A WORD LIKE PHONETIC SEQUENCE STATISTICAL SOURCE MODEL FOR LARGE LEXICON AUTOMATIC SPEECH RECOGNITION SYSTEMS

A. Falaschi

ABSTRACT

Here is presented a phonetic source model whose parameters, estimated from phonetically transcribed texts, reflect the non-stationary phoneme conditional probability which is proper of a given language. Such a model will give a priori knowledges about the allowed phonetic sequences probabilities for a very large vocabulary speech recognizer, where the lexical access is made after phonetic decoding. After a discussion about the probability estimation method, model features and performances are given.

THE PHONETIC SOURCE MODEL

Motivations and Concepts. The speech events probability of occurrence knowledge has been proved to be very useful in improving the performances available from acoustical features based speech recognizer (SR) (ref 1-2). As the lexicon size and the pragmatic focus of a SR increase, the language model probabilities estimation becomes more complicated, and often the problem is decomposed in a hierarchical manner by using broad equivalence classes. Moreover, the improvement of the SR capabilities is not a trivial task, and often the algorithms fail abruptly when applied to more ambitious projects.

We try to derive a truly general source model by estimating the probability of a general phonetic sequence once given the spoken language; this is done by means of a first-order markov approximation of the phoneme non-stationary conditional probability. The non stationary feature comes out from considerations about the morphological structure of a word (positional distribution of suffixes, prefixes, etc.), suggesting to take into account as conditioning events the source internal state represented by the syllable number and syllable internal state.

Structure and Properties. Let we define a generalized word as a sequence of syllables, each syllable being composed of a vowel surrounded by two optional consonant clusters; only one vowel in a word is stressed. Such a formal description is represented by the Word Source Model (WSM) main structure of Fig.1 and the Syllabic Source Models (SSM) of Figg. 2-3. Every S_i of the WSM is a phonetic sequence statistical model for the i-th syllable position, given that the stress has yet been observed (a=1, lower row) or not (a=0, upper row). The upper backward transition in Fig.1 accounts for apostrophed words. The SSM shows the italian syllabic allowable phoneme class sequences as a transition diagrams, describing only structural properties of the source; the first order markov approximation will be taken into account by adding to every state of the SSMs as many states as the number of phonemes that can reach such position, and providing for each state as many transitions as the number of phonemes-destinations pair which can take place from there.

We further point out that the model do not account for intra-word dependencies, so every phonetic

INFOCOM Dept., University of Rome "La Sapienza", Via Eudossiana 18, 00184 ROME, ITALY
sequence ends with a blank which constitutes the memory for every word that can be generated. The states can thus be ordered in such a way all the transitions ending in state number \( n \) comes from states with lower number; the only exceptions are due to states which accounts for the memory of apostrophed words.

**THE ESTIMATION METHOD**

**Maximum likelihood from training set.** A set of 14 training text files, giving a total of 4981 word occurrences and 2073 different lexical items is phonetically transcribed and syllabized by a computer program, in agreement to the rules described in ref 3. A maximum likelihood estimate of the model transition probability is then derived from this corpus. Such an estimate results unsatisfactory because the source is unable to generate sequences not observed in the training set analysis, in contradiction with the desired very large lexicon modelling capability. It is then necessary to make use of another probability estimation method; our approach is very similar to the one discussed in ref 4.

**Back-off philosophy.** The method is based on the Turing conjecture (ref 5), which hypothesizes that the probability of never seen events and only one time observed events should be equal. Since every event, when observed for the first time, can be regarded as unique, the observation of \( N_e \) different events along an \( N \) symbol sequence will produce an unseen event probability \( P_U = N_e/N \). Denoting \( n_i \) as the number of occurrences of the \( i \)-th event \( (N = \sum_i n_i, \ i=1,\ldots,N_e) \) and \( P_i \) its probability, the stochastic constraint \( P_U + \sum_i P_i = 1 \) is easily respected if the ML estimate for observed events is substituted with the value \( P_i = (n_i-1)/N \). The probability \( P_U \) is then distributed among the unseen events according with a priori information about their probability. Such knowledge is obtained by a hierarchical grouping of the states of our source model.

**Structure-driven probability estimation.** Let us indicate as \( P_{S,d}(a,i,n,r) \) the probability that the source state labeled \( (a,i,n,r) \) will emit the symbol \( s \) through a transition ending in a destination \( d \) (\( n \) represents \( S_i^a \) internal state number and \( r \) the previous phoneme identity). By taking the simplifying assumption of statistical independence between \( s \) and \( d \), the transition probability estimation can be decomposed in the \( P_{S}(a,i,n,r) \) and \( P_d(a,i,n,r,s) \) estimations, whose product will give the desired probability estimate. The back off procedure quickly described above is then recursively applied to both the conditional probabilities as reported in the scheme of table I, where ^ and ~ denote respectively ML.
Symbol probability
\[ \hat{P}_s(a) \quad \hat{P}_s(a,n) \quad \hat{P}_s(a,r) \quad \hat{P}_s(a,n,r) \quad \hat{P}_s(a,i,n,r) \]

Destination probability
\[ \hat{P}_d(a,n) \quad \hat{P}_d(a,n,s) \quad \hat{P}_d(a,i,n,s) \quad \hat{P}_d(a,i,n,r,s) \]

Table I - Back Off Estimation Method Hierarchical Scheme

and back-off estimates. The falling arrows come from the ML estimate to be "corrected", and the right arrows come from the "a priori" probability knowledge about unseen events.

In the method implementation the sequence of approximated models changes as function of the syllable number: in fact for \( i+a > 2 \) is preferable to take as a priori information the previous syllable probability estimate instead of the general one, because the syllable type distribution becomes heavily influenced by the suffixes (see also below).

Phonological constraints filters. The probability estimation method above outlined gives a lot of impossible phoneme sequences, even if with a probability lower than the allowed ones. In order to avoid the modelling of non-existent phoneme sequences, a set of phonological rules have been used as filters in the back-off procedure.

ESTIMATED SOURCE FEATURES

Non stationary source entropy. A first evaluation of the estimated word source model is done in Fig.4, by comparing its non stationary entropy after both the ML and BO transition probability estimation algorithms. The main mismatch is due to the last syllables entropy values. In fact very few words are so long in the training corpus, and the ML estimate (black bars) gives a near deterministic behaviour for such syllable indices. On the contrary, the loss of statistical significativity suggest to the back off estimation method (grey bars) of taking as probability estimates values not related to the morphological state. A tentative solution to this situation is to set a probability threshold to be passed by the probability estimates before they are inserted into the model. This threshold varies linearly from \( 10^{-9} \) to \( 10^{-1} \) when the syllable number goes from 1 to 9. Other variation laws are under study.

Source and test set incremental entropy. The capability of the phonetic source model of accounting also for words outside the training set is tested by means of a comparison between the logprob of a test file 670 words long and the conditional source entropy, as a function of the training set size. Fig 5 reports such values for both ML and BO estimates. The entropy saturation effect of the ML estimated source evidences the significativity of the training set used, and the stability of the BO estimated source entropy indicates a relative insensitivity of the method from the training set composition. Finally, the test set logprob decrease indicates how the addition of more training data make the source probability closer to the real situation also for unknown sequences.

Source cardinality and generable sequences. After ML estimation the source model has 505 states and 2682 transitions: these numbers become 1707 and 25449 after BO estimation. As an example of the source modelling capability, Fig. 6 reports some word-like phonetic sequences obtained in a free evolution of the source, together with their logprob/symbol.
CONCLUSIONS

A generalized phonetic source model has been presented, together with the method adopted for estimating the model parameters. It has been demonstrated its modelling capability for never seen events and morphological non-stationarity of word-like phonetic sequences. The next step in this research will be to prove the utility of such of a model in the framework of a phonetic recognizer.

REFERENCES

2. LR Bahl et al, Proc of ICASSP-84, 26.5.1, S. Diego 1984
3. RA Hall Jr, La Struttura dell'Italiano, Armando Armando Ed, Roma, 1971

Fig. 4 - Conditional Non Stationary Source Entropy

Fig. 5 - Incremental Source Entropy and Test Set LogProb

Fig. 6 - Source Generated Word-Like Phonetic Sequences