9*	1.8	*1.8*	1.1	*1.1*	2.7	*2.7*
F2	/i/		/a/		/o/		/ʉ/		/ɜ/	
Set	**B**	***C***	**B**	***C***	**B**	***C***	**B**	***C***	**B**	***C***
SS	45.45	*54.55*	81.82	*81.82*	54.55	*54.55*	63.64	*72.73*	81.82	*81.82*
DS	82.27	*80.91*	85	*84.55*	68.64	*68.64*	77.27	*75.91*	82.27	*80.91*
LR$_{test}$	2.6	*2.9*	5.5	*5.3*	1.7	*1.7*	2.8	*3.0*	4.6	*4.3*
F3	/i/		/a/		/o/		/ʉ/		/ɜ/	
Set	**B**	***C***	**B**	***C***	**B**	***C***	**B**	***C***	**B**	***C***
SS	36.36	*36.36*	36.36	*36.36*	54.55	*54.55*	63.64	*72.73*	81.82	*81.82*
DS	79.55	*77.27*	66.36	*65.91*	80	*82.27*	61.36	*56.36*	67.73	*66.36*
LR$_{test}$	1.8	*1.6*	1.1	*1.1*	2.7	*3.1*	1.6	*1.7*	2.5	*2.4*

The results show that there is not much difference in performance between the two data sets as reference distributions for FSI. For the majority of the parameters the percentage of successfully discriminated same-speaker pairs remains constant between reference samples. F2 of /i/ did show better same-speaker discrimination using the Cox data, as did F2 and F3 of /ʉ/. F2 of /ʉ/ has one of the largest mean differences between the data sets; for the tense monophthongs only F3 of /o/ exhibited a larger difference between the data sets (Cox 1999). In all of three instances the use of the Cox data as the reference distribution yielded 1 more correctly discriminated same-speaker pair than when using the Bernard data.

However, for the individual tests, the Cox data does not generally discriminate different-speaker pairings as well as the Bernard data. Out of the 15 individual parameters, only F1 of /i/, F3 of /o/, and F1 of /ʉ/ have better different-speaker discrimination rates when using the Cox data. The reasons for this are not apparent, but may have to do with sample size or other factors. This is important, especially in forensic contexts, where the future freedom of a suspect in part depends on the results of such an analysis.

Individual LRs can be combined by simply multiplying them together. The implicitly “Idiot’s Bayes” approach used in this experiment does not need to account for any correlations between parameters. Table 2 presents the discrimination rates for optimum combinations of parameters, when using the two reference distributions. Optimum combinations are those with the highest LR$_{test}$ values while maintaining high levels of different-speaker discrimination, which as stated is an important consideration in forensic contexts. Combining more parameters yields theoretically larger LRs (through multiplication of individual LRs), so

Accepted after abstract only review

number of parameters is also salient in the selection of optimum combinations.

Table 2: LRs for Optimum Combinations of Parameters (n.d. = not defined)

	All F2, /ɜ/&/a/ F1		All F2, /ɜ/&/a/ F1, /ʉ/&/ɜ/ F3		All parameters combined	
Set	**B**	**C**	**B**	**C**	**B**	**C**
SS	7	7	7	8	4	5
DS	219	219	219	219	220	220
LR$_{test}$	140	140	140	160	n.d.	n.d.

The performance of the two data sets as reference distributions remains very similar for combinations of the individual LRs. It can be noted again that the level of same-speaker discrimination is the same or better when using the Cox data as the reference distribution, and that, when combining parameters, the level of different-speaker pair discrimination equals that of the results when using the Bernard data as the reference distribution.

The LR$_{test}$ for the experiment varies by combination of parameters, but again both data sets can be seen to perform well as a reference distribution for the discrimination of pairs of recordings from male speakers of General AE. For the Cox data, the two highest LR$_{test}$s were generated by combining the LRs yielded by F2 for all vowels, F1 of /a/ and /ɜ/, and F3 of /ʉ/ and /ɜ/, for a total of 9 parameters combined. This combination resulted in an LR$_{test}$ of 160, with 72.7% (8/11) same-speaker pairs correctly discriminated, and 99.55% (219/220) different-speaker pairs correctly discriminated.

This can be compared to a maximum LR$_{test}$ of 140 for the Bernard data when using only the data classed as from speakers of General AE. In terms of verbal equivalents of LRs used by the British Forensic Science Service, both of these values would be expressed as providing moderate support for the hypothesis that same-speaker pairs will be resolved with LR values greater than 1 (González-Rodríguez et. al 2002).

A method of graphically expressing both the magnitude of LRs calculated, and the relative performance of the SS and DS pairings is to use cumulative distribution functions (Drygajlo, Meuwly, and Alexander, 2003). Figure 1 presents the cumulative distribution functions of the Log$_{10}$(LR) values for the 9 parameter combination yielding the highest LR$_{test}$ scores. The Log$_{10}$ transformation of the LRs changes the threshold of discrimination to 0 (Log$_{10}$(1)). The analysis using the Bernard data as the reference distribution is shown in the left panel, with the results using the Cox data shown on the right. Each panel also shows the equal error rate for the test (labelled EER). Log$_{10}$(LR)s are shown along the x axis, with percentage of sample shown on the y axis.

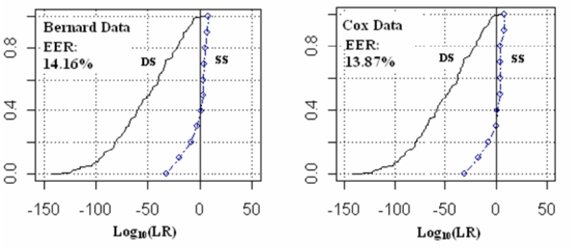

Figure 1: Cumulative Distribution Function of Log$_{10}$(LR) values for combination of all F2, F1 of /a/ and /ɜ/*, and F3 of* /ʉ/ and /ɜ/*, using Bernard data (left) and Cox data (right).*

In both panels the DS curve shows almost 100% of cases correctly discriminated. For both data sets there is a much larger magnitude of DS LR values than SS for both sets of LRs shown in Figure 1. The LR values for the same-speaker pairings are plotted as dots along the curve. The extra correctly discriminated same-speaker pairing using the Cox data can be seen close to the vertical line marking the discrimination threshold. The similarity in the curves between the two panels highlights the similarity in the results using the different reference distributions. For both sets of LR calculations the different-speaker pairings are resolved with Log$_{10}$(LR)s of nearly -150, while same-speaker pairings are resolved with values quite near to threshold (a maximum of 8 for the Cox data, and 7 for the Bernard data).

4. Conclusion

The two data sets separated by 30 years perform similarly in terms of successful discrimination of same- and different-speaker pairs of male speakers of AE. A number of observations can be made here.

The Bernard data set of over thirty years of age can be seen to be still useful as a reference distribution for FSI in Australia.

The Cox data shows that in some parameters undergoing change in target (such as /ʉ/), a more recent reference distribution can result in improved discrimination of speaker pairs.

Differences in performance of parameters can be exploited to maximise the accuracy of the result in a real forensic situation. For example, if the data to be analysed contained frequent tokens of /ʉ/, then the results obtained here suggest that the Cox data would be a better reference distribution. This is also of course reliant on the classification of the speech samples to be compared as General AE (as this is the only accent category collected in Cox 1999).

The similarity in result across the majority of parameters suggests that both data sets could be used in some contexts, as a way of verifying and strengthening results of analyses.

Overall, the Cox data can be seen to perform better than the Bernard data in terms of LR_{test} values, although the difference between 140 and 160 in the LR_{test} is not particularly large. While not providing conclusive evidence, this suggests that the newer data set outperforms the Bernard data as a reference distribution for Bayesian FSI on male speakers of General AE. This supports the notion that an accurate LR requires an accurate reference sample that reflects the population being tested.

This leads to the inference that while Bernard's data set is still useful at this stage for the discrimination of male speakers of AE, as more time passes and further change occurs its usefulness as a reference distribution will diminish. The Cox data, while newer and performing better than the Bernard data for General AE speakers, does not cover the breadth of accent type presented in the Bernard data. A data set capturing non-contemporaneous tokens of the different vowels of contemporary AE, and encompassing speakers of different accent classifications, ages, and genders is thus an extremely important project, not only as a reference distribution for FSI in Australia, but also as a way of documenting and recording the sound change ongoing in Australian English.

5. Acknowledgments

The experiments conducted for this paper would not have been possible without the speakers who allowed me to record their speech during 2003 for Alderman (2004). Thanks are extended to the Australian National Dictionary Centre, which has funded my participation in this conference and also provided much support. Thanks must also be extended to Phil Rose, who has provided much help and advice.

6. References

Aitken, C.G.G. (1995) *Statistics and the Evaluation of Evidence for Forensic Scientists*. Wiley: Chichester.

Alderman, T. (2004). Refining the Likelihood Ratio Approach to Forensic Speaker Identification – Effects of Non-Normality in the Background Distribution as Modelled with the Bernard Data for Australian English. Unpublished First Class Honours Thesis, Australian National University.

Bernard, J.R.L. (1970) "Towards the acoustic specification of Australian English". *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikations-forschung*. 23. pp113-28.

Bernard, J.R.L. (1989) "Quantitative aspects of the sound of Australian English". Blair & Collins (Eds.) A*ustralian English: The Language of a New Society.* University of Queensland Press: St. Lucia. pp 187-204.

Cox, F (1999). "Vowel Change in Australian English." *Phonetica* 56: 1-27.

Drygajlo, A., Meuwly, D. and Alexander, A. (2003) "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition". *Proceedings of 8th European Conference on Speech Communication and Technology.*

González-Rodríguez, Ortega-García & Sánchez-Bote (2002). "Forensic Identification Reporting using Automatic Biometric Systems". In D. Zhang (Ed), *Biometric Solutions for Authentication in an e-World.* Kluwer Academic Publishers.

Kirkland, J (2003) Forensic Speaker Identification Using Australian English Fucken: A Bayesian Likelihood Ratio-based Auditory and Acoustic Phonetic Investigation. Unpublished Honours Thesis. Australian National University.

Lindley, D. V. (1977). "A problem in forensic science." *Biometrika* 64(2): 207-213.

Mitchell, A.G. & Delbridge, A. (1965). *The Pronunciation of English in Australia (Revised Edition).* Angus & Robertson: Sydney.

Robertson, B. & Vignaux, G.A. (1995) *Interpreting Evidence.* Wiley: Chichester.

Rose, P. J. (2002). *Forensic Speaker Identification*. Taylor & Francis: London.

Rose, P.J. (2003). *The Technical Comparison of Forensic Voice Samples*. Issue 99, *Expert Evidence*, (series eds Freckleton, I and Sydney, H.). Thomson Lawbook Company, Sydney, 2003.

Rose, P. J., Lucy, D, and Osanai, T. (2004). "Linguistic-Acoustic Forensic Speaker Identification With Likelihood Ratios From A Multivariate Hierarchical Random Effects Model: A "Non-Idiot's Bayes" Approach." *Proceedings of Tenth Australian International Conference on Speech Science and Technology*. ASSTA.